Estudio exploratorio de la aplicación de la ciencia de datos en las materias del ciclo profesional del pénsum de la carrera de ingeniería de petróleos de la Universidad Industrial de Santander

Brian Ferney Ballesteros Barragán y William Felipe Sánchez Rondón

Trabajo de Grado para Optar el Título de Ingenieros de Petróleos

Director
Fernando Enrique Calvete González
Magíster en Informática

Universidad Industrial de Santander
Facultad de Ingenierías Fisicoquímicas
Escuela de Ingeniería de Petróleos
Bucaramanga
2021

Dedicatoria

A mi mamá, por todos los sacrificios que hizo, y sigue haciendo, para darme la oportunidad de educarme y tener un mejor futuro, además de inculcarme los valores que definen mi persona.

Al amor de mi vida, Ángela Torrealba, por ser mi apoyo en todo momento, por animarme cuando el desaliento se presenta, por ser mi motivación para dar todo de mí en cada meta que me propongo y en cada proyecto que emprendo, por estar siempre para mí y hacer mi vida feliz.

A mi abuelita, q.e.p.d., quien siempre creyó en mí, siempre me animó a cumplir mis sueños y siempre me demostró con su ejemplo lo que es ser una buena persona.

A mi hermano, quien ha estado en mi vida desde siempre y ha alegrado mis días con su presencia.

Brian Ballesteros.

Dedicatoria

Dedico con tu mi corazón mi proyecto de grado a Dios, a mi madre Esmeralda, a mi abuelita Fanny y a mi tía Sandra por su constante e incondicional apoyo en todo mi proceso educativo y en la realización de este trabajo.

Atte. Felipe Sánchez

Agradecimientos

Expreso mis más sinceros agradecimientos a las siguientes personas, quienes de una u otra manera contribuyeron al éxito de esta importante meta:

Al ingeniero Fernando Calvete, nuestro director de tesis, por su compromiso con el proyecto, su paciencia, su buena disposición y sus invaluables y siempre sabias orientaciones.

A la profesora Martha Parra y su esposo, el ingeniero Héctor, por todo el apoyo que siempre me prestaron y por la confianza que depositaron en mí.

A mi amigo y compañero de tesis, Felipe Sánchez, por todo el esfuerzo que dedicó en el desarrollo de este proyecto y por su buen humor, que siempre alegró hasta los más complicados momentos.

A la Universidad Industrial de Santander y a todo el cuerpo docente que me transmitió sus conocimientos y me formó para ser el profesional que hoy soy.

Brian Ballesteros.

Agradecimientos

Llegar a una meta implica recorrer un camino superando obstáculos, viviendo buenos y malos momentos, teniendo aciertos y cometiendo errores, pero, sobre todo, aprendiendo durante el recorrido de dicho camino. Al llegar a este punto sólo puedo definir esta parte de mi vida como una aventura y, como toda aventura, culmina con un tesoro aguardándonos al final de la misma, pero mi tesoro no es mi título de pregrado, sino que son las personas que estuvieron conmigo en todo este proceso, sin más preámbulo, sólo me queda agradecerles:

Primero, gracias a Dios por darme los talentos, temple, fortaleza, sabiduría, salud y determinación para cumplir mi objetivo.

A mi mamá, mi tía y mi abuelita que me apoyaron económicamente desde mi primer semestre hasta la fecha.

A mis amigos Amaury, Jennifer, Cristián, Brandon y Helmer por siempre creer en mí y darme palabras de aliento además de asesorarme y brindarme su ayuda cuando la necesité.

A los profesores que lograron inspirarme, guiarme y transmitirme el conocimiento. Entre los docentes que más aprecio y respeto puedo destacar a: Zuly Calderón, Fernando Calvete, Wilson Carreño, Adolfo Polo, Jimena Gómez, Olga Ortiz, Carmen Lugo y Samuel Muñoz.

A Fabiana, una persona especial e importante para mí que aportó a mi proceso estos últimos semestres.

A Iván Blanco, compañero, amigo y soporte en todo este tiempo. Alguien que nunca perdió la fe en mí y que, desde el ejemplo y su incondicionalidad, siempre logró inspirarme y mantenerme enfocado.

Por último y no menos importante, menciono a Brian Ballesteros, mi compañero de proyecto y estudios y también mi amigo, al cual destaco por su excelente desempeño en todas las actividades académicas en las que alguna vez participamos juntos, su comprensión ante las diferentes problemáticas que tuve y su apoyo en todo el proceso.

A todos, ¡Muchas gracias!

Atte. Felipe Sánchez

Tabla de Contenido

Pág.

Lista de Tablas

Pág.

Lista de Figuras

Pág.

Lista de Apéndices

Pág.

Glosario

Agrupamiento o Clustering: método de clasificación que se basa en reunir objetos (observaciones, datos, etc.), que comparten una característica en común, en clústeres o grupos. Es comúnmente utilizado en la técnica de lógica difusa.

Algoritmo: serie de pasos finitos, precisos y ordenados cuya intención es dar solución a un problema. En el ámbito de la programación, es una herramienta de comunicación entre programadores.

Automatizar: delegarle a un programa o máquina que lleve a cabo tareas repetitivas, simples, complejas o extensas que se realizaban antes por personas o de manera manual reduciendo el trabajo humano.

Axón: parte de la neurona que sirve como canal, paso o transmisor de la información, en forma de impulso eléctrico, hacia las otras neuronas.

Breakout: la rotura o quiebre de la pared del pozo, alrededor de la broca de perforación, debido a un esfuerzo horizontal desigual en la formación. Se forma preferencialmente en la dirección del esfuerzo mínimo horizontal ($S_{h\ min)}$. La identificación de su orientación es muy importante ya que a partir de ella se determina la orientación del esfuerzo horizontal máximo (es perpendicular al breakout).

Característica: cualidad o peculiaridad que puede ser cualitativa o cuantitativa y que sirve para distinguir a un individuo de un grupo.

Clasificar: acción de integrar un individuo u objeto a una categoría o grupo teniendo en cuenta ciertos criterios de selección y las características propias del individuo.

Código: en programación, es el conjunto de instrucciones de un algoritmo que son escritas en un lenguaje de programación, respetando una sintaxis determinada, con el objetivo de que sea interpretado por la máquina para que se lleven a cabo dichas instrucciones. Puede ser modificado.

Coeficiente de determinación (R^2): cuando se construye un modelo para un conjunto de datos dispersos se espera que este se acople a todos los datos, sin embargo, esto no sucede a menudo. El coeficiente de determinación es la medida del ajuste del modelo a los datos y va de 0 a 1, donde 1 significa que se ajusta perfectamente y 0 significa que no se acopla.

Covarianza: valor que posibilita la comparación de 2 variables distintas pero que a su vez tienen un grado de dependencia. Es la variación conjunta de 2 variables aleatorias teniendo como base en sus promedios.

Datos: representación de la información, en forma cualitativa o cuantitativa, que permite realizarle diferentes tratamientos con el objetivo de clasificarla, analizarla, compararla, estudiarla, tabularla, etc.

Dendrita: parte de la neurona encargada de la recepción de información que viene en forma de impulso eléctrico. También es la encargada de la alimentación celular en la neurona.

Entrenamiento: cuando se aplican modelos con base en ciencia de datos o aprendizaje automático, se le denomina entrenamiento a la fase de desarrollo, adaptación, mejora y estructuración del modelo a través de datos suministrados.

Epoch: concepto que se usa en el sistema de inferencia adaptativa neuro-difusa (ANFIS) y define la cantidad de iteraciones que se ejecutan durante el entrenamiento del modelo. Este valor se define con base en el error tolerable y el número de conjuntos de datos.

Error porcentual absoluto promedio (MAPE): herramienta estadística que permite medir la precisión de un conjunto de estimaciones o predicciones respecto a datos experimentales o de referencia. Permite visualizar de manera global el error porcentual absoluto presente en el modelo, fórmula o regresión que se quiera implementar.

Estocástico: hace referencia a la probabilidad, es decir, cuando un proceso es estocástico es equivalente a decir que el suceso es probabilístico o que tiene una probabilidad asociada.

Estructura: para el caso de un programa informático, se define como la cantidad, complejidad y el orden de los bloques o partes que lo conforman, donde cada bloque es un conjunto de objetos y/o funciones que tienen una tarea determinada.

Interpolación: operación que consiste en encontrar nuevos datos a través de información disponible; lo anterior se lleva a cabo a través de una aproximación matemática y gráfica del comportamiento de los datos. La interpolación se centra en determinar un nuevo punto que está dentro del intervalo de datos suministrados.

Iteración: repetición de una instrucción o proceso de cálculo. En un programa o procedimiento numérico, el número de iteraciones está definido por una meta que a su vez se representa como un error máximo permitido o tolerable para el valor calculado.

Lógica: en el contexto de la programación, es un concepto que se basa en que las instrucciones de un programa, método o procedimiento se desarrollen de forma coherente, de acuerdo con el contexto actual y sin que haya contradicciones entre ellas.

Máquina: en términos generales, es cualquier objeto fabricado que lleve a cabo o facilite un trabajo, por ejemplo, la polea. Dentro del contexto de las técnicas de la ciencia de datos y el aprendizaje automático, la palabra máquina hace referencia a dispositivos electrónicos donde, en la mayoría de los casos, se refiere a computadores.

Modelo: solución o abstracción matemática de un fenómeno determinado, es decir, una fórmula, relación o ecuación que busca representar o generar una teoría sobre el suceso o fenómeno que se quiere estudiar.

Patrón: conjunto de objetos o sucesos que se repiten o que son recurrentes respetando una secuencia determinada.

Predicción: se define como una estimación o cálculo que se proyecta al futuro con el fin de pronosticar un suceso, evento o resultado.

Procedimiento: sucesión de pasos; método o manera de llevar a cabo una acción específica.

Procesamiento: dentro de la computación, es la aplicación de una serie o conjunto de operaciones sobre un grupo de datos, esto a través de un procesador, con el objetivo de obtener la información o respuesta deseada.

Programa: serie de instrucciones o pasos para que el computador desarrolle una tarea o solucione un problema. Se diferencia del algoritmo en la necesidad de estar escrito en algún lenguaje de programación, es decir, el algoritmo es sólo una descripción del programa en lenguaje humano.

Regresión: en términos generales, es la aproximación para definir la relación entre 2 o más variables a través de un conjunto de datos y un modelo matemático que puede ser lineal, polinómico, exponencial, etc. Una buena herramienta para realizar una regresión es graficando la información suministrada para visualizar una posible tendencia.

Sensor: dispositivo electrónico capaz de detectar estímulos externos y transformar esas detecciones en señales eléctricas para transmitirlas a un computador u otro dispositivo para su posterior análisis.

Simulación: visualización y ejecución de un experimento digital en una computadora utilizando datos, modelos, suposiciones, restricciones y bases teóricas con el objetivo de extraer nueva información a partir de los resultados de este proceso.

Sistema: en términos generales, se puede definir como un objetivo que tiene una frontera o límite demarcado, sea de manera física o imaginaria, además posee una composición, estructura y entorno definidos.

Técnica: aquel procedimiento que requiere conocimientos previos para ser ejecutado, además de que en algunos casos es necesario precisar de recursos o cualidades específicas de quienes lo ejecutan.

Telemetría: es una de las bases del internet de las cosas (IoT) y comprende los procesos, funciones y dispositivos que se pueden medir o monitorear a distancia a través de una conexión inalámbrica (internet).

Validación: fase final del desarrollo de un modelo donde se utilizan datos experimentales o de referencia, que han sido previamente verificados, con el fin de determinar la desviación de los

resultados obtenidos a través del modelo y la información suministrada. Si el error no excede el criterio establecido, el modelo será aceptado.

Variable: en computación, se define como el espacio de memoria que contiene un dato al cual se le hace referencia a través de un identificador llamado nombre de variable.

Varianza: valor estadístico que expresa la dispersión de un conjunto de datos, es decir, qué tan separados están los unos de los otros si se graficaran y se analizara su tendencia.

Resumen

Título: Estudio exploratorio de la aplicación de la ciencia de datos en las materias del ciclo profesional del pénsum de la carrera de ingeniería de petróleos de la Universidad Industrial de Santander[*]

Autor: Brian Ferney Ballesteros Barragán, William Felipe Sánchez Rondón[**]

Palabras Clave: data science, machine learning, inteligencia artificial, O&G, redes neuronales artificiales, algoritmos.

Descripción: La ciencia de datos es una disciplina que combina la estadística, las matemáticas y la computación para procesar y analizar grandes volúmenes de datos y extraer información útil de ellos. Promete beneficios muy interesantes que, en general, contribuyen a la toma de decisiones, a la optimización de procesos y a la automatización de tareas; lo anterior se traduce en un incremento de la productividad y una disminución de costos, por lo que diferentes industrias la han estado implementando con el fin de hacer más rentable su actividad económica. La industria petrolera no podría ser la excepción.

El presente proyecto de investigación se enfoca en estudiar las posibles aplicaciones de la ciencia de datos en los procedimientos de ingeniería que se realizan en las asignaturas profesionales del programa de ingeniería de petróleos de la UIS. Para ello, se empieza por definir detalladamente los conceptos fundamentales relacionados con esta disciplina; luego, se continúa revisando el pénsum del programa de ingeniería de petróleos y se identifican los procedimientos que son potenciales candidatos para aplicar en ellos técnicas de la ciencia de datos, además, se eligen algunos de esos procedimientos y se proponen ejemplos en los cuales se compara la metodología tradicional y la metodología basada en la previamente mencionada ciencia para llevar a cabo tales procedimientos y los resultados obtenidos con cada una, todo ello apoyándose en estudios publicados en la literatura científica de la industria petrolera; posteriormente, se describe metodológicamente el proceso de implementación de las más importantes técnicas y algoritmos de la ciencia de datos; finalmente, con el propósito de demostrar los beneficios de su aplicación, se presentan algunos casos que se encuentran publicados en diversas revistas científicas de renombre cuyos resultados ponen de manifiesto el potencial de la estudiada disciplina.

[*] Trabajo de Grado

[**] Facultad de Ingenierías Fisicoquímicas. Escuela de Ingeniería de Petróleo. Director: Fernando Enrique Calvete González. Magíster en Informática.

Abstract

Title: Exploratory study of the application of data science in the subjects of the professional cycle of the pensum of the petroleum engineering career of the Industrial University of Santander[*]

Author(s): Brian Ferney Ballesteros Barragán, William Felipe Sánchez Rondón[**]

Key Words: data science, machine learning, artificial intelligence, O&G, artificial neural networks, algorithms.

Description: Data science is a discipline that combines statistics, mathematics, and computing to process and analyze large volumes of data and extract useful information from them. It promises very interesting benefits that, in general, contribute to decision making, process optimization and task automation; this leads to an increase in productivity and a decrease in costs, which is why different industries have been implementing it in order to make their economic activity more profitable. The oil industry could not be the exception.

This research project focuses on studying the possible applications of data science in the engineering procedures carried out in the professional subjects of the UIS petroleum engineering program. To do this, it begins by defining in detail the fundamental concepts related to this discipline; then, the pensum of the petroleum engineering program is reviewed and the procedures that are potential candidates for applying data science techniques are identified, in addition, some of these procedures are chosen and examples are proposed in which they are compared the traditional methodology and the methodology based on the aforementioned science to carry out such procedures and the results obtained with each one, all based on studies published in the scientific literature of the oil industry; later, the process of implementation of the most important techniques and algorithms of data science is methodologically described; finally, in order to demonstrate the benefits of its application, some cases, that are published in various renowned scientific journals, are presented and the results of which reveal the potential of the studied discipline.

[*] Degree Work

[**] Faculty of Physicochemical Engineering. School of Petroleum Engineering. Director: Fernando Enrique Calvete González. MSc in Informatics.

Introducción

La situación actual de la industria del petróleo, caracterizada por una disminución de las reservas de hidrocarburos convencionales y un costo creciente en las operaciones de exploración y explotación como consecuencia de la mayor dificultad para el aprovechamiento de fuentes alternativas como los yacimientos en aguas ultraprofundas, las formaciones de shale o los crudos extrapesados o la necesidad de implementar técnicas y tecnologías para seguir produciendo los campos maduros, sumada al hecho de que esta industria es responsable de suplir el 56% de la demanda energética mundial (BP, 2021), constituyen una fuerte motivación para los profesionales de la industria petrolera que se ven impulsados a buscar soluciones innovadoras y eficientes que contribuyan a aumentar la productividad y disminuir los costos.

En tal sentido, la ciencia de datos se presenta como una excelente opción a considerar, ya que promete interesantes beneficios como lo son el aumento de la producción mediante la automatización de los procesos, operaciones más seguras gracias al monitoreo constante de equipos y procedimientos y el reemplazo de personal humano por máquinas autónomas en labores de alto riesgo y, finalmente, la reducción de costos asociada a un menor requerimiento de personal, mantenimientos preventivos a los equipos que evitan daños mayores en los mismos y la anticipación a problemas graves como podrían ser una falla severa en las estructuras, un reventón o el colapso del pozo, problemas tales que podrían costar vidas humanas y hasta el fracaso del proyecto.

Atendiendo a los motivos previamente expuestos, surge la necesidad de realizar un estudio sobre la potencial aplicación de la ciencia de datos en los procedimientos (generalmente, de cálculo o diseño) que se llevan a cabo en el desarrollo de las diferentes asignaturas que forman parte del ciclo profesional del programa de ingeniería de petróleos de la Universidad Industrial de

Santander. Tal estudio podría suponer el punto de partida para realizar una reestructuración del plan de estudios del programa en la cual se considere la inclusión de la formación en ciencias de datos dentro del mismo con el fin último de dotar de herramientas y habilidades a los profesionales que se gradúan de la carrera y que se enfrentan a un panorama laboral en el que cada vez cobran más relevancia las tecnologías de la cuarta revolución industrial, dentro de las que se incluye, evidentemente, la ciencia de datos.

El estudio se desarrolla en cuatro etapas bien diferenciadas: la primera etapa consiste en definir detalladamente cada una de las técnicas de la ciencia de datos (a saber, Machine Learning, redes neuronales artificiales, minería de datos, Big Data Analytics, algoritmos genéticos y lógica difusa) y algunos conceptos que son aplicaciones directas de la ciencia de datos a la industria petrolera, tales como los pozos inteligentes y los campos de aceite y gas digitales, entre otros; la segunda etapa se enfoca en determinar a cuáles de los procedimientos de ingeniería de las asignaturas del programa de ingeniería de petróleos de la Universidad Industrial de Santander se les puede aplicar técnicas de la ciencia de datos; la tercera etapa presenta la metodología general para la implementación de las técnicas y algoritmos de la ciencia de datos; finalmente, en la cuarta etapa se presentan algunos casos documentados de la aplicación de la ciencia de datos a diferentes procedimientos de la industria petrolera que demuestran la utilidad de emplear esta ciencia y los beneficios que ello conlleva.

1. Objetivos

1.1 Objetivo General

Realizar un estudio que permita definir la aplicación de las técnicas de la ciencia de datos a los diferentes procedimientos de ingeniería que se efectúan en las materias del ciclo profesional del pénsum de la carrera de ingeniería de petróleos de la Universidad Industrial de Santander.

1.2 Objetivos Específicos

Definir, de manera clara, concisa y detallada, las técnicas de la ciencia de datos que existen en la actualidad.

Definir los procedimientos de diseño de ingeniería en las materias del pensum de la carrera de ingeniería de petróleos donde se puedan aplicar las técnicas de la ciencia de datos.

Describir metodológicamente la aplicación de las técnicas de la ciencia de datos en cada uno de los procedimientos de diseño de ingeniería de petróleos seleccionados.

Comprobar con ejemplos de la literatura los resultados obtenidos en la aplicación de las técnicas de la ciencia de datos a algunos procedimientos de la ingeniería de petróleos analizados en este proyecto de grado.

2. Marco teórico

2.1 Ciencia de datos

2.1.1 Definición

Como su nombre lo sugiere, la ciencia de los datos o *data science,* que es como se le conoce más comúnmente, se fundamenta en los datos y es absolutamente dependiente de ellos, por tal motivo se hace necesario que, antes de pasar a definir formalmente esta ciencia, se explique el concepto de los datos en sí.

Según (Kelleher y Tierney, 2018), los datos, en su forma más básica, son abstracciones de una entidad del mundo real (personas, objetos o eventos). Se suelen usar indistintamente los términos *variable, característica* o *atributo* para denotar abstracciones individuales. Es importante tener en cuenta que cada entidad es descrita por un número de atributos. En otras palabras, los datos son información sobre los atributos que describen a una entidad.

Los datos pueden presentarse en dos formas diferentes:

- Datos estructurados: son datos que pueden ser almacenados en una tabla, y cualquier instancia en la tabla tiene la misma estructura, es decir, el mismo conjunto de atributos. Por ejemplo, los datos demográficos de una población estarán presentados en una tabla donde cada fila describe a una persona y consiste de los mismos atributos (nombre, edad, fecha de nacimiento, dirección, etc.) que se aplicarán a todas las personas y en el mismo orden. Los datos estructurados son más fáciles de almacenar, organizar, buscar, registrar y unir con otros datos estructurados.

- Datos no estructurados: son datos donde cada instancia en el conjunto de datos puede tener su propia estructura interna y la estructura no es necesariamente la misma entre las diferentes instancias. Por ejemplo, un conjunto de datos de páginas web, donde cada página web tiene su propia estructura, pero tal estructura puede ser diferente en las demás páginas.

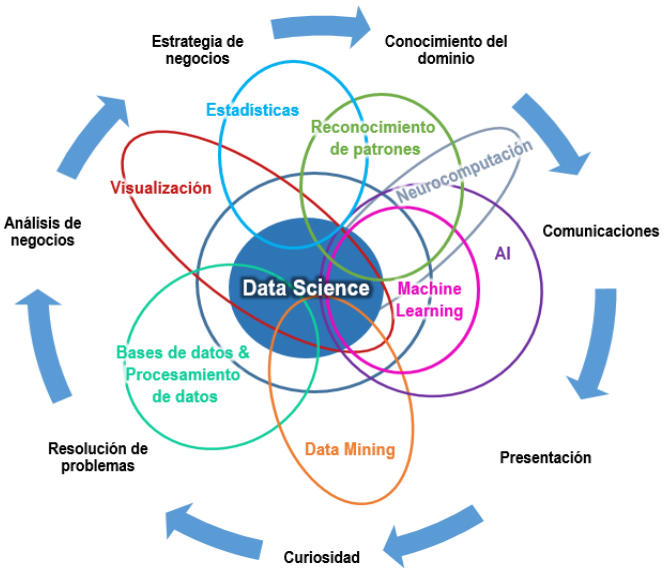
Ahora que está claro lo que son los datos, es posible definir la ciencia de los datos de forma satisfactoria. En palabras de (Ozdemir, 2016, pág. 4), data science es "el arte o la ciencia de adquirir conocimiento a partir de los datos".

Este autor ofrece una definición un poco más detallada y dice que la ciencia de los datos involucra todo lo relacionado con cómo se obtienen los datos, la manera como se obtiene conocimientos de ellos y los usos que se le dan a ese conocimiento, los cuales son, principalmente, tomar decisiones, hacer predicciones, entender el pasado o el presente y crear nuevas industrias o productos.

Por otra parte, según (Sarkar et al., 2018), la ciencia de datos es un campo muy diverso e interdisciplinario que engloba múltiples campos como lo son la inteligencia artificial, la minería de datos, la estadística, la neurocomputación, entre otros. Básicamente, la ciencia de los datos se ocupa de los principios, metodologías, procesos, herramientas y técnicas para obtener conocimiento o información de los datos. Lo anterior puede evidenciarse en la Figura 1.

Figura 1

La ciencia de datos como un campo multidisciplinario



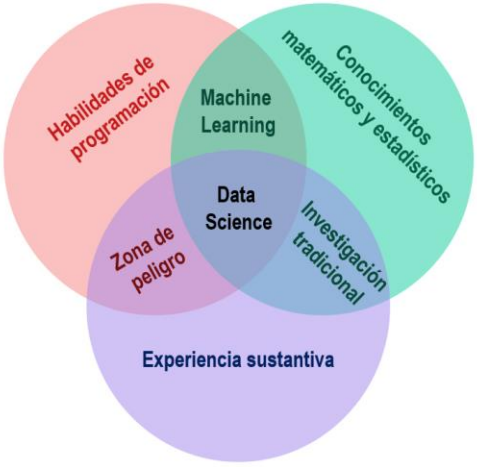
Nota. La figura muestra a la ciencia de datos como un campo multidisciplinario donde convergen diferentes y variados campos y disciplinas. Adaptado de (Tierney, 2012).

2.1.2 Componentes de la ciencia de datos

De acuerdo con (Conway, 2010), la ciencia de datos posee tres componentes principales y corresponde, además, a la intersección de dichos componentes, tal como se ilustra en la Figura 2.

Figura 2

Los tres componentes de la ciencia de datos



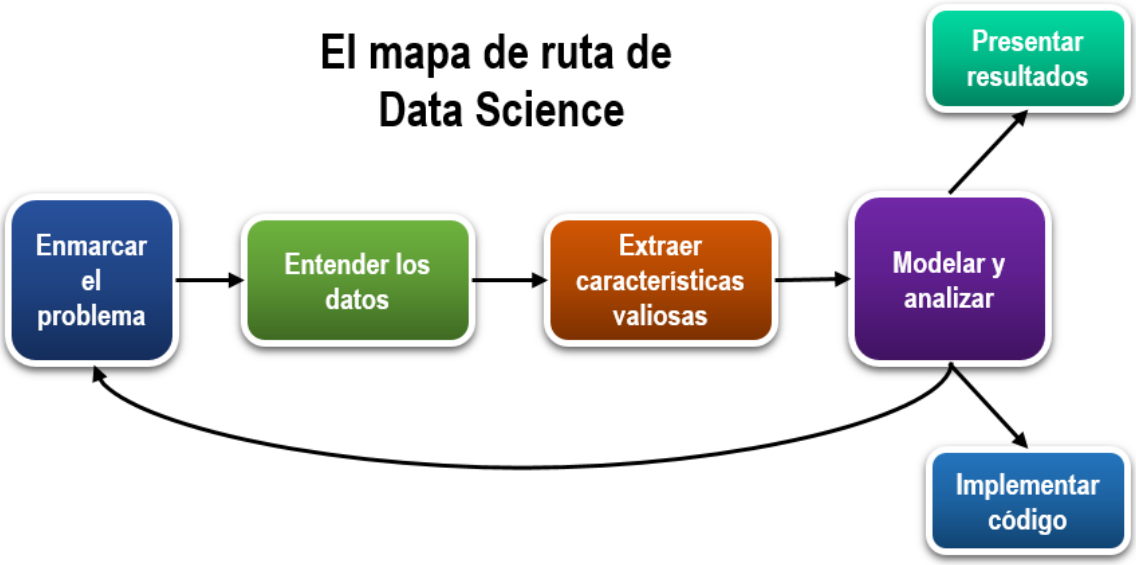
Nota. Adaptado de (Conway, 2010)

En la Figura 2, los conocimientos matemáticos y estadísticos tienen que ver con las diversas técnicas matemáticas y estadísticas computacionales y cuantitativas que permiten extraer conocimiento de los datos. Las habilidades de programación hacen referencia a la capacidad de manejar, procesar, manipular y transformar los datos en formatos fáciles de entender y de analizar. Finalmente, la experiencia sustantiva es, básicamente, el conocimiento que se posee sobre el área a la cual pertenece el problema a tratar y es la que permite aplicar los conceptos e interpretar los resultados de una manera significativa y eficaz.

2.1.3 Proceso para resolver problemas de ciencia de datos

Figura 3

El mapa de ruta de la ciencia de datos



Nota. La figura resume el proceso a seguir para resolver problemas de ciencia de datos. Adaptado de (Cady, 2017).

Para resolver un problema de ciencia de datos, el primer paso es enmarcar el problema, es decir, comprender la situación que se tiene y definir uno o más problemas analíticos a partir de esta. Posteriormente, se continúa con una etapa extensa donde se tratan y procesan los datos teniendo presente siempre las cosas del mundo real que describen, para que se puedan extraer características significativas. Finalmente, las características extraídas se incorporan a herramientas analíticas que generan resultados numéricos concretos (Cady, 2017).

De la Figura 3, es importante aclarar dos cosas: lo primero es que en el paso de "Modelar y analizar" se genera un bucle que regresa al inicio del proceso. Ello se hace para refinar el proceso y obtener mejores resultados. Lo segundo es que hay dos posibilidades para terminar el mapa de ruta: presentar resultados e implementar código; la elección de una u otra posibilidad depende de la naturaleza del problema, es decir, si lo que se está haciendo es un análisis de datos y se requiere conocer respuestas sobre el problema, entonces se presentan resultados. Si, por el contrario, el problema tiene que ver con el desarrollo de un software para una máquina, entonces se hace implementación de código.

2.1.4 Ciencia de datos e Inteligencia Artificial

Los campos de la ciencia de datos y de la inteligencia artificial están estrechamente relacionados debido a que el primero aprovecha técnicas de análisis de datos que pertenecen a la inteligencia artificial, tales como la minería de datos o el aprendizaje automático, para hacer el respectivo procesamiento a los datos y extraer de ellos información útil. Por tal motivo, en esta sección se tratará lo relacionado con la inteligencia artificial con el fin de proporcionar una visión general sobre la misma.

2.1.4.1 Definición. En palabras de (Frankish y Ramsey, 2014), la inteligencia artificial es "un enfoque interdisciplinario para comprender, modelar y replicar la inteligencia y los procesos

cognitivos humanos aprovechando diversos dispositivos y principios computacionales, matemáticos, lógicos, mecánicos e incluso biológicos".

Por su parte, (Franceschetti, 2018) define la inteligencia artificial de acuerdo con la disciplina a la cual se aplica, así:

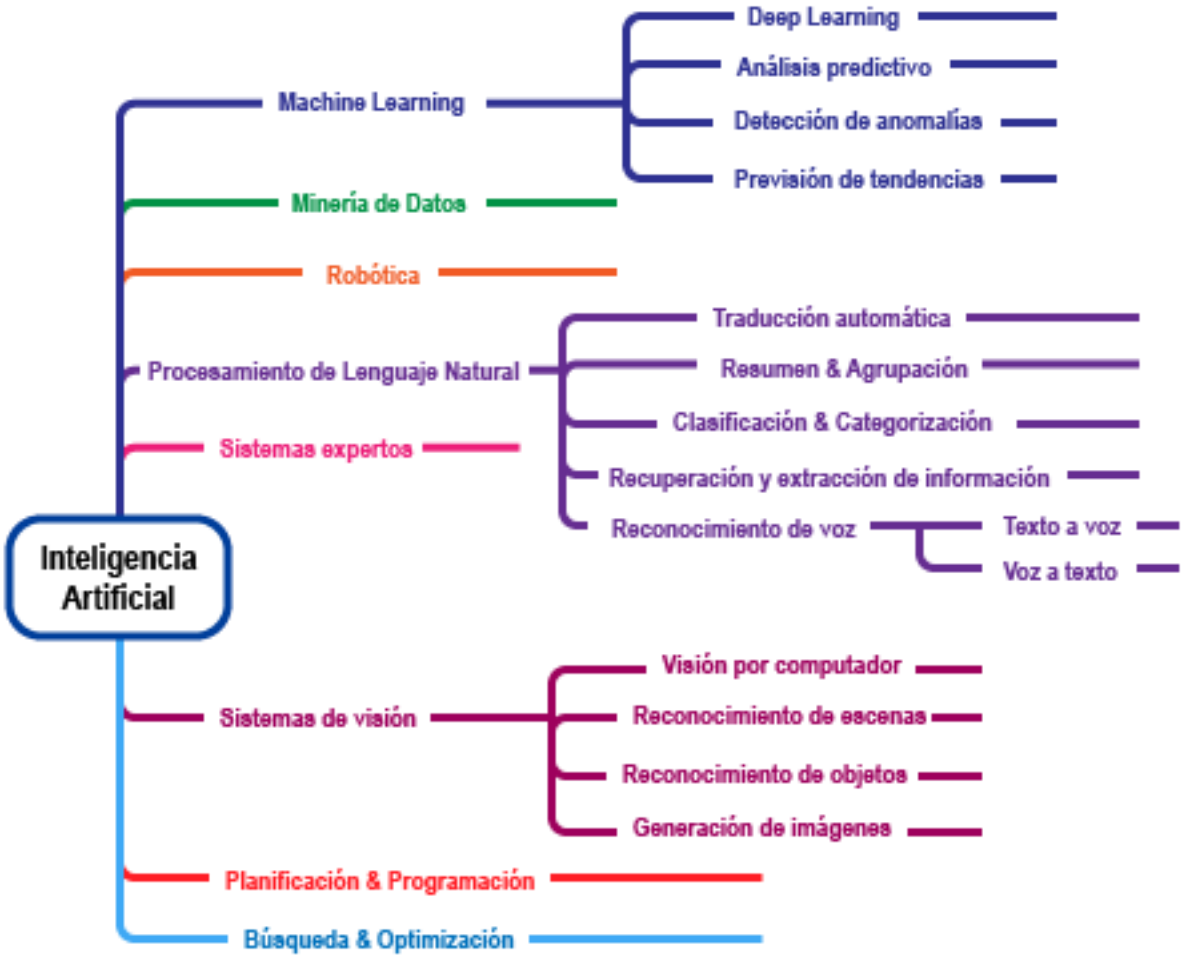
- Para las ciencias computacionales, inteligencia artificial hace referencia al desarrollo de programas que presenten comportamiento humano. Dichos programas pueden planificar inteligentemente, traducir idiomas, actuar como expertos, por ejemplo, al seleccionar el mejor vino para la cena, entre otras habilidades.
- En ingeniería, la inteligencia artificial tiene que ver con construir máquinas que ejecuten tareas a menudo realizadas por humanos. Esas tareas pueden ser actividades específicas en una fábrica, sistemas de vigilancia, labores en condiciones extremas de presión o temperatura, etc.
- Para la ciencia cognitiva, inteligencia artificial se refiere a desarrollar modelos de inteligencia humana para entender mejor el comportamiento de los humanos. Inicialmente, los modelos eran simbólicos y se enfocaban en la psicología cognitiva y la filosofía partiendo de la idea de que las regiones del cerebro realizan un razonamiento complejo procesando símbolos. Los modelos más recientes son modelos de cognición humana desarrollados para reflejar el funcionamiento del cerebro como una computadora electroquímica.

2.1.4.2 Ramas de la inteligencia artificial. La inteligencia artificial es un campo muy amplio que abarca una gran variedad de subcampos o disciplinas orientadas a resolver problemas

muy diversos, que van desde tareas de búsqueda y optimización hasta procesos de automatización industrial a gran escala. La Figura 4 ilustra el alcance de la inteligencia artificial.

Figura 4

Principales ramas de la inteligencia artificial



Nota. Adaptado de (Sarkar et al., 2018).

A continuación, se presenta una breve descripción de cada uno de los subcampos que se aprecian en la ilustración.

2.1.4.2.1 Machine Learning. El aprendizaje automático o *Machine Learning* es un campo de investigación proveniente de la intersección de la estadística, la inteligencia artificial y las

ciencias computacionales y tiene como finalidad la obtención de conocimiento a partir de datos. También es conocido como “análisis predictivo” o “aprendizaje estadístico” (Müller y Guido, 2016). Consiste en desarrollar algoritmos de predicción precisos y eficientes que aprovechan la experiencia para mejorar el desempeño de los mismos; acá, el término “experiencia” hace referencia a información pasada, disponible usualmente en datos electrónicos, que es analizada y procesada (Mohri et al., 2012).

2.1.4.2.2 Data Mining. La minería de datos, a veces también llamada “descubrimiento de conocimiento a partir de datos” (*Knowledge Discovery from Data*, KDD, en inglés), es el proceso de descubrir patrones y conocimientos interesantes a partir de grandes cantidades de datos. Las fuentes de datos pueden incluir bases de datos, depósitos de datos, la web, otros repositorios de información o datos que se transmiten dinámicamente al sistema (Han et al., 2011).

2.1.4.2.3 Robótica. Es un campo científico interdisciplinario relacionado con el diseño, desarrollo, operación y valoración de dispositivos electrónicos que ejecutarán acciones que normalmente requieren de intervención humana. Generalmente, se usan estos dispositivos en automatización de procesos consistentes en tareas repetitivas. Los dispositivos reciben el nombre de robots y consisten, generalmente, de al menos tres partes: una estructura metálica (a menudo un brazo que ejecuta determinadas labores) que es la que permite al robot interactuar con el medio; sensores que recogen las propiedades físicas tales como sonido, temperatura, presión, movimiento, etc.; y algún tipo de sistema de procesamiento que transforma los datos adquiridos por los sensores en instrucciones sobre las acciones a realizar (Franceschetti, 2018).

2.1.4.2.4 Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés). Es un campo multidisciplinario que combina conceptos de lingüística computacional, ciencias computacionales e inteligencia artificial y consiste en la habilidad de hacer que las máquinas

procesen, entiendan e interactúen con los lenguajes naturales humanos. El principal objetivo de las aplicaciones o sistemas basados en NLP es permitir la interacción entre las máquinas y los lenguajes naturales que han evolucionado a lo largo del tiempo (Sarkar et al., 2018).

2.1.4.2.5 Sistemas expertos. Son sistemas capaces de ofrecer soluciones a problemas específicos en un dominio determinado o que son capaces de asesorar, en una forma y nivel comparables a los de los expertos en la materia. Los problemas en los campos para los cuales son desarrollados los sistemas expertos son aquellos que requieren de una considerable experiencia humana para su desarrollo y, entre estos, algunos de los más comunes son: el diagnóstico médico de enfermedades, el asesoramiento financiero o el diseño de productos (Lucas y van der Gaag, 1991).

2.1.4.2.6 Sistemas de visión. En inteligencia artificial, un sistema de visión es una combinación de software y hardware que pretende imitar la visión humana. En el caso de la visión humana, el ojo captura la imagen y el cerebro la procesa y, al final, se obtiene como resultado una predicción de los componentes de la imagen basada en los datos extraídos de la misma. De manera análoga, en un sistema de visión artificial existen dispositivos sensores que hacen las veces del ojo, pudiendo ser estos una cámara, un radar, rayos X o cualquier dispositivo que le proporcione al sistema la escena completa del entorno donde ejecutará la tarea. Además, el sistema cuenta con un poderoso algoritmo que imita las funciones del cerebro y se encarga de interpretar y clasificar el contenido de la imagen (Elgendy, 2020).

2.1.4.2.7 Planificación y programación. La planificación es el proceso de poner en una secuencia u orden parcial un conjunto de actividades o acciones que cumplan unas restricciones de tiempo y recursos requeridas para lograr un objetivo determinado. Por su parte, la programación

es la asignación de tales actividades o acciones a lo largo del tiempo de acuerdo con ciertos criterios de desempeño (Spyropoulos, 2000).

2.1.4.2.8 Búsqueda y optimización. En inteligencia artificial, encontrar la solución a un problema se considera como un proceso de búsqueda por el espacio de las posibles soluciones. Por otro lado, en matemáticas e ingeniería, a esto se le considera como un proceso de optimización, es decir, encontrar la mejor solución o una solución óptima al problema. Así las cosas, la búsqueda y optimización cosiste en buscar en el espacio de las posibles soluciones la solución a un problema y que esta sea, además, la más eficiente (Chandel y Sood, 2014).

2.1.4.3 Tipos de inteligencia artificial. De acuerdo con (Mueller y Massaron, 2018), la inteligencia artificial puede categorizarse en cuatro grupos:

2.1.4.3.1 Sistemas que actúan como humanos. Un sistema que actúa como humano refleja el test de Turing, el cual mide la capacidad del sistema para presentar un comportamiento inteligente similar al de un ser humano y es aprobado cuando la diferenciación del comportamiento del sistema y del humano no es posible. Se emplea en tecnologías tales como procesamiento de lenguaje natural, representación de conocimiento, razonamiento automatizado y aprendizaje automático.

2.1.4.3.2 Sistemas que piensan como humanos. Cuando un sistema piensa como humano, puede realizar tareas que requieren la inteligencia de un humano para tener éxito, como conducir un carro. Para determinar si un sistema piensa como humano, se requiere disponer de algún método para determinar cómo piensan los humanos, lo cual lo define el enfoque de modelado cognitivo. Dicho modelo se fundamenta en tres técnicas: introspección, pruebas psicológicas e imagen mental (monitorear la actividad cerebral). Este tipo de IA es utilizado a menudo en psicología y otros

campos en los cuales el modelamiento del pensamiento humano es esencial para crear simulaciones realistas.

2.1.4.3.3 Sistemas que piensan racionalmente. Un sistema que piensa racionalmente se basa en los comportamientos registrados para crear una guía sobre cómo interactuar con un entorno de acuerdo con los datos que tenga disponibles. El objetivo de este enfoque es resolver los problemas de manera lógica, siempre que sea posible. Muchas veces, este enfoque permite la creación de una técnica de referencia para resolver un problema, la cual luego puede ser modificada para resolver realmente el problema, lo que implica que, a menudo, resolver un problema en la práctica es diferente de como se hace en teoría. Los sistemas expertos pertenecen a este grupo.

2.1.4.3.4 Sistemas que actúan racionalmente. El estudio de cómo los humanos actúan en circunstancias particulares bajo condiciones específicas permite determinar cuáles técnicas son más eficientes y efectivas. Un sistema que actúa racionalmente se apoya en las acciones registradas para interactuar con el medio en función de las condiciones, los factores ambientales y los datos existentes.

2.1.4.4 Aplicaciones de la inteligencia artificial. La inteligencia artificial, al ser un campo de las ciencias computacionales tan amplio, como ya se evidenció previamente, tiene gran cantidad de aplicaciones, muchas de ellas en el sector industrial, especialmente, en la automatización de procesos, aunque también se utiliza en gran medida en el sector médico, financiero y en la industria de los videojuegos.

A continuación, se presentan algunas de las aplicaciones más frecuentes de la inteligencia artificial (Mueller y Massaron, 2018, págs. 70-71):

Figura 5

Aplicaciones más comunes de la inteligencia artificial

2.1.5 Técnicas de la ciencia de datos

De acuerdo con la definición dada a la ciencia de datos en la sección 2.1.1, esta ciencia consiste en un enfoque multidisciplinario que aprovecha técnicas de otros campos, particularmente, de la estadística, de la inteligencia artificial y de la minería de datos, para extraer información valiosa y útil mediante el procesamiento y análisis de grandes volúmenes de datos. Por lo tanto, las técnicas de la ciencia de datos son, en esencia, técnicas o disciplinas de los otros campos.

2.1.5.1 Machine Learning.

2.1.5.1.1 Definición. La definición de Machine Learning (o aprendizaje automático, en español) más ampliamente aceptada es la proporcionada por Arthur Samuel en 1959: Machine Learning es un campo de estudio que le da a los computadores la capacidad de aprender sin ser explícitamente programados (Samuel, 1959).

La definición ofrecida por Samuel contiene la esencia del aprendizaje automático, es decir, que gracias a ello los computadores poseen la habilidad de aprender por sí mismos, sin una programación previa, sin embargo, esta definición es muy general, razón por la cual varios autores posteriores han desarrollado y presentado sus propias definiciones, más específicas y detalladas. A continuación, se exponen algunas de ellas:

- Según (Müller y Guido, 2016), el aprendizaje automático o *Machine Learning* es un campo de investigación proveniente de la intersección de la estadística, la inteligencia artificial y las ciencias computacionales y tiene como finalidad la obtención de conocimiento a partir de datos. También es conocido como "análisis predictivo" o "aprendizaje estadístico".
- Para (Bhavsar et al., 2017, pág. 283), Machine Learning es "una colección de métodos que permiten a las computadoras automatizar la creación y programación de modelos basados en datos a través de un descubrimiento sistemático de patrones estadísticamente significativos en los datos disponibles".
- (Mohri et al., 2012) definen Machine Learning como un conjunto de métodos computacionales que utilizan la experiencia para mejorar el rendimiento o para hacer predicciones más precisas. Para ellos, experiencia se refiere a información previa disponible para el software aprendiz, la cual, generalmente, se encuentra en

forma de datos recopilados y puestos a disposición para su análisis. La calidad y precisión de las predicciones realizadas están determinas por la cantidad y calidad de la información que se le suministró al sistema durante su entrenamiento.

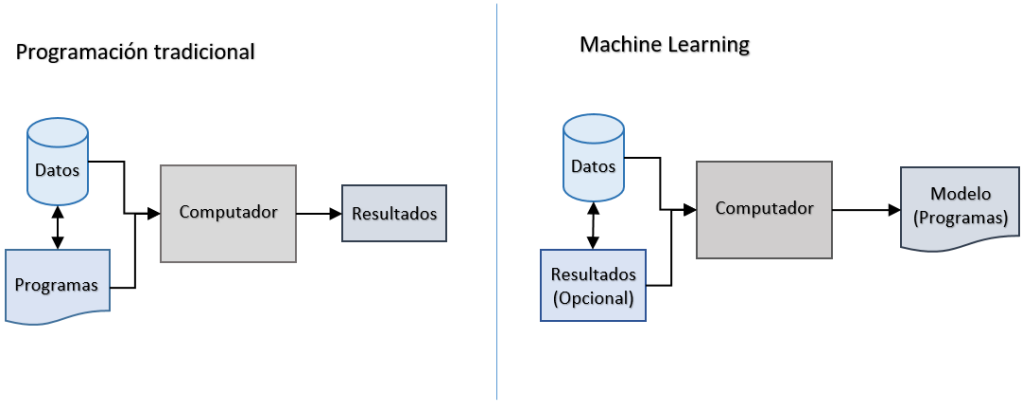
2.1.5.1.2 Machine Learning vs Programación tradicional. En la programación tradicional, al computador se le proporcionan como datos de entrada los datos del tema a estudiar y uno o más programas, generalmente codificados mediante algún lenguaje de programación. Los programas le permiten al computador trabajar sobre los datos, realizar cálculos y, finalmente, generar resultados.

Por otra parte, en Machine Learning, al computador se proporcionan como datos de entrada los datos del tema a estudiar y los posibles resultados (aunque a veces estos últimos no se proporcionan, lo que se denomina "aprendizaje no supervisado") y con ellos el computador intentar descubrir los patrones inherentes a los datos para, finalmente, generar un modelo, que sería el equivalente al programa en programación tradicional.

En resumen, en programación tradicional con los datos y uno o varios programas se generan unos resultados, mientras que en Machine Learning se usan los posibles resultados y los datos para generar un modelo (o programa). Lo anterior se puede apreciar mejor en la Figura 6.

Figura 6

Programación tradicional vs Machine Learning



Nota. Adaptado de (Sarkar et al., 2018).

2.1.5.1.3 Tipos de Machine Learning. De acuerdo con (Géron, 2019), los diferentes tipos de modelos de Machine Learning pueden ser clasificados en las siguientes 4 categorías:

- Aprendizaje supervisado: incluye algoritmos de aprendizaje que toman muestras de datos y los resultados asociados a cada muestra durante el proceso de entrenamiento del modelo. La finalidad es aprender un mapeo o relación entre los datos de entrada y sus respectivos resultados y de esa manera poder predecir un resultado para cualquier nuevo dato de entrada basándose en el conocimiento adquirido previamente. El método se denomina “supervisado” porque el modelo aprende de muestras de datos cuyos resultados se conocen de antemano. Su mayor aplicación es en análisis predictivo. Se utiliza en tareas de regresión y clasificación.
- Aprendizaje no supervisado: este método se aplica cuando no se dispone de resultados ni datos de entrenamiento preetiquetados. Se denomina no supervisado porque el modelo intenta aprender estructuras latentes inherentes, patrones y relaciones a partir de los datos disponibles y sin disponer de ninguna ayuda o supervisión como las que proporcionan los resultados conocidos de antemano.

Este método de aprendizaje tiene más que ver con intentar extraer información o conocimientos útiles de los datos antes que intentar predecir los resultados basándose en datos de entrenamiento supervisado previamente disponibles. Los resultados presentan mayor incertidumbre. Puede ser utilizado para agrupamiento (clustering), asociación y detección de anomalías.

- Aprendizaje semisupervisado: es un método híbrido que utiliza datos etiquetados (datos de entrada con su respectivo resultado) y datos no etiquetados (los datos de entrada no poseen un resultado conocido). En este método, primero se entrena el modelo usando los datos etiquetados de la misma manera que se hace en el aprendizaje supervisado y luego se aplica el modelo entrenado a los datos no etiquetados para predecir sus posibles resultados y de esa manera obtener más datos etiquetados.
- Aprendizaje por refuerzo: este método es diferente de los métodos de aprendizaje supervisado y no supervisado, ya que su objetivo es optimizar un sistema en cierto contexto. El aprendizaje por refuerzo se enfoca en encontrar estrategias y acciones óptimas en una determinada situación basándose en la retroalimentación o en la recompensa que obtiene del entorno al ejecutar ciertas acciones o tomar determinadas decisiones. Este método se suele utilizar para desarrollar juegos de lógica como ajedrez, Go o póker.

2.1.5.1.4 Algoritmos de Machine Learning más utilizados. Según (Ray, 2017), los algoritmos de Machine Learning más comúnmente utilizados son los que se listan a continuación:

- Regresión lineal: este algoritmo se usa para estimar valores reales basándose en variables continuas. Para ello, se establece una relación entre las variables

dependientes e independientes ajustando la mejor línea, la cual se denomina línea de regresión y es representada por la ecuación lineal y = mx + b, donde “y” es la variable independiente, “m” es la pendiente, “x” es la variable dependiente y “b” es el intercepto. Los términos “m” y “b” se obtienen mediante la minimización de la suma de la diferencia al cuadrado de la distancia entre los puntos de datos y la línea de regresión. Una vez conocidos “m” y “b”, es posible estimar el valor de la variable independiente a partir de los valores de la variable dependiente.

- Regresión logística: pese a su nombre, la regresión logística no es un algoritmo de regresión sino de clasificación. Se usa para estimar valores discretos (específicamente, valores binarios tales como Sí/No, 0/1, Verdadero/Falso) a partir de un conjunto dado de variables independientes. Este algoritmo predice la ocurrencia de un evento al ajustar los datos a una función logística, motivo por el cual sus valores de salida están entre 0 y 1. En otras palabras, predice la probabilidad de que la muestra dada pertenezca a un grupo o clase particular.
- Árboles de decisión (Decision Trees, DT): pertenece a los algoritmos de aprendizaje supervisado y se usan principalmente en problemas de clasificación, aunque también pueden usarse en problemas de regresión. Pueden usarse con variables dependientes categóricas y continuas. La clase de un objeto en particular está determinada por una serie de preguntas, usualmente, con respuestas Sí/No.
- Máquinas de Vectores de Soporte (Support Vector Machine, SVM): son algoritmos de aprendizaje supervisado que pueden ser utilizados tanto en tareas de clasificación como en tareas de regresión. Consiste en graficar cada elemento del conjunto de datos como un punto de un espacio n-dimensional (siendo n el número

de características disponibles) con el valor de cada característica siendo el valor de una coordenada particular.

La finalidad de este algoritmo es encontrar la línea (llamada hiperplano) que separe los datos en diferentes grupos, cumpliendo como criterio que la distancia que haya entre los puntos (de ambos grupos) más cercanos al hiperplano sea la máxima posible. Las líneas paralelas al hiperplano que contienen los puntos más cercanos a este reciben el nombre de vectores de soporte.

- Naive Bayes: es un algoritmo de clasificación que está basado en el teorema de Bayes y que asume una independencia entre las características, es decir, que la presencia o ausencia de una característica no tiene relación con la presencia o ausencia de las demás características. En este algoritmo, se considera que cada característica contribuye de manera independiente a que la variable de interés pertenezca a una determinada clase. Es muy útil para conjuntos de datos muy grandes.
- k-Nearest Neighbors (KNN): en español se conoce como algoritmo de los k vecinos más cercanos y se utiliza en problemas de clasificación y de regresión, sin embargo, su uso más amplio es en clasificación. Se fundamenta en almacenar una gran cantidad de casos conocidos y clasifica a los nuevos casos en el mismo grupo al que pertenecen la mayor cantidad de sus vecinos más próximos. Por ejemplo, un nuevo elemento cuyos vecinos son 3 círculos y dos triángulos será clasificado como círculo, ya que la mayoría de sus vecinos son círculos. Los elementos son considerados vecinos de acuerdo con una función de distancia; algunas funciones de distancia son: euclidiana, Manhattan, Minkowski, Hamming, entre otras.

- K-medias (K-means): es un algoritmo de aprendizaje no supervisado para problemas de agrupamiento o clustering (en inglés). Clasifica los datos agrupándolos de manera que los elementos dentro de un grupo son homogéneos entre sí, pero heterogéneos a los de otros grupos. Su finalidad es clasificar un nuevo elemento dentro de los grupos (o clústeres) generados durante el entrenamiento del algoritmo.
- Bosque aleatorio (Random Forest, RF): este algoritmo está compuesto por un conjunto o colección de árboles de decisión (de ahí que se llame "bosque"). Para clasificar un elemento que posee cierta cantidad de características, se selecciona aleatoriamente algunas de sus características y se evalúa en uno de los árboles que componen el bosque para que este haga su clasificación particular. Posteriormente se repite el proceso con otro árbol, pero tomando otras características (ya que se hace aleatoriamente) y así se obtiene su clasificación; el proceso se repite con todos los árboles del bosque. Finalmente, se clasifica el elemento en la clase que más se repita, por ejemplo, si 5 árboles clasificaron un elemento como "fruta" y otros 3 como "verdura", entonces el elemento se clasifica como fruta.
- Algoritmos de reducción de dimensionalidad: como se puede inferir de su nombre, la finalidad de estos algoritmos es reducir la cantidad de variables aleatorias de un problema a fin de obtener ventajas tales como facilitar el procesamiento, disminuir el ruido (distorsión) de los datos, facilitar la interpretación de los resultados, entre otras. Para reducir la dimensión, se pueden aplicar métodos como: selección de características (se escoge un subconjunto de las características originales de acuerdo con ciertos criterios), derivación de características (consiste en crear

nuevas características a partir de las originales) y agrupación o clustering de muestras (se agrupan elementos que posean características similares).

- Algoritmos de Aumento de Gradiente (Gradient Boosting Algorithms): estos algoritmos son, en esencia, una colección de algoritmos de aprendizaje que, por sí solos, generan predicciones débiles (poco precisas) pero que al estar combinados producen una única predicción robusta y fuerte (alta precisión). Algunos algoritmos que pertenecen a este grupo son los siguientes: Gradient Boosting Machine (GBM), XGBoost, LightGBM y Catboost.

Además de los algoritmos de Machine Learning previamente mencionados, existen otros que son muy utilizados en las diferentes industrias y que en la industria de los hidrocarburos se les da un uso extenso en aplicaciones de ciencia de datos. Tales algoritmos son los siguientes:

- Redes Neuronales Artificiales (RNA): estos algoritmos muy populares de Machine Learning están inspirados en las redes neuronales biológicas. Están compuestas por unidades de procesamiento conectadas entre sí llamadas neuronas, las cuales pueden ser representadas como una función que posee varias entradas y una salida. Las RNA constituyen una de las técnicas de la ciencia de datos más utilizadas por lo que son tratadas en profundidad en la siguiente sección.
- Regresión de Procesos Gaussianos (Gaussian Process Regression, GPR): Williams y Rasmussen (2006) definen el algoritmo GPR como “una regresión no lineal o técnica de interpolación que modela los nuevos valores estimados basándose en un proceso gaussiano determinado por una función de covarianza” (Como se citó en Shadravan et al., 2015, pág. 4).

- Análisis de componentes principales (Principal Components Analysis, PCA): este algoritmo pertenece al grupo de los algoritmos de reducción de dimensionalidad y consiste de un conjunto de métodos utilizados en tareas de clasificación al intentar realizar una reducción dimensional en los datos. También puede utilizarse para acelerar otros algoritmos al disminuir el tamaño del vector de la muestra. Los principales componentes de un conjunto de datos son las direcciones en las cuales la varianza es mayor y representan, de algún modo, la estructura subyacente (Merayo et al., 2019).
- Regresión de mínimos cuadrados parciales (Partial Least Squares Regression, PLSR): es un algoritmo similar al de PCA, lo que implica que es también un algoritmo de reducción de dimensionalidad. Este algoritmo identifica componentes principales basados tanto en los datos de entrada como en los resultados, con lo cual consigue una reducción de dimensionalidad que está supervisada por los resultados. Los componentes principales son usados para ajustar el modelo de regresión.

  Es conveniente para conjuntos de datos cuyas variables de entrada están altamente relacionadas entre sí. El número de componentes principales se suele elegir mediante validación cruzada (cross-validation). Los datos de entrada y los resultados deben ser estandarizados para hacer comparables las variables (Kassambara, 2017).

2.1.5.2 Redes Neuronales Artificiales (RNA).

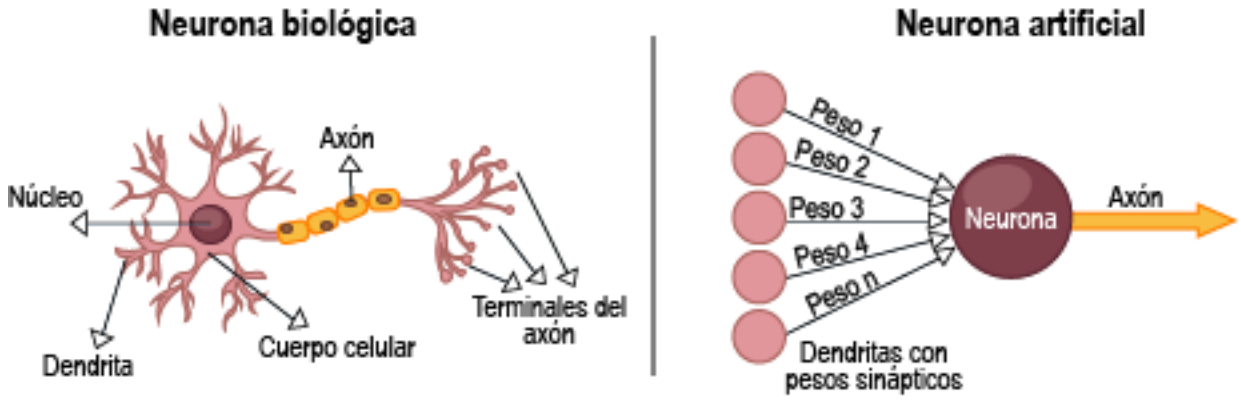
2.1.5.2.1 Definición. Son uno de los algoritmos más populares de Machine Learning y se fundamentan en intentar simular el mecanismo de aprendizaje de los organismos biológicos. El

sistema nervioso humano contiene células, denominadas neuronas, las cuales se conectan entre sí mediante axones y dendritas; la región entre los axones y las dendritas se llama sinapsis. Las fortalezas de las conexiones sinápticas a menudo cambian en respuesta a estímulos externos y gracias a estos cambios es cómo se lleva a cabo el aprendizaje en los organismos vivos.

Este mecanismo biológico se simula en las redes neuronales artificiales, que contienen unidades de cálculo que se denominan "neuronas". Las unidades de cálculo están conectadas unas a otras mediante "pesos", los cuales desempeñan el mismo papel que las fortalezas de las conexiones sinápticas en organismos biológicos. Cada entrada a una neurona se escala con un peso, que afecta la función calculada en esa unidad. La Figura 7 ilustra lo anteriormente mencionado.

Figura 7

Neurona biológica vs Neurona artificial



Nota. La figura relaciona los componentes en una neurona biológica con los componentes de una neurona artificial. Adaptado de (Aggarwal, 2018).

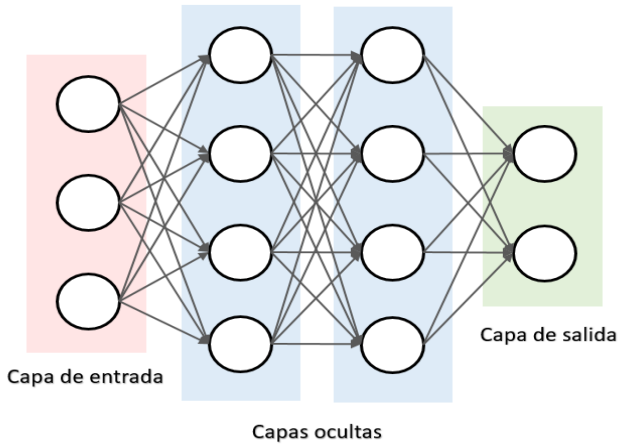
Una red neuronal artificial calcula una función de las entradas al propagar los valores calculados desde las neuronas de entrada a las neuronas de salida y utilizando los pesos como

parámetros intermedios. El aprendizaje ocurre cuando se cambian los pesos que conectan las neuronas. Así como se necesitan estímulos externos para aprender en organismos biológicos, el estímulo externo en redes neuronales artificiales es proporcionado por los datos de entrenamiento que contienen ejemplos de pares de entrada-salida de la función a aprender (Aggarwal, 2018).

2.1.5.2.2 Estructura de una red neuronal artificial. Normalmente, una red neuronal artificial se compone de una capa de entrada, una capa de salida y una o más capas ocultas. Los datos con los que se alimenta a la red neuronal son recibidos en la capa de entrada, donde se detectan las características amplias. Las capas ocultas son las encargadas de analizar y procesar los datos y están compuestas por varias neuronas. De acuerdo con los cálculos realizados previamente, los datos se van simplificando a medida que pasan por cada capa oculta. Finalmente, los resultados o salidas son mostrados en la capa de salida (Theobald, 2017).

Figura 8

Estructura de una red neuronal artificial



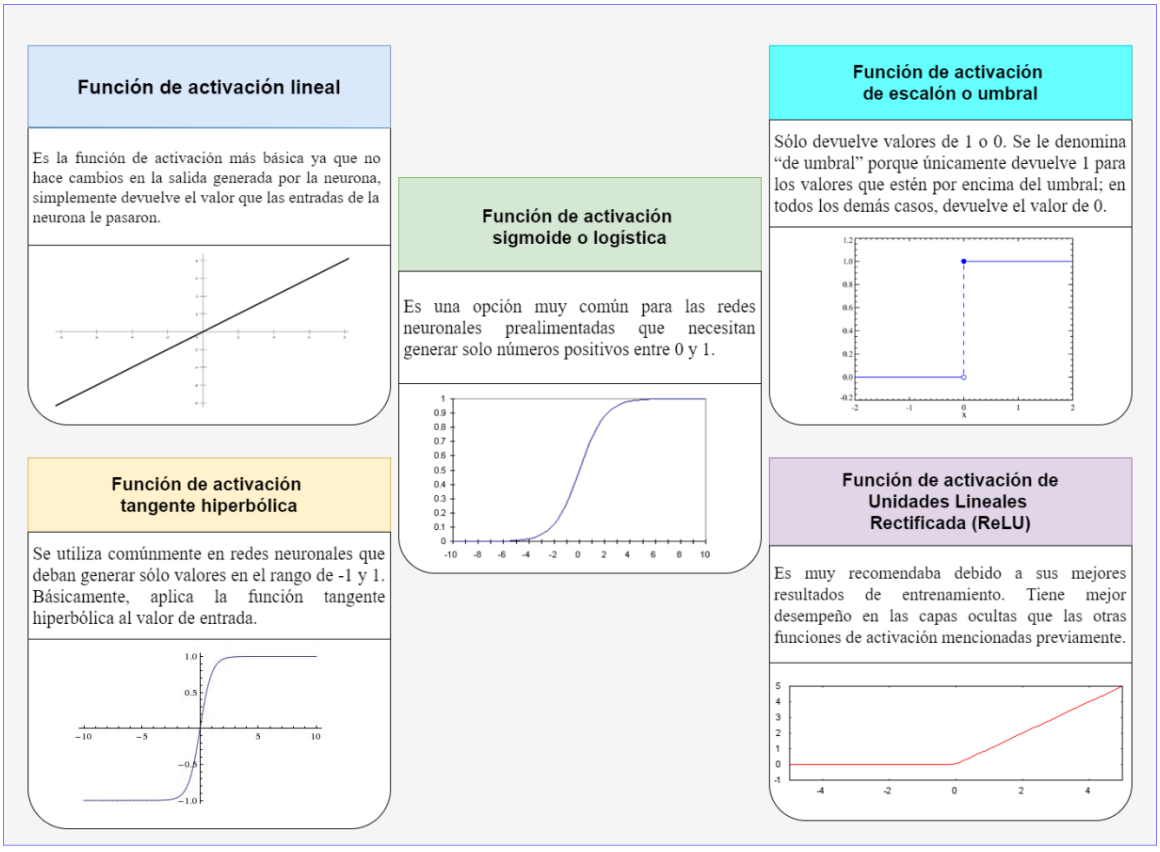
Nota. La figura muestra una red neuronal artificial con una capa de entrada que recibe tres parámetros, dos capas ocultas con 4 neuronas cada una y una capa de salida con 2 resultados. Adaptado de (Theobald, 2017).

2.1.5.2.3 Funciones de activación. Son uno de los componentes más importantes de una red neuronal artificial puesto que se encargan de decidir si se activa o no una neurona. Específicamente, la función de activación tiene el cometido de evaluar la salida de cada neurona y decidir si ese valor de salida debe ser considerado o conectado a las demás neuronas.

En sí, las neuronas no poseen ninguna información sobre qué tan razonable es el resultado que obtuvo tras hacer los cálculos ni cuáles son los rangos aceptables para ese valor. Así, es la función de activación la que proporciona esa información dentro de una red neuronal (Géron, 2019). Existen diversas funciones de activación y, según (Heaton, 2015), las más comunes son las que se muestran en la Figura 9.

Figura 9

Funciones de activación



2.1.5.2.4 Tipos de redes neuronales artificiales. Las redes neuronales pueden ser de muy diversos tipos, sin embargo, y de acuerdo con (Géron, 2019), los tipos más comunes son las redes neuronales prealimentadas o *feed-forward*, las redes neuronales recurrentes (RNR) y las redes neuronales convolucionales (RNC). Cada tipo de RNA posee sus características particulares y su propio nivel de complejidad, lo que la hace aplicable a determinados problemas. La Figura 10 proporciona información más detallada sobre los tipos de RNA.

Figura 10

Tipos de redes neuronales artificiales

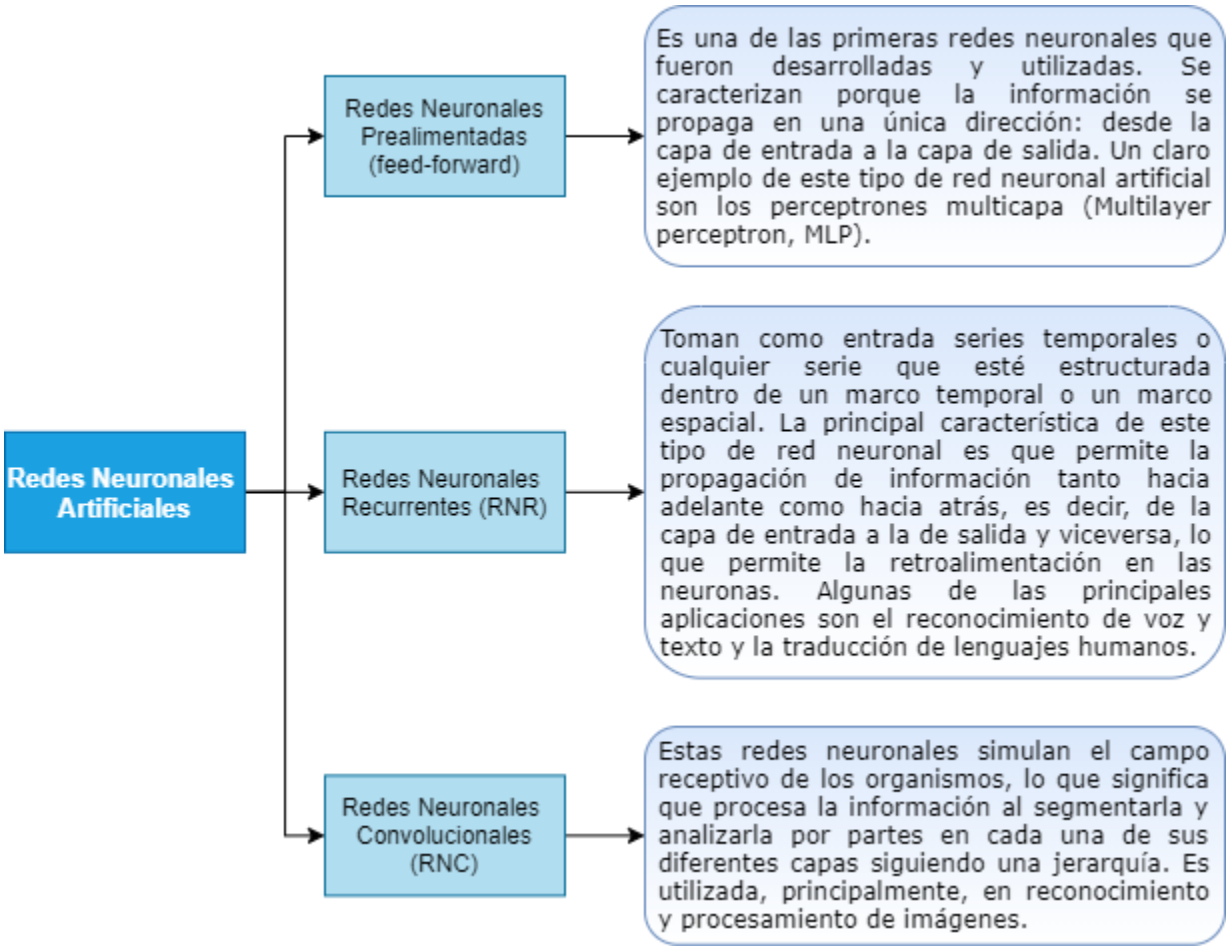
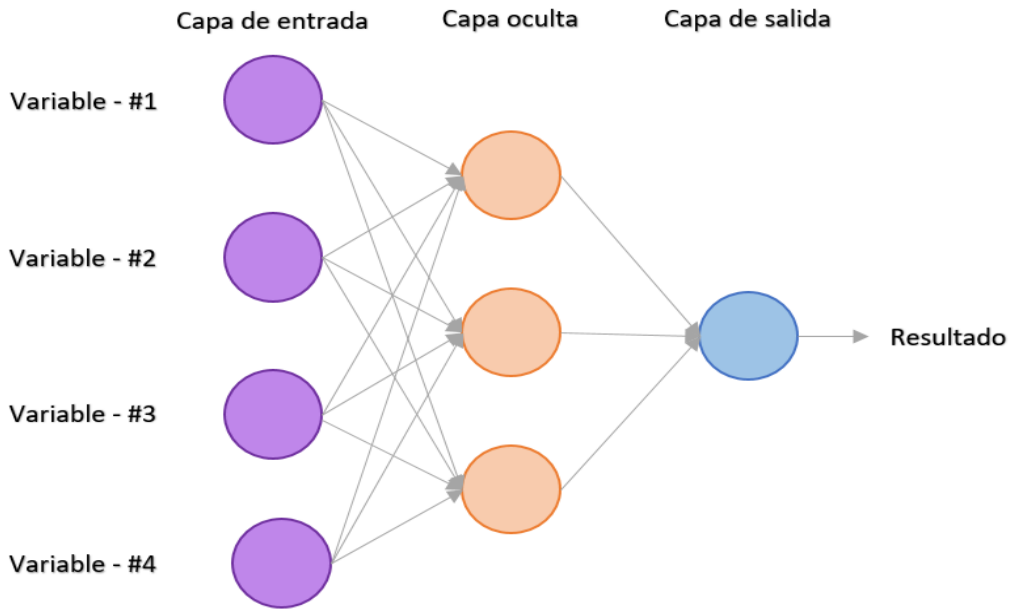


Figura 11

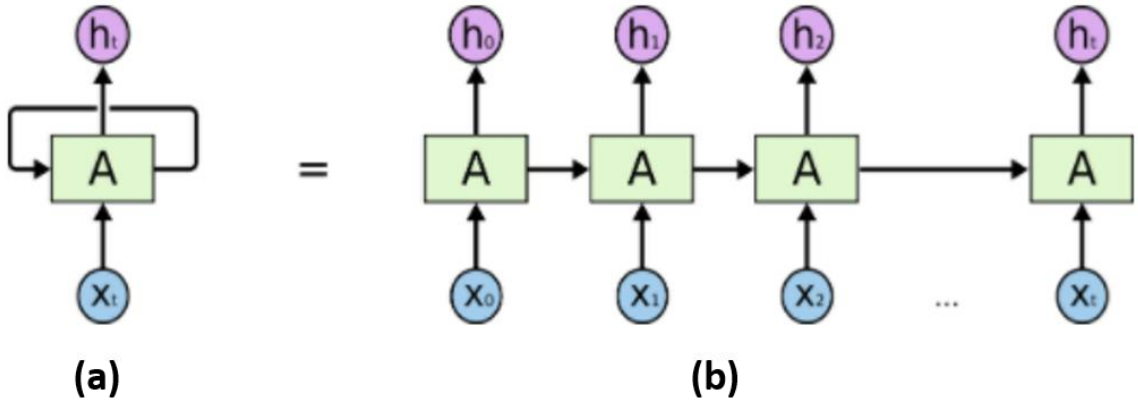
Red Neuronal prealimentada (feed-forward)



Nota. La dirección de las flechas en la figura ilustra el flujo de la información dentro de la red neuronal: desde la capa de entrada hacia la capa de salida. Adaptado de (Müller y Guido, 2016).

Figura 12

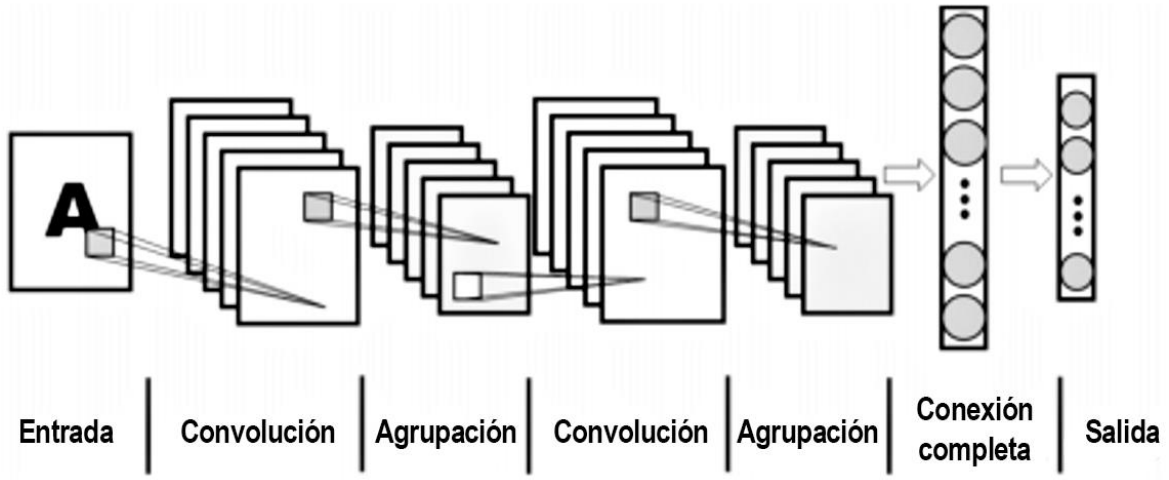
Red Neuronal Recurrente (RNC)



Nota. (a) La figura muestra una red neuronal recurrente en su representación compacta y resalta su naturaleza iterativa. (b) Se ilustra la red neuronal recurrente en su forma expandida para entender mejor su funcionamiento: aplicar las mismas operaciones (representadas por "A") a cada elemento que pertenece a la serie de tiempo (indicados por X_o hasta X_t). Adaptado de (Olah, 2015).

Figura 13

Red Neuronal Convolucional



Nota. La figura muestra las diferentes capas que componen una red neuronal convolucional y expone el principio básico de este tipo de redes neuronales: segmentar la información en partes cada vez más pequeñas y numerosas y procesarla en las diferentes capas ocultas que posee, que son las de convolución y de agrupación. En la capa de conexión completa se recolecta toda la información procesada en las capas anteriores y se le aplica una función de activación para, finalmente, generar el resultado. Adaptado de (Zhan et al., 2020).

2.1.5.2.5 Deep Learning. En español se traduce como aprendizaje profundo y es un subcampo de las redes neuronales artificiales que se fundamenta en las redes neuronales profundas. No existe un consenso claro sobre cuándo una red es considerada profunda ya que, mientras algunos expertos consideran que toda red que tenga más de una capa oculta debe ser considerada como profunda, otros tantos consideran que la red debe tener varias capas para ser considerada como tal (entiéndase por "varias" a más de dos capas) (ActiveWizards, 2019).

El aprendizaje profundo se basa en la cantidad de capas de representaciones (son formas diferentes de representar o codificar la información), que a su vez facilitan las tareas impuestas. Básicamente, este enfoque resuelve problemas al introducir representaciones que se expresan en términos de representaciones más simples, por ejemplo, se puede representar el concepto de

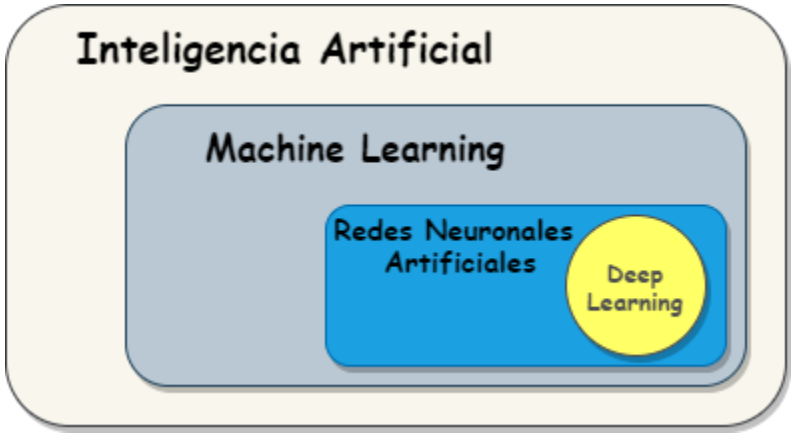
imagen de una persona combinando conceptos más simples, como esquinas y contornos, que a su vez se definen en términos de bordes (Goodfellow et al., 2016).

Las redes neuronales recurrentes (RNR) y las convolucionales (RNC) pertenecen al aprendizaje profundo, ya que poseen múltiples capas ocultas para procesar los datos. Así, algunas de las principales aplicaciones de este subcampo son el reconocimiento de voz y texto (aplicaciones propias de las RNR) y el procesamiento y reconocimiento de imágenes (aplicaciones de las RNC).

La Figura 14 ilustra la posición del aprendizaje profundo dentro del campo de la inteligencia artificial.

Figura 14

Aprendizaje profundo y la inteligencia artificial



Nota. Adaptado de (ActiveWizards, 2019).

2.1.5.3 Minería de datos (Data Mining).

2.1.5.3.1 Definición. La minería de datos o *data mining* puede definirse, en breves palabras, como el proceso de descubrir patrones y conocimientos interesantes a partir de grandes cantidades de datos. Acá, la palabra clave es “descubrir”, ya que la esencia de la minería de datos

es, a partir de datos en bruto, extraer información útil, no trivial y, sobre todo, **nueva**. De hecho, a la minería de datos también se le conoce como *Knowledge Discovery from Data* (KDD), lo que en español se traduce como "descubrimiento de conocimiento a partir de los datos", una expresión que engloba muy bien la esencia de la técnica (Han et al., 2012).

En la práctica, los dos objetivos principales de la minería de datos tienden a ser la predicción y la descripción. La predicción implica el uso de algunas variables o campos en el conjunto de datos para predecir valores desconocidos o futuros de otras variables de interés. La descripción, por otro lado, se enfoca en encontrar patrones que describen los datos que pueden ser interpretados por humanos. Las técnicas de extracción pueden ser clasificadas en cualquiera de estas 2 categorías (Kantardzic, 2020).

2.1.5.3.2 El proceso de la minería de datos. El procedimiento de descubrimiento de conocimientos se muestra como una secuencia iterativa de los siguientes pasos:

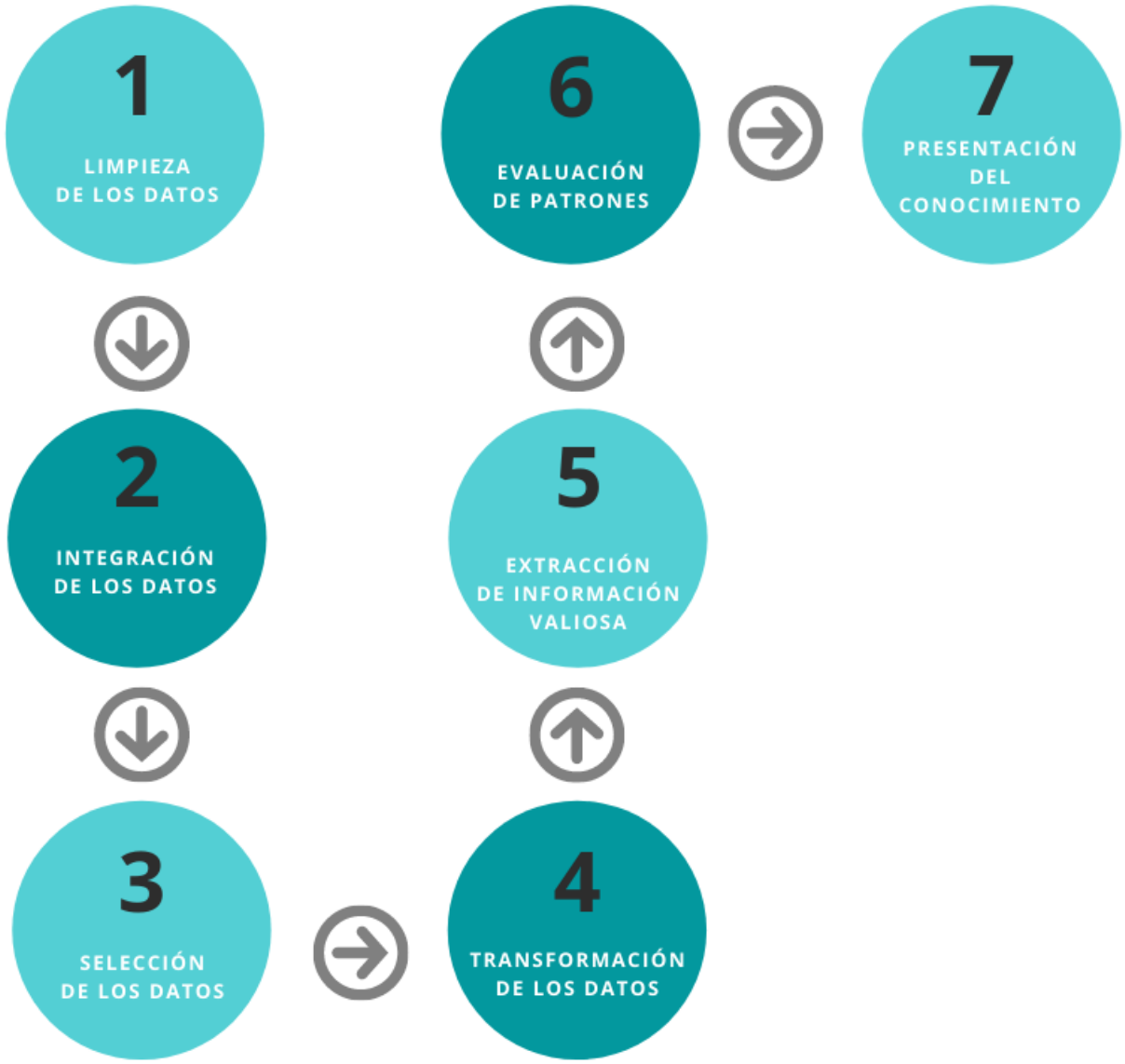
1) Limpieza o purificación de datos: consiste en la remoción de datos redundantes o inconsistentes.
2) Integración de datos: permite la combinación de múltiples fuentes de datos.
3) Selección de datos: los datos relevantes para la tarea de análisis se recuperan de la base de datos.
4) Transformación de datos: donde los datos se transforman y consolidan en formas apropiadas mediante la realización de operaciones de resumen o agregación.
5) Extracción de patrones: un proceso esencial donde se aplican métodos inteligentes para extraer patrones de datos.
6) Evaluación de patrones: se centra en identificar los patrones verdaderamente útiles que representan el conocimiento basado en medidas de interés.

7) Presentación del conocimiento: donde se utilizan técnicas de visualización y representación de la información tratada y extraída que se mostrará a los usuarios.

La Figura 15 ilustra el anterior proceso en forma resumida.

Figura 15

Proceso de minería de datos



2.1.5.3.3 Técnicas de la minería de datos. La minería de datos se apoya en las técnicas o algoritmos que provienen de la inteligencia artificial y la estadística, tales como los árboles de decisión, las redes neuronales, el algoritmo K-means, el algoritmo K-Nearest Neighbors, las regresiones, etc. Tales algoritmos ya fueron definidos en la sección 2.1.5.1.4.

2.1.5.4 Algoritmos genéticos.

2.1.5.4.1 Definición. (AlJuboori et al., 2020, pág. 4) definen los algoritmos genéticos como "un tipo de técnica de optimización que resuelve problemas de optimización restringidos y no restringidos a través del proceso de selección natural basado en el concepto de biología evolutiva, incluidos los procesos fundamentales de selección, cruce y mutación".

Para una mejor y más detallada definición de los algoritmos genéticos es primordial tener claro el concepto de algoritmo evolutivo (AE). Este algoritmo evolutivo se basa en un proceso iterativo y estocástico que opera sobre un conjunto de soluciones (población) enfocadas a un problema determinado. Cada solución (individuo) se obtiene mediante un mecanismo de codificación y decodificación. Inicialmente la población se genera aleatoriamente y a cada individuo se le asigna, mediante una función de bondad o aptitud (fitness), un valor que es una manera de cuantificar la eficacia de dicha solución con respecto al problema en cuestión, además de que el algoritmo lo utiliza como una guía para la búsqueda.

Por su parte, los algoritmos genéticos son un tipo especial de algoritmo evolutivo que no sólo lleva a cabo los procesos mencionados anteriormente, sino que tienen la particularidad de generar nuevas soluciones a partir de la selección, cruce y mutación de los primeros individuos, similar a la creación de nuevos individuos en el campo de la ingeniería genética (Sivanandam y Deepa, 2008).

La mayor utilidad de aplicar algoritmos genéticos es para encontrar respuestas a problemas que tienen espacios de búsqueda realmente grandes generando continuamente soluciones candidatas, evaluando qué tan bien se ajustan las soluciones al resultado deseado y refinando las mejores soluciones. Al resolver un problema con un algoritmo genético, en lugar de pedir una solución específica, se proporcionan características que debe tener la solución o reglas que debe pasar para ser aceptada. Cuantas más restricciones agregue, más soluciones potenciales se bloquearán y mejor será el resultado (Sheppard, 2016).

2.1.4.5.2 Proceso del algoritmo genético. Un proceso de implementación de algoritmo genético tiene los siguientes pasos:

a) Generación de la población inicial: se genera el conjunto de individuos (soluciones), generalmente, en forma aleatoria.
b) Evaluación de los individuos: se aplica la función fitness para evaluar los individuos.
c) Criterio de optimización: se aplica un criterio de optimización que determina si el individuo, es decir, la solución, cumple con las condiciones deseadas. Si cumple, se detiene el proceso; si no, se procede al paso siguiente.
d) Selección de los individuos: los mejores individuos, según el paso anterior, son seleccionados.
e) Reproducción: se cruzan los individuos seleccionados mediante la función de cruce (crossover), lo cual genera nuevos individuos con características ligeramente diferentes basadas en las características de sus padres.
f) Mutación: se realizan pequeños cambios en las características de algunos de los nuevos individuos.

g) Nueva población: los individuos mutados constituyen un nuevo conjunto de soluciones que son, por lo general, mejores que las anteriores.

h) Se procede al paso b.

2.1.5.5 Big Data Analytics.

Este término hace referencia a la aplicación de técnicas, métodos y algoritmos que permitan procesar y analizar volúmenes inmensos de datos (lo que se denomina *big data*) a fin de obtener de ellos patrones, tendencias, correlaciones o cualquier otra información valiosa que ayude a la toma de decisiones o a entender mejor los eventos pasados.

El término clave en esta disciplina es *big data*, el cual se define como grandes colecciones de conjuntos de datos cuyo volumen, velocidad o variedad son tan grandes que es difícil (puede que imposible) almacenar, administrar, procesar y analizar utilizando bases de datos y herramientas de procesamiento de datos tradicionales. Una solución para el tratamiento de esta información es el uso de sistemas informáticos basados en la nube que facilita el almacenamiento, acceso y manejo de dicha cantidad de datos (Bahga y Madisetti, 2019).

Es importante resaltar que el término “análisis” abarca los procesos, tecnologías, marcos y algoritmos para extraer información significativa de los datos. Los datos brutos en sí mismos no tienen significado hasta que se contextualizan y procesan en información útil (Bahga y Madisetti, 2019). Dentro del actual contexto, se encuentran 4 tipos de análisis:

- Análisis descriptivo: comprende el estudio de datos pasados generalmente en forma de reportes o alertas con el objetivo de definir lo que ha acontecido.
- Análisis diagnóstico: al igual que el anterior tipo se centra en el estudio de eventos pasados, sin embargo, su fin es diagnosticar las razones por las cuales se presentaron ciertos sucesos.

- Análisis predictivo: a través del uso de modelos de predicción busca pronosticar o predecir la ocurrencia de un evento o el resultado probable de un evento; también es posible obtener valores futuros de un proceso.
- Análisis prescriptivo: mientras que el anterior tipo se basa en un modelo de predicción para obtener el resultado probable de un evento, esta categoría utiliza múltiples modelos para determinar varios resultados, y a su vez, definir el mejor curso de acción para cada resultado.

La Figura 16 ilustra el proceso de Big Data Analytics, el cual puede resumirse como el tratamiento que se aplica a los grandes volúmenes de datos para obtener de ellos información útil y completa.

Figura 16

Proceso de Big Data Analytics

2.1.5.6 Lógica difusa (Fuzzy Logic, FL).

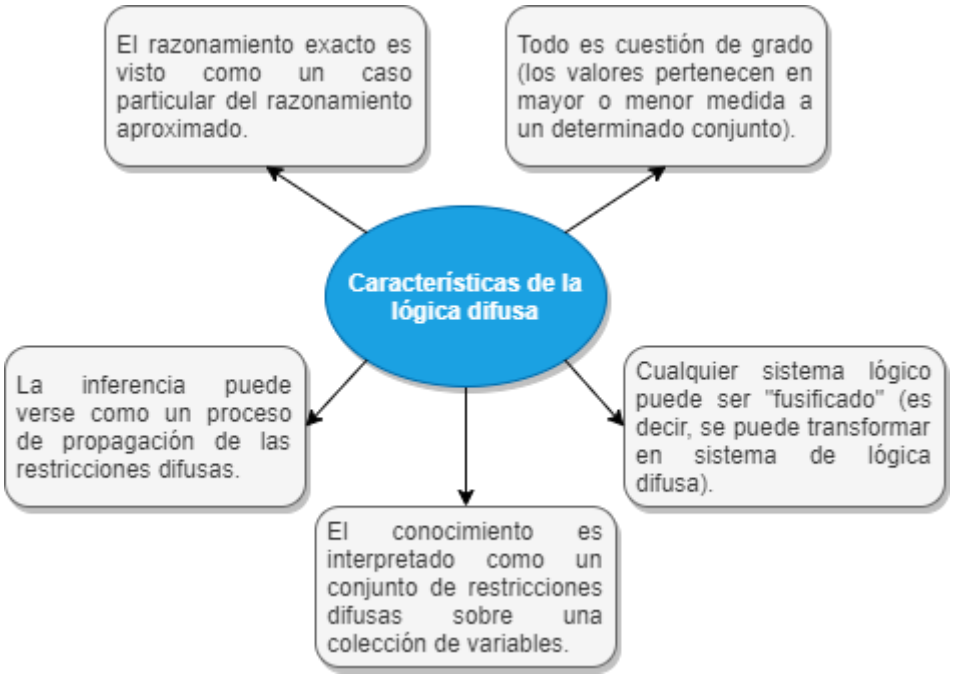
2.1.5.6.1 Definición. (Alavala, 2007) define la lógica difusa como una lógica multivalor que permite definir valores intermedios a las evaluaciones convencionales tales como Verdadero/Falso, Sí/No, Alto/Bajo, 1/0, etc. Gracias a esta lógica es posible utilizar nociones como "bastante grande", "muy rápido", "poco espacio", etc., las cuales serán formuladas matemáticamente y procesadas por los computadores.

Según el mismo autor, la lógica difusa está relacionada con la incertidumbre, la imprecisión y la vaguedad de las variables de entrada y proporciona herramientas formales para tratar con ello.

2.1.5.6.2 Características de la lógica difusa.

Figura 17

Características de la lógica difusa



Nota. Adaptado de (Alavala, 2007)

2.1.5.6.3 Conjuntos difusos. Son aquellos donde un objeto puede tener un grado de pertenencia a un conjunto que oscila entre la pertenencia total (indicada por un valor de 1) y la no pertenencia (indicada por un valor de 0); lo anterior introduce una nueva teoría: la de los conjuntos difusos. Dicha teoría permite que un elemento pertenezca de forma parcial a un conjunto y no de forma absoluta, como en la teoría clásica de conjuntos, lo que implica que una proposición no es completamente cierta ni completamente falsa sino que estará expresada en un grado que va desde 0 hasta 1, donde 0 se le asigna a elementos que están completamente fuera del conjunto y 1 a los que están completamente dentro del conjunto; los valores entre 0 y 1 se asignan a elementos que están parcialmente dentro del conjunto (Pinzón y Martínez, 2019).

2.1.5.6.4 Sistema de Inferencia Adaptativa Neuro-Difusa (Adaptive Neuro-Fuzzy Inference System, ANFIS). Según (Elzenary et al., 2018), ANFIS es un tipo de red neuronal artificial que se basa en el sistema de inferencia difusa de Takagi-Sugeno. El sistema integra los principios de la lógica difusa y de las redes neuronales artificiales por lo que posee los beneficios de ambas técnicas.

Básicamente, lo que hace es aproximar los parámetros de lógica difusa mediante la utilización de redes neuronales artificiales. Su sistema de inferencia corresponde a un conjunto de reglas SI - ENTONCES (IF -THEN) con la capacidad de aprendizaje para aproximar funciones no lineales, lo que implica que ANFIS es considerado como un estimador universal.

2.1.6 Otros conceptos relacionados con la ciencia de datos

2.1.6.1 Internet de las cosas (Internet of Things, IoT). Encapsula una visión de un mundo en el que billones de objetos con inteligencia integrada, medios de comunicación y capacidades de detección y actuación se conectarán a través de redes IP (Protocolo de Internet). Los objetos de la vida diaria, mejorados con microcontroladores, transceptores ópticos y/o de radio, sensores,

actuadores y pilas de protocolos adecuados para la comunicación, se conectarían a internet, permitiendo así la generación de información y, además, la interacción del mundo virtual con el mundo físico (Cirani et al., 2018).

El también llamado "internet de los objetos" se basa en la conexión de todos los objetos electrónicos presentes en un conjunto determinado a través de la internet, para lo cual es necesario no sólo que estos estén adaptados, sino que además posean una forma única de identificación para que puedan ser abordados de manera individual o colectiva, es decir, que posean capacidad de detección y actuación gracias a dicha conexión para trabajar de manera singular o en grupo. Este enfoque promete no sólo iniciar una nueva era de la comunicación, también proyecta un cambio en la cotidianidad de las personas facilitando mucho más sus tareas diarias (Tan y Wang, 2010).

El enfoque de esta tendencia radica en máquinas, sensores o dispositivos que permitan recopilar, procesar y distribuir datos por la red a la que estén conectados con el objetivo de mejorar diferentes procesos a través de una toma de decisiones más precisa donde el usuario no necesariamente debe intervenir de manera directa. Con lo anterior en cuenta, se puede afirmar que el campo de aplicación de esta tecnología es bastante amplio (fabricación, transporte, servicios públicos, etc.) y que su impacto en la economía será notable (Kranz, 2017).

Según (Barrio, 2018), los dispositivos u objetos dentro de una red encaminada a IoT deben presentar los siguientes rasgos o características:

- Comunicación y cooperación: se basa en la conexión de los objetos que no sólo es entre ellos mismos, sino también con los recursos de internet, lo que a su vez permite aprovechar los datos y servicios que hay disponibles para una constante actualización.
- Identificación: es importante que los objetos puedan identificarse de manera única, es decir, que puedan distinguirse como individuos. Lo anterior es con el fin de que dichos objetos

estén relacionados con la información asociada a un criterio o bien concreto y de esa manera los datos puedan ser recuperados de un servidor aprovechando la conexión a una red.

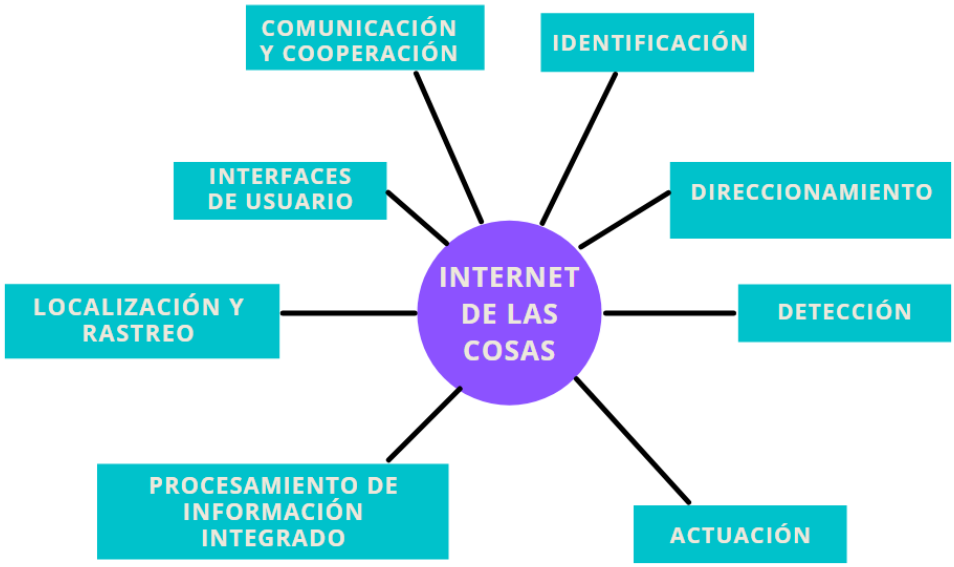
- Direccionamiento: se centra en que los dispositivos puedan ser ubicados y dirigidos de tal manera que se les pueda dar un manejo remoto para extraer información y/o configurarlos.
- Detección: es la capacidad de los objetos de recopilar datos del entorno que los rodea a través de sensores, lo que permite inclusive que estos reaccionen de manera automática.
- Actuación: permite que los dispositivos ejecuten acciones de manera remota gracias a que pueden alterar físicamente su entorno a través de actuadores, por ejemplo, cuando convierten las señales eléctricas en movimiento mecánico.
- Procesamiento de información integrado: este ítem se basa en el procesador del objeto que permite guardar, tratar e interpretar los datos adquiridos además de llevar un registro de las acciones o procesos ejecutados.
- Localización y rastreo: se basa en que los dispositivos puedan ser ubicados o ellos mismos almacenen información de su propia ubicación física; un ejemplo de lo anterior es el GPS (Global Positioning System) en automóviles o teléfonos móviles.
- Interfaces de usuario: es el puente de comunicación entre el usuario y el objeto, que puede ser de manera directa (suponiendo que el dispositivo en cuestión presente un teclado o una pantalla para ese fin) o de manera indirecta (que puede ser a través de un Smartphone o un computador aprovechando la conexión a internet).

El IoT no sólo aporta a la transformación digital del mundo, también ayuda a reducir costos optimizando los procesos en lo que estos estén involucrados. A lo anterior se le agrega la ventaja de la automatización de diferentes tareas que requieren mano de obra, volviéndolas más eficientes

y precisas, lo que se traduce en una reducción de tiempo y una mayor calidad en el producto final (Kranz, 2017).

Figura 18

Esquema de las características de los dispositivos en IoT



Nota. Adaptado de (Barrio, 2018)

2.1.6.2 Sistemas inteligentes. Se definen como sistemas informáticos basados en técnicas no convencionales provenientes de la inteligencia artificial aplicados a máquinas tecnológicamente avanzadas (Pollo Cattaneo et al., 2016). El valor de los sistemas inteligentes radica en que el dispositivo o la máquina tenga la capacidad de recibir información de su entorno para posteriormente procesarla y dar un diagnóstico preciso, en algunos casos toma una decisión respecto a la ejecución de una acción o protocolo frente a una eventualidad.

Para profundizar este concepto se utilizará un ejemplo de aplicación a casas inteligentes, aunque, en general, cualquier edificio de tipo administrativo, industrial o residencial posee un grupo de sistemas diferentes que se encargan de cumplir unas funciones o tareas determinadas. Entre más grande sea el conjunto o volumen de estas actividades, mayor será la complejidad al momento de gestionar dichos edificios, lo que se traduce como un aumento en los costos de mantenimiento, reparaciones y personal encargado de tal fin.

La solución ante la anterior problemática es la de automatizar la mayor cantidad de funciones que se puedan, con el fin de obtener un sistema integrado autónomo de gestión que permita centralizar y proporcionar una interfaz común de acceso y control para todos los equipos, electrodomésticos, dispositivos, sensores o similares, lo que facilita la ejecución de dichas tareas pero, además, permite un ahorro de recursos, una simplificación de la gestión, una reducción en el error humano, respuestas más rápidas y precisas frente a eventos catastróficos (incendios, terremotos, inundaciones, etc.) y, en general, un entorno más seguro y cómodo para los huéspedes. Lo anterior se conoce como casa o edificio inteligente y proporciona al usuario un nuevo tipo de interacción entre la persona y los sistemas de la casa basados en una interfaz fácil de usar que permite manipular toda la red de dispositivos mediante una pequeña cantidad de comandos (Lytvyn et al., 2019).

En los sistemas inteligentes, se destaca como principal característica su capacidad de adaptación al entorno, pero también es necesario que posean 3 facultades básicas que son: razonar (interpretar información para definir sus propias conclusiones y tomar decisiones), aprender (se centra en la adquisición de nuevos datos de manera empírica) e interactuar (ya sea con otros sistemas o con el usuario). Con base en lo anterior, el sistema puede recibir datos y generar respuestas que luego determinará cuáles de estas tuvieron mayor aceptación por parte del usuario

para, posteriormente, guardar las experiencias en su memoria y generalizar, con el fin de crear protocolos frente a diferentes situaciones (Ticona, 2017).

En la Tabla 1 se presentan las diferencias entre un sistema convencional y un sistema inteligente:

Tabla 1

Diferencias entre un sistema inteligente y un sistema convencional

Sistema inteligente	Sistema convencional
Su estructura de funcionamiento es independiente de la base de conocimiento.	Se integra la funcionalidad y la información.
Suele incluir estructuras de explicación de las conclusiones.	No existen estructuras de explicación.
Algunos problemas son resueltos utilizando conocimiento heurístico. Puede presentar errores.	Los problemas son resueltos por algoritmos específicos. No presenta errores.
Métodos declarativos y no determinísticos.	Métodos procedimentales y determinísticos.
Intentan seguir líneas de razonamiento similares a las del ser humano aprendiendo e interactuando con su entorno.	Se centran en la solución y no en la forma en que se obtiene.
Interpretan datos.	Manipulan datos.
Usa heurística y lógica como estrategias de resolución.	Se ejecuta paso a paso mediante procesos predecibles, fiables y exactos.
Puede operar con información incompleta.	Opera con información completa.
Selecciona un lenguaje de representación del conocimiento y lo escribe en dicho lenguaje.	Selecciona un lenguaje de programación y escribe el algoritmo.
Usa las consecuencias del conocimiento para resolver el problema.	Ejecuta el programa.

Nota. Tomado de (Ticona, 2017).

2.1.6.3 Data-driven modeling (DDM). En español se conoce como modelado basado en datos y es una técnica que se basa en el análisis de los datos sobre un sistema, en particular, en la búsqueda de conexiones entre las variables de estado del sistema (variables de entrada, internas y de salida) sin conocimiento explícito del comportamiento físico (Solomatine et al., 2008).

Estos modelos se basan en los métodos de inteligencia computacional y aprendizaje automático y, por lo tanto, implican la presencia de una cantidad considerable de datos que describen la física del sistema modelado (Solomatine y Ostfeld, 2008).

DDM construye modelos que pueden complementar e, incluso, reemplazar los modelos físicos. Los modelos pueden ser generados empleando un algoritmo de Machine Learning que determina la relación entre las entradas y salidas de un sistema utilizando un conjunto de datos que es representativo del comportamiento encontrado en todo el sistema. Posteriormente, el modelo puede ser probado utilizando un conjunto de datos no utilizados con anterioridad para determinar qué tan bien generaliza el modelo los datos no vistos.

Las técnicas más populares en DDM pertenecen a la inteligencia artificial y al aprendizaje automático y son las siguientes: redes neuronales artificiales, sistemas basados en lógica difusa, algoritmos genéticos (para optimización), entre otras.

2.1.6.4 Digital oil and gas fields (DOF).

2.1.6.4.1 Definición. (Carvajal et al., 2018) definen los campos de aceite y gas digitales (DOF) como “un sistema tecnológico que integra la adquisición y transmisión de datos de gran volumen en tiempo real para usar datos en centros de operaciones, sistemas informáticos distribuidos y tecnologías móviles. Desde estos destinos, los datos se reproducen en modelos virtuales y se visualizan en un entorno colaborativo multidisciplinar mediante flujos de trabajo automatizados, comunicación de máquina a máquina, agentes inteligentes y sistemas analíticos

predictivos. Lo anterior permite a la empresa mantener sus operaciones de petróleo y gas en condiciones operativas óptimas y seguras y, en última instancia, maximizar el potencial financiero con una intervención humana mínima".

2.1.6.4.2 Principales componentes y áreas de un sistema DOF

La adopción e implementación de un sistema DOF requiere de tres componentes imprescindibles: los procesos de trabajo, la arquitectura tecnológica y la organización (personas). El conjunto de tales componentes constituye la esencia del sistema DOF.

Por otra parte, un sistema DOF posee 5 áreas fundamentales que deben trabajar completamente sincronizadas para que la adopción del sistema sea exitosa. Según (Carvajal et al., 2018), estas áreas son:

- Instrumentación, sensores remotos y telemetría de procesos en tiempo real: se centra en el equipo y la tecnología de las operaciones de producción de gas y aceite necesarios la telemetría, la recolección remota y la transmisión de los datos que permitirán el monitoreo, la optimización y la automatización de procesos. Los principales elementos de esta área son los sensores, los paneles de control, los cables y los dispositivos de transmisión de datos de manera inalámbrica (routers, por ejemplo).
- Manejo y transmisión de datos: el principal elemento de esta área es la terminal SCADA (Supervisión, Control y Adquisición de Datos) que se encarga de recolectar y procesar la información proveniente de los sensores del campo. Un software denominado "historiador" (*historian*, en inglés) se encarga de limpiar y acondicionar los datos mediante algoritmos que eliminan el ruido, los valores

atípicos y demás anormalidades a fin de garantizar su calidad y luego los organiza y almacena en diferentes capas estructuradas para su posterior uso.

- Automatización del flujo de trabajo: esta área se encarga de realizar de manera automática tareas repetitivas que, de hacerlas manualmente, tomarían demasiado tiempo. Algunas de estas tareas son: recolección de información de diferentes fuentes y filtrado, realización de pruebas de ensayo y error, validación de datos y calibración de un modelo, ejecución de diferentes escenarios de un mismo modelo, entre otras.

  La automatización del flujo de trabajo se hace mediante el uso de rutinas de lenguajes de programación de alto nivel que se aplican sobre los procesos manuales y que pueden, por ejemplo, poblar y actualizar modelos de forma automática.

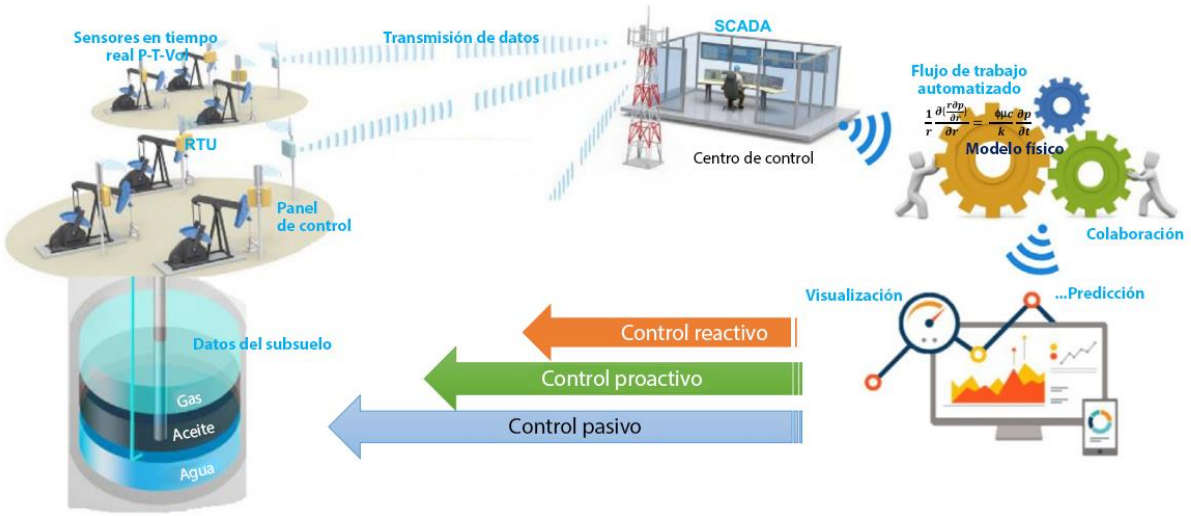
  Por otra parte, la automatización del flujo de trabajo debe poder captar en tiempo real alarmas y alertas para generar acciones inmediatas, realizar monitoreos, diagnósticos y optimizaciones de procesos. Además, a menudo también se requiere que esta automatización posea un carácter predictivo y capacidad de prever futuros problemas operacionales.
- Interfaces de usuario y visualización: muestran la información más relevante de los procesos mediante gráficas, tablas y mapas altamente interactivos, además de ofrecer diagnósticos y análisis relacionados con las acciones operacionales para facilitar la toma de decisiones. También poseen los botones que permiten el control operacional.
- Colaboración y organización de las personas: la comunicación eficiente y la colaboración entre los miembros que componen las diferentes disciplinas en un

proyecto O&G es esencial para la implementación exitosa de un sistema DOF. La colaboración implica estar conectados con el centro de operaciones, en diálogo abierto, contribuyendo a la discusión, cooperando con los demás miembros para encontrar soluciones y comunicando los resultados; en tal aspecto, la tecnología actual facilita la comunicación al permitir videollamadas y videoconferencias entre los miembros del equipo sin importar dónde se encuentren. Por otra parte, la organización busca formar equipos de trabajo eficientes mediante la identificación de habilidades, la definición de roles y la asignación de responsabilidades, tanto de los profesionales como del personal del apoyo.

Las 5 áreas que conforman un sistema DOF y sus elementos más representativos son ilustradas en la Figura 19.

Figura 19

Elementos de un sistema DOF



Nota. La figura ilustra los principales elementos de las 5 áreas que conforman un sistema DOF: adquisición, transmisión y recolección de datos, procesamiento de datos, automatización del flujo de trabajo, colaboración, visualización y generación de modelos de predicción y, finalmente, la elaboración de planes de acción con diferentes modos de control. Adaptado de (Carvajal et al., 2018).

2.1.6.4.3 Ventajas del uso del sistema DOF

- Reduce los riesgos y minimiza la exposición de los operadores y profesionales a accidentes laborales.
- Aumenta la producción: para ello, monitorea continuamente los procesos para asegurar que están funcionando eficientemente; si se presentan fallos en algún proceso, aplica los planes correctivos y/o alerta al personal sobre los inconvenientes. También puede prever fallos y aplicar medidas preventivas antes de que sucedan, así se evitan paradas en la producción.
- Aumenta la eficiencia del proceso de trabajo: ello es conseguido mediante la anticipación de fallos (que evita que el flujo de trabajo se detenga) y también mediante la automatización de los procesos.
- Aumenta la eficiencia del personal del proyecto: la automatización de tareas repetitivas permite al personal centrarse en otras labores más importantes como la solución de fallos o la búsqueda de alternativas para aumentar la productividad. Además, el sistema DOF puede indicarle al personal cuáles son los problemas con mayor prioridad y la manera más eficiente de abordarlos.
- Disminuye costos operacionales: esto se logra mediante diagnósticos preventivos que evitan daños severos en el equipo. Así mismo, el incremento de la eficiencia de trabajo disminuye los costos operacionales de cada proyecto ya que se requiere menos tiempo para llevarlo a cabo.

2.1.6.5 Pozos inteligentes (Smart Wells).

2.1.6.5.1 Definición. Son pozos que utilizan dispositivos mecánicos que permiten controlar la presión y las tasas en el fondo del pozo con el fin de optimizar el rendimiento de la producción

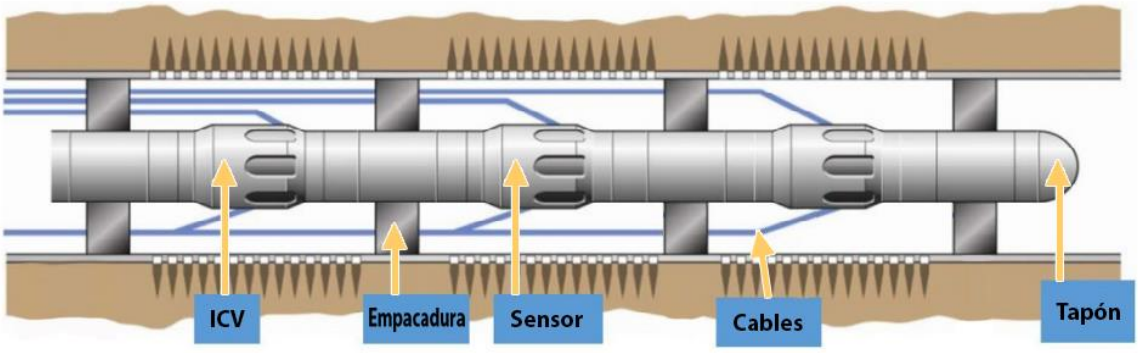
y, en última instancia, mejorar el recobro del yacimiento. Los pozos inteligentes se instalan con dispositivos de fondo de pozo, es decir, son equipos mecánicos y electrónicos que permiten a los operadores controlar los pozos de forma remota, sin intervención, utilizando plataformas o tubería flexible (*coiled tubing*). (Carvajal et al., 2018).

2.1.6.5.2 Principales componentes de un pozo inteligente.

- Dispositivos de control de flujo en fondo de pozo: como su nombre lo indica, son los encargados de regular el flujo en el fondo de pozo y lo hacen de acuerdo con los datos que se reciben en el sistema SCADA provenientes de los sensores instalados en el pozo. Son, en esencia, válvulas con tamaño variable de orificio y controladas remotamente, siendo la válvula ICV (Interval Control Valve) una de las más utilizadas.
- Sensores de fondo de pozo: son dispositivos electrónicos o mecánicos que miden parámetros como presión, temperatura, pH, etc., y envían los valores en forma de señales eléctricas.
- Cables: son los encargados de llevar desde el fondo de pozo hacia superficie los datos medidos por los sensores y también para llevar desde el centro de control hasta el fondo de pozo el pulso que controla las válvulas ICV. Suelen estar integrados en la empacadura (*packer).*
- Empacaduras: conocidas en inglés como *packers*, son utilizadas para aislar entre sí las secciones productoras del pozo perforado; también llevan embebidos los cables de transmisión de datos y de control.

Figura 20

Principales componentes de un pozo inteligente

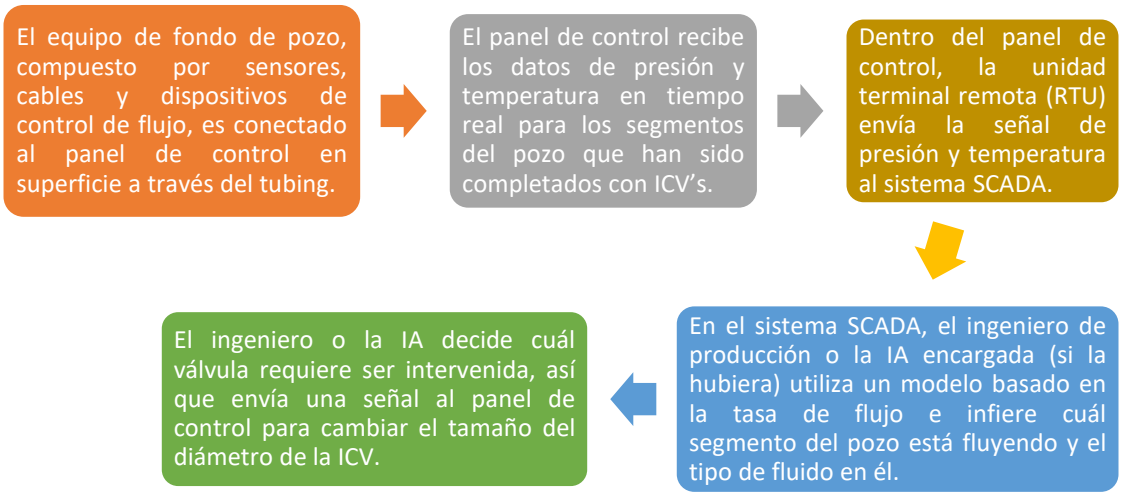


Nota. La figura presenta el esquema de un pozo vertical con completamiento inteligente usando como dispositivo de control de flujo un conjunto de válvulas ICV. Adaptado de (Carvajal et al., 2018).

2.1.6.5.3 Principio de funcionamiento de los pozos inteligentes

Figura 21

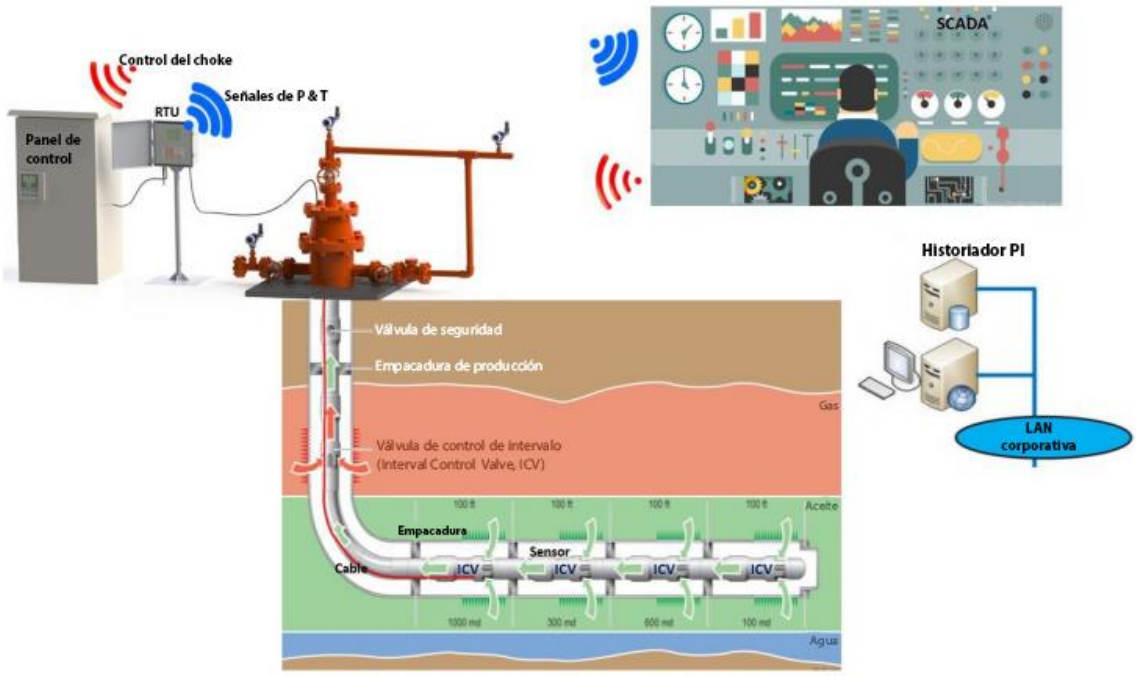
Funcionamiento de los pozos inteligentes



El proceso representado en el diagrama de flujo anterior se ilustra en la Figura 22.

Figura 22

Operación de un pozo inteligente controlado por ICV.



Nota. Adaptado de (Carvajal et al., 2018).

2.1.6.5.4 Aplicaciones de los pozos inteligentes

Tabla 2

Aplicaciones de los pozos inteligentes

Aplicación	Descripción
Control del influjo de gas o agua	Uno de los principales usos de los pozos inteligentes es el control de la invasión temprana de agua o gas en el recobro primario o secundario. Es particularmente útil en pozos horizontales, ya que el agua o el gas se encuentran en la parte superior debido a la diferencia de densidad. El uso de controladores de flujo, como válvulas ICV e ICD, ayudan a reducir la caída de presión o, al menos, la igualan a lo largo de la sección lateral.

Aplicación	Descripción
Control de mezcla de fluidos producidos en pozos verticales	Los pozos inteligentes aumentan la eficiencia de explotación de pozos verticales que producen de múltiples zonas, ya que permiten aislarlas unas de otras mediante las empacaduras y, gracias al uso de los dispositivos de control de flujo, pueden bloquear o controlar el flujo en cualquier zona que esté presentando invasión por gas o agua, con lo cual se mantiene el GOR y/o el corte de agua en los niveles deseados.
Auto-inyección de gas lift	La técnica de auto-inyección de gas lift se puede emplear en pozos que producen de una zona donde la parte superior contiene gas y la parte inferior, aceite. En estos casos, se aíslan las dos zonas y se pone una válvula ICV en la sección superior para controlar el volumen de gas necesario para elevar el aceite desde yacimiento hasta superficie. Cuando el pozo produce de una zona con aceite y agua, puede usarse un arreglo similar, aislando la sección de agua de la sección de aceite. Se utiliza el agua de la sección superior para reinyectarla en la sección de aceite y así energizar el yacimiento.
Optimización EOR/IOR	En este caso, el uso de pozos inteligentes puede contribuir en gran medida al incremento del factor de recobro. La aplicación puede llevarse a cabo en dos formas: por un lado, se pueden instalar válvulas ICV y las respectivas empacaduras para aislar zonas con diferentes permeabilidades, así, se puede disminuir o cerrar el orificio de las ICV para evitar excesiva producción de agua o gas; por otra parte, pueden usarse válvulas ICV en pozos inyectores a fin de mejorar la distribución del agua inyectada y lograr que esta realice el barrido, específicamente, en la zona de interés.
Monitoreo de la producción en tiempo real	Los pozos inteligentes equipados con sensores de presión y de temperatura pueden analizar las pérdidas de presión, minimizar los períodos de cierre del pozo durante pruebas de presión de tipo *build-up* y recolectar instantáneamente datos para cierros de pozo no planeados. Lo anterior contribuye a reducir los tiempos de inactividad en la producción.

3. Metodología para la selección de los procedimientos de diseño de ingeniería

En este capítulo se presenta la metodología aplicada para la selección de los procedimientos de diseño de ingeniería de las asignaturas de la carrera de ingeniería de petróleos de la Universidad Industrial de Santander.

En primer lugar, se presenta una tabla con el contenido todas las asignaturas que componen el ciclo profesional de la carrera de ingeniería de petróleos de la Universidad Industrial de Santander. Posteriormente, se lista un conjunto de procedimientos (de diseño, de cálculo y de medición) que se realizan en la industria petrolera, cuyas fuentes son el "Petroleum Engineering Handbook" de la Sociedad de Ingenieros de Petróleo (SPE) y el "Standard Handbook of Petroleum and Natural Gas Engineering" (Lyons et al., 2015); la tabla con el contenido de las asignaturas presentada al comienzo es cotejada con el conjunto de procedimientos previamente mencionado a fin de extraer los temas que corresponden a procedimientos de diseño de ingeniería (y relacionados). Finalmente, a dichos procedimientos se les aplican ciertos criterios de selección para obtener únicamente aquellos procedimientos de ingeniería a los cuales se les puede aplicar alguna técnica de la ciencia de datos.

3.1 Contenido de las asignaturas del ciclo profesional de ingeniería de petróleos

Para empezar, es fundamental tener en cuenta que, de acuerdo con (Escuela de Ingeniería de Petróleos - UIS, 2015), el ciclo profesional de la ingeniería de petróleos de la UIS está compuesto por 4 áreas fundamentales, a saber: área de operaciones, área de procesos, área de yacimientos y área de asignaturas de administración y complementarias.

Para el desarrollo de este proyecto de investigación se decidió unir las áreas de procesos y de operaciones en una única área denominada "de operaciones", debido a que las asignaturas que

componen cada una de las áreas mencionadas están fuertemente relacionadas entre sí y al estar en conjunto permiten entender el proceso que posibilita la explotación de los yacimientos de hidrocarburos; en otras palabras, estas asignaturas son las que describen todas las operaciones y procesos que se efectúan para llevar los hidrocarburos descubiertos desde el yacimiento hasta la superficie.

De acuerdo con lo anteriormente establecido, las áreas que componen el ciclo profesional de la ingeniería de petróleos y las respectivas asignaturas que forman parte de cada una de las áreas son las siguientes:

- Área de yacimientos: análisis petrofísicos, propiedades de los fluidos, evaluación de formaciones, ingeniería de yacimientos, laboratorio de fluidos, análisis de presiones, simulación de yacimientos, métodos de recobro.
- Área de operaciones: lodos de perforación, perforación de pozos, completamiento de pozos, métodos de producción, ingeniería del gas, facilidades de superficie.
- Área de asignaturas de administración y complementarias: ingeniería económica, evaluación de proyectos, política petrolera, responsabilidad social, gestión integral de la industria petrolera, electivas profesionales (corrosión y su control, estructura y propiedades de los materiales, geomecánica y su aplicación a la industria del petróleo, operaciones costa afuera, simulación avanzada de yacimientos, medición de hidrocarburos).

En la Tabla 3 se presenta el contenido de las asignaturas del ciclo profesional que pertenecen al pénsum de la carrera de ingeniería de petróleos de la Universidad Industrial de Santander. La información está basada en el documento denominado “Proyecto Educativo de Reforma Académica del Programa Ingeniería de Petróleos” (Escuela de Ingeniería de Petróleos -

UIS, 2015) y complementada con datos del plan de estudios que se encuentra disponible en la página de la escuela de ingeniería de petróleos[††].

Tabla 3

Contenido de las asignaturas del ciclo profesional de ingeniería de petróleos

Asignatura [área]	Contenido
Análisis petrofísicos [yacimientos]	Descripción litológica Análisis granulométrico Porosidad Permeabilidad Fabricación de tapones sintéticos Saturación Propiedades eléctricas de la roca Permeabilidad relativa Mojabilidad y tensión interfacial Presión capilar
Lodos de perforación [operaciones]	Fluidos de perforación Propiedades físicas del lodo de perforación Química de las arcillas Reología de los fluidos de perforación Propiedades de filtración Contaminantes y tratamiento Lodos salados Lodos dispersos y no dispersos Lodos base calcio Lodos base aceite Prueba de azul de metileno y resistividad del lodo Propiedades de una lechada de cemento
Ingeniería económica [administración y complementarias]	Conceptos básicos Interés Equivalencias Anualidades Gradiente

[††] Más información en: http://petroleos.uis.edu.co/eisi/, en la ruta /Pregrado/Ingeniería de Petróleos/Escuela de Ingeniería de Petróleos/1.3 Plan de estudios.

Asignatura [área]	Contenido
	Inflación y devaluación Punto de equilibrio Evaluación económica y financiera
Propiedades de los fluidos [yacimientos]	Comportamiento de fases Fluidos del yacimiento Propiedades de los fluidos (aceite, agua y gas) Fundamentos de equilibrio de fases Depositación de orgánicos
Evaluación de formaciones [yacimientos]	Fundamentos de petrofísica Anatomía de un registro Registros litológicos en hueco abierto Registros de porosidad en hueco abierto Registros resistivos en hueco abierto Registro de resonancia magnética nuclear Evaluación de formaciones en pozos entubados Correlación de registros para su interpretación integral Evaluación de cemento Evaluación de corrosión del revestimiento
Perforación de pozos [operaciones]	Descripción del equipo de perforación Cálculos adicionales sobre el equipo Hidráulica y diseño Técnicas de detección de presiones anormales Drill Stem Test Estudio de brocas Corazonamiento Gradiente de presión y de fractura Pruebas de Leak Off Test Problemas en pozos durante la perforación Control de pozo Perforación direccional y horizontal Perforación costa afuera Coil tubing Instrumentación, sistemas de control y medición

Asignatura [área]	Contenido
Laboratorio de fluidos [yacimientos]	Toma de muestras de fluidos Salinidad, BSW & °API Viscosidad Punto de fuego y relampagueo Pruebas de botella Punto de fluidez Presión de vapor REID Análisis cromatográfico Análisis y tratamiento de aguas residuales Análisis PVT
Evaluación de proyectos [administración y complementarias]	Gestión de proyectos Evaluación financiera Evaluación bajo incertidumbre y riesgo Evaluación económica y social tasa social de descuento Evaluación ambiental Herramienta OILPROJ-G
Política petrolera [administración y complementarias]	Política petrolera mundial Política petrolera nacional Fiscalización petrolera Entidades gubernamentales y no gubernamentales involucradas en el sector de hidrocarburos Legislación petrolera
Responsabilidad social [administración y complementarias]	Responsabilidad social individual Responsabilidad social empresarial en el sector de los hidrocarburos
Completamiento de pozos [operaciones]	Descripción de herramientas Manejo de manuales y tablas de elementos tubulares Diseño de revestimiento Operaciones de cementación Análisis de la calidad de la cementación Operaciones de cañoneo Fluidos de completamiento Configuración de los completamientos Árboles de navidad Arenamiento de los pozos de petróleo Tratamiento de acidificación

Asignatura [área]	Contenido
	Operaciones de fracturamiento Otras técnicas de estimulación y limpieza de pozos
Ingeniería de yacimientos [yacimientos]	Yacimientos de hidrocarburos Yacimiento de gas Yacimientos de gas condensado Yacimientos de aceite subsaturados Yacimientos de aceite con empujes simultáneos Intrusión de agua Curvas de declinación Manejo y uso de la información Caracterización de yacimientos
Gestión integral de la industria petrolera [administración y complementarias]	Seguridad industrial Salud ocupacional Medio ambiente Norma NTC ISO 9001:2008 Normas líderes nacionales e internacionales en HSE
Análisis de presiones [yacimientos]	Fundamentos de flujo en medios porosos Pruebas de ascenso de presión (buildup) Pruebas de descenso de presión (drawdown) Curvas tipo Pruebas en pozos de gas Otras pruebas Diseño y herramientas de prueba Pruebas de presión en yacimientos naturalmente fracturados Síntesis directa de Tiab Manejo de software e interpretación de pruebas de presión (Fast Welltest, PanSystem)
Métodos de producción [operaciones]	Curvas de declinación de la producción Comportamiento de entrada de los fluidos al pozo (IPR) Flujo multifásico en tuberías Análisis nodal Bombeo mecánico

Asignatura [área]	Contenido
	Sistema de levantamiento artificial por bombeo neumático (Gas lift) Cavidades progresivas Bombeo hidráulico tipo pistón y tipo jet Bombeo electrosumergible
Ingeniería del gas [operaciones]	Generalidades Equipos de proceso en la industria del gas Plantas de tratamiento de gas Plantas de procesamiento de gas Sistemas de transporte de gas
Simulación de yacimientos [yacimientos]	Antecedentes Información necesaria para utilizar un simulador Clasificación de los simuladores Principios básicos y ecuaciones fundamentales Modelo numérico usando diferencias finitas Solución de sistema de ecuaciones Modelos en diferencias finitas Aspectos prácticos de simulación de yacimientos Simulador de aceite negro
Facilidades de superficie [operaciones]	Problemas de pozo Depósitos de parafinas y asfáltenos Limpieza de la arena Generalidades de recolección y tratamiento de aceite y gas Proceso de separación Tratamiento de sistemas hidrocarburos Tanques Bombas Tratamiento del agua producida
Métodos de recobro [yacimientos]	Factores que influyen en un proyecto de recuperación mejorada Heterogeneidad en yacimientos Eficiencias de desplazamiento Generalidades método de recobro mejorado de aceite

Asignatura [área]	Contenido
	Facilidades de superficie Patrones de inyección Mecanismos de desplazamiento con fluidos inmiscibles Relación tasas inyección vs. Presiones de inyección Métodos de predicción del comportamiento de la inyección de agua en los yacimientos Tratamiento de aguas de inyección y de producción Problemas asociados a la inyección de agua Fuentes de agua de inyección Casos históricos – pruebas en pozos de inyección Inyección de vapor
Corrosión y su control [administración y complementarias]	Introducción Termodinámica y potencial de electrodo Cinética electroquímica en corrosión Pasivación de aleaciones Métodos para medir la velocidad de corrosión Tipos de corrosión Corrosión marina y en suelos Protección catódica
Estructura y propiedades de los materiales [administración y complementarias]	Introducción a la ingeniería de los materiales Estructura atómica y enlaces Estructura cristalina y amorfa en los materiales Propiedades mecánicas de metales Aleaciones más usadas en ingeniería de petróleos Integridad de materiales en la industria de los hidrocarburos Corrosión de materiales
Geomecánica y sus aplicaciones a la industria del petróleo [administración y complementarias]	Generalidades de la geomecánica Esfuerzo Deformación

Asignatura [área]	Contenido
	Relaciones constitutivas y constantes elásticas Resistencia de las rocas y criterios de falla Presión de poro Compactación y subsidencia Estabilidad de pozos Fracturamiento hidráulico y producción de arena
Operaciones costa afuera [administración y complementarias]	Introducción operaciones offshore Introducción operaciones costa afuera Perforación en costa afuera Producción en costa afuera Economía petrolera
Simulación avanzada de yacimientos [administración y complementarias]	History matching Simulación de yacimientos cerca al punto crítico Principios fundamentales para la simulación de procesos de recobro Datos necesarios para la simulación de procesos de recobro Simulación de procesos químicos Simulación de procesos térmicos de recuperación mejorada Gerenciamiento y predicción del comportamiento del yacimiento
Medición de hidrocarburos [administración y complementarias]	Conceptos básicos, regulaciones y estándares Medición estática Medición dinámica de líquido Medición dinámica de gas Análisis de laboratorio Balances y conciliaciones

3.2 Procedimientos de diseño de ingeniería

Se entiende por *procedimiento de diseño*, en ingeniería de petróleos, a aquel procedimiento que permite efectuar cálculos numéricos y no numéricos a fin de obtener resultados con variables

que son de importancia para cumplir con un diseño en la mencionada ingeniería; tales diseños son fundamentales para cumplir el objetivo del campo de la ingeniería de petróleos donde aplican. Acá, la palabra *diseño* hace referencia a la actividad de proyectar el aspecto, características, funcionalidad y posibles resultados de un objeto, sistema o proceso.

Cabe aclararse que en el presente proyecto de investigación se tendrán en cuenta no sólo procedimientos de diseño de ingeniería sino también aquellos procedimientos de cálculo, de análisis y/o de medición que estén fuertemente relacionados con los procedimientos de diseño, toda vez que su ejecución es indispensable para realizar un determinado diseño; por ejemplo, la determinación de las propiedades de los fluidos presentes en un yacimiento no es un procedimiento de diseño sino un procedimiento de medición, sin embargo, tal procedimiento es fundamental para el diseño de las facilidades de superficie.

3.2.1 Procedimientos de ingeniería más importantes en la industria petrolera

A continuación, se presentan tres tablas que contienen procedimientos de cálculo, análisis, medición y diseño que se llevan a cabo en el sector upstream (exploración y producción) de la industria petrolera. Los procedimientos son agrupados por las áreas definidas en la Sección 3.1 del presente documento.

Los procedimientos fueron extraídos de dos fuentes confiables y que constituyen un referente dentro de la industria O&G, como lo son el “Petroleum Engineering Handbook” de la SPE (Lake, 2007) y el “Standard Handbook of Petroleum and Natural Gas Engineering” (Lyons et al., 2015). Se hace la aclaración de que la lista de procedimientos presentados no es exhaustiva, sino que tiene como propósito exponer aquellos procedimientos que son más frecuentes y/o importantes y que, por tanto, tienen mayor impacto en la industria del petróleo.

Tabla 4

Procedimientos de ingeniería en el área de yacimientos

Área de yacimientos	
Procedimiento	**¿Presente en el plan de estudios? ***
Análisis de núcleos.	Sí
Análisis de presión usando curvas tipo.	Sí
Análisis de pruebas de pozo: *buildup, drawdown, falloff*, multitasa, interferencia.	Sí
Análisis del movimiento de fluidos en inyección de agua.	Sí
Análisis SARA del crudo.	No
Cálculo de la eficiencia de barrido (de desplazamiento, volumétrica, de área o de patrón, vertical o de invasión).	Sí
Cálculo de la eficiencia de recobro.	No
Cálculo de la inyectividad e índice de inyectividad.	Sí
Cálculo del almacenamiento del pozo.	Sí
Cálculo del factor de daño (skin).	Sí
Cálculo del factor de recobro.	Sí
Cálculo/estimación de las propiedades de la roca (porosidad, volumen de poro, permeabilidad, permeabilidad relativa, capacidad, transmisibilidad, resistividad y conductividad eléctrica, compresibilidad).	Sí
Cálculo/medición de la presión capilar.	Sí
Cálculos en yacimientos naturalmente fracturados: intensidad de fractura, apertura de la fractura, porosidad de la fractura, permeabilidad de la fractura, longitud de la fractura.	No
Cálculos para la inyección de agua.	Sí
Cálculos para la inyección de vapor.	Sí
Determinación de heterogeneidades en yacimientos.	Sí
Corazonamiento (extracción de núcleos).	Sí
Determinación de la tensión superficial e interfacial.	Sí
Determinación o cálculo de las propiedades de los fluidos del yacimiento (agua, aceite, gas): densidad, viscosidad, compresibilidad, factor de formación.	Sí
Diseño de los patrones de inyección.	Sí
Diseño del espaciamiento de pozos.	Sí

Área de yacimientos	
Procedimiento	**¿Presente en el plan de estudios? ***
Diseños de laboratorio para recobro mejorado.	No
Estimación de hidrocarburos en sitio mediante la ecuación de balance de materiales para diferentes tipos de yacimientos.	Sí
Estimación de reservas en fuentes no convencionales.	No
Estimación del aceite residual (tras inyección de agua): balance de materiales, pruebas de pozo, análisis de núcleos.	No
Estimación del recobro en inyección de agua mediante balance de materiales.	No
Estimación/medición directa de la saturación de fluidos.	Sí
Exploración geofísica	No
Generación de curvas de permeabilidad relativa.	Sí
Interpretación de registros de pozo en hueco abierto: potencial espontáneo (SP), laterolog, microrresistivos, gamma ray, sónicos, densidad (density), neutrón (neutron), resonancia magnética nuclear (NMR), registro de buzamiento (*dipmeter).*	Sí
Interpretación de registros de pozo en hueco revestido: registro de neutrones pulsados, registro de espectroscopía natural gamma, registro de adherencia del cemento y registro de densidad variable (CBL-VDL), registro *caliper*, registros electromagnéticos de inspección, registros de potencial eléctrico, registros de producción.	Sí
Interpretación sísmica.	No
Medición directa de las propiedades de la roca.	Sí
Medición directa de las propiedades de los fluidos del yacimiento.	Sí
Método volumétrico para la estimación del aceite y del gas original en sitio (OOIP, OGIP).	Sí
Métodos de recobro en yacimientos naturalmente fracturados	No
Métodos de recobro mejorado.	Sí
Modelamiento numérico de yacimientos.	Sí
Predicción del recobro primario en yacimientos con empuje por agua.	Sí
Predicción del recobro primario en yacimientos con empuje por gas en solución.	Sí
Preservación de núcleos.	Sí
Pruebas de laboratorio para medición de las propiedades de los fluidos.	Sí
Selección de la técnica de recobro mejorado.	Sí
Simulación de procesos de recobro mejorado	Sí

Área de yacimientos	
Procedimiento	**¿Presente en el plan de estudios? ***
Selección de yacimientos donde es aplicable el recobro mejorado.	Sí
Simulación dinámica de yacimientos.	No
Toma de muestras de fluidos del yacimiento.	No

Nota. * Esta columna indica si el procedimiento en cuestión se encuentra dentro del contenido de alguna de las asignaturas del programa de Ingeniería de Petróleos de la Universidad Industrial de Santander.

Tabla 5

Procedimientos de ingeniería en el área de operaciones

Área de operaciones	
Procedimiento	**¿Presente en el plan de estudios? ***
Análisis de curvas de declinación.	Sí
Análisis de curvas IPR.	Sí
Análisis de dinagramas.	
Análisis del rendimiento de las arcillas en los fluidos de perforación.	Sí
Análisis nodal.	Sí
Balanceo de tapones de lechada de cemento.	No
Cálculo de la trayectoria del pozo en perforación direccional.	Sí
Cálculo del desempeño de la unidad de bombeo.	Sí
Cálculo del posicionamiento del pozo (coordenadas geográficas, sistema UTM, sistema de coordenadas geodésicas).	Sí
Cálculo del volumen de lechada requerido.	Sí
Cálculos de hidráulica de brocas.	Sí
Cálculos de hidráulica de perforación y completamiento (regímenes de flujo, pérdidas de presión, etc.).	Sí
Cálculos para coiled tubing.	No
Cálculos y procedimientos para despegar la sarta.	Sí
Completamientos no convencionales: pozos horizontales con fracturamiento hidráulico.	No
Criterios de diseño de casing: colapso, estallido, tensión, compresión.	Sí
Detección de patadas de pozo.	Sí

Área de operaciones	
Procedimiento	**¿Presente en el plan de estudios? ***
Determinación de la presión y temperatura a la cual se forman los hidratos.	Sí
Determinación de las pérdidas de filtrado.	Sí
Diseño de “árboles de navidad”.	Sí
Determinación de las propiedades de la lechada.	Sí
Diseño de acidificación de matriz (operaciones de remediación).	Sí
Diseño de BHA para control direccional.	Sí
Diseño de bombas (reciprocantes, giratorias y centrífugas).	Sí
Diseño de brocas.	No
Diseño de completamientos inteligentes (Smart Wells).	No
Diseño de compresores (centrífugos, reciprocantes, etc.).	Sí
Diseño de deshidratadores de gas con desecantes sólidos.	Sí
Diseño de deshidratadores de gas con glicol.	Sí
Diseño de equipos de separación de sólidos.	No
Diseño de estabilizadores.	No
Diseño de estaciones de compresión.	Sí
Diseño de intercambiadores de calor.	Sí
Diseño de la sarta de perforación: tamaños de drill collar y drill pipe, selección de conectores, selección de grados de la tubería, etc.	Sí
Diseño de la torre de perforación (estructura, capacidad de carga, material de fabricación, etc.).	Sí
Diseño de lechada para cementación remedial (*squeeze*).	No
Diseño de levantamiento artificial con bombeo hidráulico.	Sí
Diseño de levantamiento artificial con gas lift.	Sí
Diseño de líneas de flujo y tuberías de recolección.	Sí
Diseño de oleoductos.	Sí
Diseño de plantas de endulzamiento con aminas.	Sí
Diseño de rimadores (reamers).	No
Diseño de separadores (verticales, horizontales, esféricos).	Sí
Diseño de tanques de almacenamiento.	Sí
Diseño de tratadores.	Sí

Área de operaciones	
Procedimiento	**¿Presente en el plan de estudios? ***
Diseño de tubing o tubería de producción: selección de grado y espesor de pared; aplicación de los criterios de diseño (colapso, estallido, tensión, corrosión)	Sí
Diseño del bombeo mecánico.	Sí
Diseño del proceso de fracturamiento: formulación del fluido de fracturamiento, selección de propantes, análisis de propagación de fracturas, etc.	Sí
Diseño del programa de casing o revestimiento: selección del tamaño del casing y de las respectivas brocas, establecimiento de las profundidades de asentamiento del casing, selección del grado del casing, etc.	Sí
Diseño del sistema de izamiento/elevación.	Sí
Diseño del sistema de levantamiento artificial con bombas de cavidades progresivas (PCP).	Sí
Diseño del sistema de levantamiento artificial con bombas electrosumergibles (ESP).	Sí
Diseño del sistema de levantamiento artificial plunger lift.	No
Diseño del sistema de recolección de gas.	Sí
Diseño del sistema eléctrico del campo.	No
Diseño e instalación de teas.	No
Diseño y selección de cabezal de pozo.	No
Diseño y selección de guaya.	Sí
Elaboración de mud report y mud record.	Sí
Estimación del índice de productividad.	Sí
Evaluación del tratamiento de acidificación.	Sí
Evaluación económica de brocas (costo por pie perforado).	Sí
Evaluación de la corrosión en el revestimiento.	Sí
Formulación de la lechada de cemento: cantidad de sacos de cemento por volumen de agua; masa de aditivos, etc.	Sí
Evaluación de las propiedades reológicas del fluido de perforación.	Sí
Formulación del fluido de completamiento y workover.	Sí
Formulación del fluido de perforación.	Sí
Formulación/diseño del fluido espaciador.	Sí
Fracturamiento con ácido.	Sí

Área de operaciones	
Procedimiento	**¿Presente en el plan de estudios? ***
Fracturamiento en pozos horizontales y desviados.	No
Instalación de sistemas de seguridad.	No
Limpieza de tuberías (pigging).	No
Mantenimiento de equipos.	No
Medición del contenido de agua en el gas.	Sí
Medición del flujo de gas.	Sí
Métodos de control de arena.	Sí
Operación de cañoneo.	Sí
Operaciones de “pesca”.	Sí
Operaciones de cementación de casing multietapa.	No
Operaciones de cementación de casings de gran diámetro.	No
Operaciones de completamiento (limpieza del hueco, desplazamiento del fluido de completamiento, completamiento superior e inferior)	Sí
Operaciones de estimulación de pozos (fracturamiento hidráulico).	Sí
Predicción de IPR: Vogel, Fetkovich, Standing (modificado), Fetkovich-Vogel, método unificado.	Sí
Predicción de la formación de hidratos.	Sí
Prevención de la formación de hidratos.	Sí
Prevención y control en caso de pérdidas de circulación.	Sí
Procedimiento de cementación primaria.	Sí
Procedimiento de cementación secundaria.	Sí
Procedimiento de la cementación tipo *squeeze*.	No
Procedimientos de control de pozo (método del ingeniero, método del perforador, método volumétrico).	Sí
Proceso de inhibición de la formación de hidratos.	Sí
Pruebas al fluido de perforación (peso, viscosidad, esfuerzo de gel, filtración, contenido líquidos y sólidos, prueba química, etc.).	Sí
Selección de brocas.	Sí
Selección de martillos de perforación (drilling jars).	Sí
Selección del equipo.	Sí
Selección del fluido de completamiento.	Sí
Selección del fluido de perforación.	Sí
Selección del procedimiento de control de pozo.	Sí

Área de operaciones	
Procedimiento	**¿Presente en el plan de estudios? ***
Selección e instalación de empacaduras (packers).	Sí
Tratamiento de emulsiones.	Sí

Nota. * Esta columna indica si el procedimiento en cuestión se encuentra dentro del contenido de alguna de las asignaturas del programa de Ingeniería de Petróleos de la Universidad Industrial de Santander.

Tabla 6

Procedimientos de ingeniería en el área de administración y complementarias

Área de administración y complementarias	
Procedimiento	**¿Presente en el plan de estudios? ***
Análisis de riesgo (cualitativo y cuantitativo).	Sí
Análisis de valor presente neto (VPN).	Sí
Árboles de decisión.	No
Cálculo de esfuerzos.	Sí
Caracterización de residuos.	Sí
Consideraciones ambientales para las operaciones de perforación: selección del lugar, identificación y comprensión de la regulación gubernamental aplicable, facilidad para obtener los recursos requeridos para la operación.	Sí
Construcción del modelo geomecánico.	Sí
Corrosión en operaciones de producción.	Sí
Corrosión en operaciones offshore.	Sí
Detección de *breakouts* y fracturas por tensión.	Sí
Determinación de la orientación de los esfuerzos.	Sí
Determinación del límite económico de un proyecto.	Sí
Determinación del retorno de la inversión (ROI) y de la tasa interna de retorno (TIR).	Sí
Diseño de BOP para perforación offshore.	No
Diseño de plataformas.	No
Diseño e instalación de facilidades offshore.	No
Estimación de la cantidad de hidrocarburos producibles.	Sí
Estimación de producción futura.	Sí

Área de administración y complementarias	
Procedimiento	**¿Presente en el plan de estudios? ***
Evaluación económica de proyectos.	Sí
Manejo de aguas de producción.	Sí
Manejo de residuos (métodos).	Sí
Obtención de permisos de explotación.	No
Predicción de la corrosión.	Sí
Predicción de la estabilidad del pozo.	Sí
Prevención de la contaminación del agua.	Sí
Prevención y control de derrames.	Sí
Producción en offshore.	No
Programa de casing offshore.	No
Pronóstico de precio de venta.	Sí
Remediación del sitio de producción.	No
Selección del método de control de corrosión.	Sí
Selección de facilidades para operaciones offshore.	Sí
Simulación Montecarlo.	No
Técnicas de mitigación de corrosión.	Sí
Transporte de hidrocarburos en offshore.	No
Tratamiento de residuos (lodos utilizados y aguas de producción) en offshore.	No

Nota. * Esta columna indica si el procedimiento en cuestión se encuentra dentro del contenido de alguna de las asignaturas del programa de Ingeniería de Petróleos de la Universidad Industrial de Santander.

3.2.2 Procedimientos de ingeniería presentes en el plan de estudios

Como resultado de cotejar el contenido de cada una de las diferentes asignaturas del plan de estudios del programa de ingeniería de petróleos de la UIS, el cual se encuentra en la Tabla 3, con los procedimientos de ingeniería presentados en las Tablas 4, 5 y 6, se extrajo aquellos temas que corresponden a procedimientos de ingeniería, ya sean de cálculo, de análisis, de medición o de diseño, que son tratados durante el desarrollo de las mencionadas asignaturas.

En las Tablas 7, 8 y 9 se presentan los temas que corresponden a procedimientos de ingeniería y las respectivas asignaturas a las cuales pertenecen dichos temas.

Tabla 7

Procedimientos de ingeniería en el área de yacimientos (plan de estudios)

Asignatura	Procedimiento
Análisis petrofísicos	– Descripción litológica (composición, textura y densidad de la roca) – Determinación de la porosidad – Determinación de la permeabilidad – Determinación de la saturación – Determinación de la permeabilidad relativa – Determinación de la mojabilidad y la tensión interfacial – Determinación de la presión capilar
Propiedades de los fluidos	– Cálculo de las propiedades de los fluidos del yacimiento (agua, aceite y gas)
Evaluación de formaciones	– Interpretación de registros en hueco abierto – Interpretación de registros en hueco entubado – Evaluación de corrosión del revestimiento
Ingeniería de yacimientos	– Estimación de reservas en yacimientos de gas mediante método volumétrico y ecuación de balance de materiales (EBM) – Estimación de reservas en yacimientos de gas condensado mediante método volumétrico y ecuación de balance de materiales (EBM) – Estimación de reservas en yacimientos de aceite subsaturados mediante método volumétrico y ecuación de balance de materiales (EBM) – Estimación de reservas en yacimientos de aceite con empujes simultáneos mediante ecuación de balance de materiales (EBM) – Estudio de la intrusión de agua en los yacimientos – Caracterización de yacimientos
Laboratorio de fluidos	– Toma de muestras de los fluidos del yacimiento

Asignatura	Procedimiento
	– Pruebas de laboratorio a los fluidos del yacimiento (salinidad, BS&W, °API, viscosidad, punto de fluidez, etc.) – Análisis cromatográfico – Tratamiento de emulsiones (prueba de botella) – Análisis y tratamiento de aguas residuales – Análisis PVT
Análisis de presiones	– Análisis de pruebas de ascenso de presión (buildup) – Análisis de pruebas de descenso de presión (drawdown) – Análisis e interpretación de curvas tipo – Análisis de pruebas de presión en pozos de gas – Análisis de pruebas de presión en pozos con fracturas – Análisis de pruebas de interferencia, multitasa y de pulso
Simulación de yacimientos	– Simulación numérica de yacimientos – Interpretación de los resultados de la simulación
Métodos de recobro	– Determinación de la heterogeneidad del yacimiento – Determinación de la eficiencia de desplazamiento – Patrones de inyección – Estudio del desplazamiento con fluidos inmiscibles – Predicción del comportamiento de la inyección de agua en los yacimientos – Tratamiento de aguas de inyección y de producción – Cálculo de la inyectividad e índice de inyectividad. – Cálculos para la inyección de agua – Cálculos para la inyección de vapor

Tabla 8

Procedimientos de ingeniería en el área de operaciones (plan de estudios)

Asignatura	Procedimiento
Lodos de perforación	– Formulación del fluido de perforación (lodo) – Pruebas al lodo (peso, viscosidad, esfuerzo de gel, filtración, contenido líquidos y sólidos, prueba química, etc.) – Determinación del rendimiento de arcillas – Evaluación de la reología del lodo – Tratamiento de contaminación del lodo

Asignatura	Procedimiento
	– Determinación de las pérdidas de filtrado
Perforación de pozos	– Cálculos sobre el equipo de perforación (torre, cables, bombas, motores, etc.) – Cálculos de hidráulica de perforación – Diseño de la sarta de perforación – Detección de presiones anormales – Detección de patadas de pozo – Estudio de brocas (diseño, hidráulica, costo por pie perforado, selección, bit record, etc.) – Métodos de control de pozo – Cálculos de perforación direccional – Soluciones a problemas de pozo (pegas, reventones, pérdidas de circulación, desviaciones)
Completamiento de pozos	– Diseño de revestimiento o casing – Diseño de tubería de producción o tubing – Procedimientos de cementación – Diseño de la lechada de cemento – Diseño de la operación de cañoneo – Formulación del fluido de completamiento – Configuración de los completamientos (simples, a hueco abierto, duales, triples, etc.) – Estudio y control del arenamiento – Diseño de árboles de navidad – Diseño del tratamiento de acidificación – Diseño de la estimulación por fracturamiento – Control de parafinas y asfaltenos
Métodos de producción	– Análisis de curvas de declinación de la producción – Comportamiento de entrada de los fluidos al pozo (IPR): determinación del índice de productividad, predicción del IPR, análisis de curvas IPR. – Estudio del flujo multifásico en tuberías – Análisis nodal – Diseño del sistema de bombeo mecánico – Diseño de instalaciones de gas lift – Diseño del sistema de levantamiento artificial por bombas de cavidades progresivas (PCP) – Diseño del sistema de bombeo hidráulico

Asignatura	Procedimiento
Ingeniería del gas	– Diseño de separadores – Diseño de intercambiadores de calor – Diseño de compresores – Diseño de plantas de endulzamiento de gas (con aminas, con lecho, con membranas) – Diseño de plantas de deshidratación de gas (con glicol, con desecantes sólidos) – Fraccionamiento de gas – Licuefacción de gas natural – Diseño de gasoductos – Medición de gas
Facilidades de superficie	– Prevención y control de la depositación de asfaltenos y parafinas – Diseño de estaciones de recolección de hidrocarburos – Diseño y selección de separadores (bifásicos y trifásicos) – Diseño y selección de tratadores (térmicos y termoelectrostáticos) – Tratamiento de emulsiones – Diseño y selección de tanques de almacenamiento – Diseño de bombas – Tratamiento del agua de producción – Diseño de oleoductos

Tabla 9

Procedimientos de ingeniería en el área de administración y complementarias (plan de estudios)

Asignatura	Procedimiento
Ingeniería económica	– Cálculo de los diferentes tipos de interés (simple, compuesto, nominal, efectivo, múltiple, continuo) – Cálculo de equivalencias – Cálculo de anualidades – Cálculo de gradientes – Cálculo de la inflación y la devaluación – Evaluación económica de proyectos
Evaluación de proyectos	– Evaluación financiera de proyectos

Asignatura	Procedimiento
	– Gestión de proyectos – Evaluación y análisis de riesgos – Evaluación económica y social – Evaluación ambiental
Gestión integral de la industria petrolera	– Diagnóstico de condiciones de trabajo o panorama de riesgos – Análisis de trabajo seguro – Plan de manejo ambiental – Plan de contingencia – Disposición y manejo de residuos – Manejo de derrames hidrocarburos
Corrosión y su control	– Métodos para control de corrosión: protección catódica – Medición de la corrosión
Estructura y propiedades de los materiales	– Diseño y selección de materiales – Técnicas de evaluación de la integridad de los materiales – Estudio de la corrosión de los materiales (en tuberías de transporte y distribución, en casing y tubing, en tanques de almacenamiento, en el proceso de refinación)
Geomecánica y su aplicación a la industria del petróleo	– Determinación/cálculo de los esfuerzos – Aplicación de los criterios de falla – Predicción de la presión de poro – Análisis de estabilidad de pozo – Determinación de la orientación de los esfuerzos – Construcción del modelo geomecánico de estabilidad del pozo – Construcción e interpretación del círculo de Mohr
Operaciones costa afuera	– Well planning en costa afuera – Selección de facilidades en costa afuera
Simulación avanzada de yacimientos	– Simulación de procesos químicos en recobro mejorado – Simulación de procesos térmicos en recobro mejorado – Predicción del comportamiento de un yacimiento
Medición de hidrocarburos	– Medición estática de líquido – Medición dinámica de líquido – Medición dinámica de gas

3.3 Selección de los procedimientos de diseño de ingeniería

Debido a que los procedimientos de ingeniería sobre los cuales se pueden emplear técnicas de la ciencia de datos son muy variados, precisamente a causa de la extensa aplicabilidad que tiene la ciencia de datos, se hace necesario escoger un limitado número de ellos para ser presentados por cada una de las áreas del ciclo profesional de la ingeniería de petróleos.

La selección de los procedimientos se lleva a cabo atendiendo a las siguientes razones: que existan estudios previos sobre la aplicación de alguna técnica de la ciencia de datos al procedimiento en cuestión, es decir, estudios publicados en la literatura científica de la industria petrolera, particularmente, en libros o revistas; que sean procedimientos representativos del área a la que pertenecen (yacimientos, operaciones, complementarias); y, finalmente, que la aplicación de la técnica de la ciencia de datos signifique una ventaja en términos de calidad de resultados y/o reducción del tiempo de análisis o ejecución, con respecto a la manera convencional.

Los procedimientos que cumplen con los criterios antes mencionados se muestran en la Tabla 10. En dicha tabla, es fundamental tener en cuenta lo siguiente:

a. Cuando la técnica corresponde a Machine Learning (abreviado como ML), se consideran todos los algoritmos de ML, excepto las redes neuronales artificiales.
b. Cuando la técnica corresponde a las redes neuronales artificiales (abreviado como RNA), se incluyen únicamente las RNA convencionales, especialmente, el perceptrón multicapa, lo que implica que se excluyen las redes neuronales profundas.
c. Cuando la técnica corresponda al aprendizaje profundo o Deep Learning, esta hará mención tanto a las redes neuronales recurrentes como a las redes neuronales convolucionales.
d. A menudo, se abrevia el nombre de la técnica empleada y, seguidamente, entre paréntesis, se especifica el algoritmo o método utilizado. Por ejemplo, ML (DT) indica que la técnica es Machine Learning y que se usa, específicamente, el algoritmo de árbol de decisión (Decision Tree, DT).

Tabla 10

Procedimientos de diseño de ingeniería seleccionados

Procedimiento	Área [Asignatura]	Técnica(s) aplicable(s)	Fuente(s)
Diseño y selección de materiales	Administración y complementarias [Estructura y propiedades de los materiales]	RNA, ML (DT, K-means, PCA)	(Merayo et al., 2019)
Evaluación económica y social de proyectos	Administración y complementarias [Evaluación de proyectos]	Data Mining	(Bravo et al., 2020)
Estimación de propiedades geomecánicas	Administración y complementarias [Geomecánica y sus aplicaciones en la industria del petróleo]	Data Mining, RNA	(Parapuram et al., 2017)
Análisis de estabilidad de pozo	Administración y complementarias [Geomecánica y sus aplicaciones en la industria del petróleo]	RNA, ML (SVM)	(Okpo et al., 2016), (Lin et al., 2018)
Criterios de falla	Administración y complementarias [Geomecánica y sus aplicaciones en la industria del petróleo]	ML (ANFIS, SVM, SVR), RNA	(Alloush et al., 2017), (Negara et al., 2018)
Análisis de trabajo seguro	Administración y complementarias [Gestión integral de la industria petrolera]	Big Data Analytics	(Loretti et al., 2019), (Tarrahi y Shadravan, 2016)
Well planning offshore	Administración y complementarias [Operaciones costa afuera]	ML (clustering), NLP	(Cumming et al., 2020)
Simulación de procesos químicos	Administración y complementarias [Simulación avanzada de yacimientos]	RNA	(Dang et al., 2018)
Simulación de procesos de recobro mejorado	Administración y complementarias	ML	(Sierra et al., 2020)

Procedimiento	Área [Asignatura]	Técnica(s) aplicable(s)	Fuente(s)
	[Simulación avanzada de yacimientos]		
Diseño del completamiento de pozos	Operaciones [Completamiento de pozos]	ML (LR, RF, GB), RNA	(Pankaj et al., 2018), (Shelley et al., 2021), (Porras et al., 2020)
Operaciones de fracturamiento hidráulico	Operaciones [Completamiento de pozos]	RNA	(He et al., 2019), (Rezaei et al., 2020), (Pankaj et al., 2018), (Shelley et al., 2021), (Porras et al., 2020)
Operaciones de cementación	Operaciones [Completamiento de pozos]	RNA, ML (GPR)	(Shadravan et al., 2015)
Diseño de plantas de endulzamiento de gas	Operaciones [Ingeniería de gas]	Big Data Analytics	(Cadei et al., 2019)
Diseño del fluido de perforación	Operaciones [Lodos de perforación]	RNA, ML (GPR)	(Shadravan et al., 2017)
Determinación de las pérdidas de circulación.	Operaciones [Lodos de perforación]	ML (PLSR), RNA, Big Data Analytics, Naive Bayes	(Al-Hameedi, Alkinani, Dunn-Norman, Al-Alwani, et al., 2019), (Hou et al., 2020), (Duarte et al., 2018)
Estimación de la producción de un pozo	Operaciones [Métodos de producción]	Data Mining, RNA, Deep Learning	(Rahmanifard et al., 2020), (Gupta et al., 2021), (H. Khan y Louis, 2021)
Diseño del equipo de bombeo mecánico	Operaciones [Métodos de producción]	ML (DT), RNA	(Bangert, 2019)
Análisis del comportamiento de afluencia (IPR)	Operaciones [Métodos de producción]	RNA	(Basfar et al., 2018)
Diseño de instalaciones de gas lift	Operaciones [Métodos de producción]	GA	(AlJuboori et al., 2020)

Procedimiento	Área [Asignatura]	Técnica(s) aplicable(s)	Fuente(s)
Cálculo del índice de productividad	Operaciones [Métodos de producción]	RNA, FL	(Alarifi et al., 2015)
Análisis de curvas de declinación	Operaciones [Métodos de producción]	RNA, ML (PCA, RF, GB), Deep Learning	(Li y Han, 2017), (Liang y Zhao, 2019), (Gupta et al., 2021), (H. Khan y Louis, 2021), (Masini et al., 2019)
Selección del sistema de levantamiento artificial	Operaciones [Métodos de producción]	ML (DT), RNA	(Ounsakul et al., 2019)
Problemas de pozos y su control	Operaciones [Perforación de pozos]	Data Mining (DT, KNN), RNA, ML (PCA, DT, SVM, RF)	(Alouhali et al., 2018), (Hou et al., 2019), (Alshaikh et al., 2019), (Shi et al., 2019)
Diseño en perforación direccional	Operaciones [Perforación de pozos]	Deep Learning	(Pollock et al., 2018)
Evaluación de las propiedades reológicas del lodo	Operaciones [Perforación de pozos]	Data Mining	(Al-Hameedi, Alkinani, Dunn-Norman, Flori, et al., 2019)
Cálculos de hidráulica de perforación (ECD)	Operaciones [Perforación de pozos]	RNA, ANFIS	(Elzenary et al., 2018)
Estudio de brocas: Selección de broca	Operaciones [Perforación de pozos]	RNA + GA, Big Data Analytics	(Abbas et al., 2019), (Tortrakul et al., 2021), (Cornel & Vazquez, 2020)
Diseño de sarta: determinación y/u optimización de ROP	Operaciones [Perforación de pozos]	RNA, ML (RF, SVR, GB), Deep Learning	(Al-AbdulJabbar et al., 2018), (Noshi, 2019), (Han et al., 2019)
Determinación de la porosidad	Yacimientos [Análisis petrofísicos]	RNA	(Al-Abduijabbar et al., 2020)
Determinación de la permeabilidad	Yacimientos [Análisis petrofísicos]	RNA, ML (SVM), FL	(Gholami et al., 2014)

Procedimiento	Área [Asignatura]	Técnica(s) aplicable(s)	Fuente(s)
Determinación de la saturación de agua	Yacimientos [Análisis petrofísicos]	RNA, ML (SVM), ANFIS	(Sambo et al., 2018), (Anifowose et al., 2019), (Khan et al., 2018)
Interpretación de registros de pozo	Yacimientos [Evaluación de formaciones]	RNA (MLP), Deep Learning	(Gupta y Soumya, 2020), (Belozerov et al., 2018)
Caracterización de yacimientos	Yacimientos [Ingeniería de yacimientos]	Data Mining, RNA, ML (LogR, RF, DT, GB, KNN, K-means, SVM), GA, Deep Learning	(Tavares et al., 2019), (Aming, 2021), (Baldini et al., 2020), (Ojukwu et al., 2020), (Mohamed et al., 2019), (Sun y Belhaj, 2019), (Miller et al., 2019)
Cálculo del factor de recobro	Yacimientos [Ingeniería de yacimientos]	RNA, ML (ANFIS-2, SVM)	(Noureldien y El-Banbi, 2015), (Ahmed et al., 2017)
Eficiencia de la inyección de agua	Yacimientos [Métodos de recobro]	ML (PSO)	(Chen et al., 2021)
Determinación de la heterogeneidad del yacimiento	Yacimientos [Métodos de recobro]	ML (clustering)	(Konoshonkin et al., 2020)
Cálculo de las propiedades de los fluidos del yacimiento	Yacimientos [Propiedades de los fluidos]	RNA	(Al-Amoudi et al., 2019), (Liu et al., 2021)
Simulación numérica de yacimientos	Yacimientos [Simulación de yacimientos]	Deep Learning, ML	(Masoudi et al., 2020)

Nota. En la tabla, las abreviaciones que aparecen en la columna "Técnica aplicable" son las siguientes:
ML: Machine Learning; RNA: Redes Neuronales Artificiales; PLSR (Partial Least Squares Regression): Regresión de mínimos cuadrados parciales; DT (Decision Tree): Algoritmo de árbol de decisión; KNN: Algoritmo de K-Nearest Neighbors; FL(Fuzzy Logic): Algoritmo de lógica difusa; SVM (Support Vector Machine): Algoritmo de máquinas de vectores de soporte; SVR (Support Vector Regression): Algoritmo de regresión de vectores de soporte; PCA (Principal Component Analysis): Algoritmo de Análisis de Componentes Principales; GA (Genetic Algorithms): Algoritmos genéticos; ANFIS (Adaptive Neuro-Fuzzy Inference System): Sistema de inferencia adaptativa neuro-difusa; GPR (Gaussian Process Regression): Algoritmo de Regresión de Procesos Gaussianos; RF (Random Forest): Algoritmo de bosque aleatorio; LogR (Logistic Regression): Algoritmo de regresión logística; LR (Linear Regression): Algoritmo de regresión lineal; GB (Gradient Boosting): Algoritmo de aumento del gradiente; NLP

(Natural Language Processing): Técnicas de procesamiento de lenguaje natural; PSO (Particle Swarm Optimization): Algoritmo de optimización por enjambre de partículas.

3.4 Ejemplos propuestos

En esta sección se proponen algunos ejemplos por cada una de las áreas que componen el ciclo profesional del programa de ingeniería de petróleos ofrecido en la Universidad Industrial de Santander. Los ejemplos fueron propuestos teniendo en cuenta las siguientes consideraciones:

1. Que sean procedimientos con abundantes estudios relacionados con la aplicación de la ciencia de datos sobre los mismos. Para ello, se tendrá en consideración los análisis bibliométricos presentados en las Figuras 22 y 23.
2. Que sean procedimientos que se encuentren dentro del ciclo profesional del plan de estudios de la ingeniería de petróleos de la UIS y que, adicionalmente, sean procedimientos representativos de cada área del ciclo profesional (Tablas 7, 8 y 9).
3. Idealmente, procedimientos que normalmente se lleven a cabo mediante el uso de correlaciones ya existentes y probadas o mediante software comercial, a fin de poder comparar resultados obtenidos mediante la aplicación de las técnicas convencionales (las correlaciones o el software comercial) y las técnicas que en este estudio se analizan (técnicas de la ciencia de datos).

Figura 23

Análisis bibliométrico de OnePetro

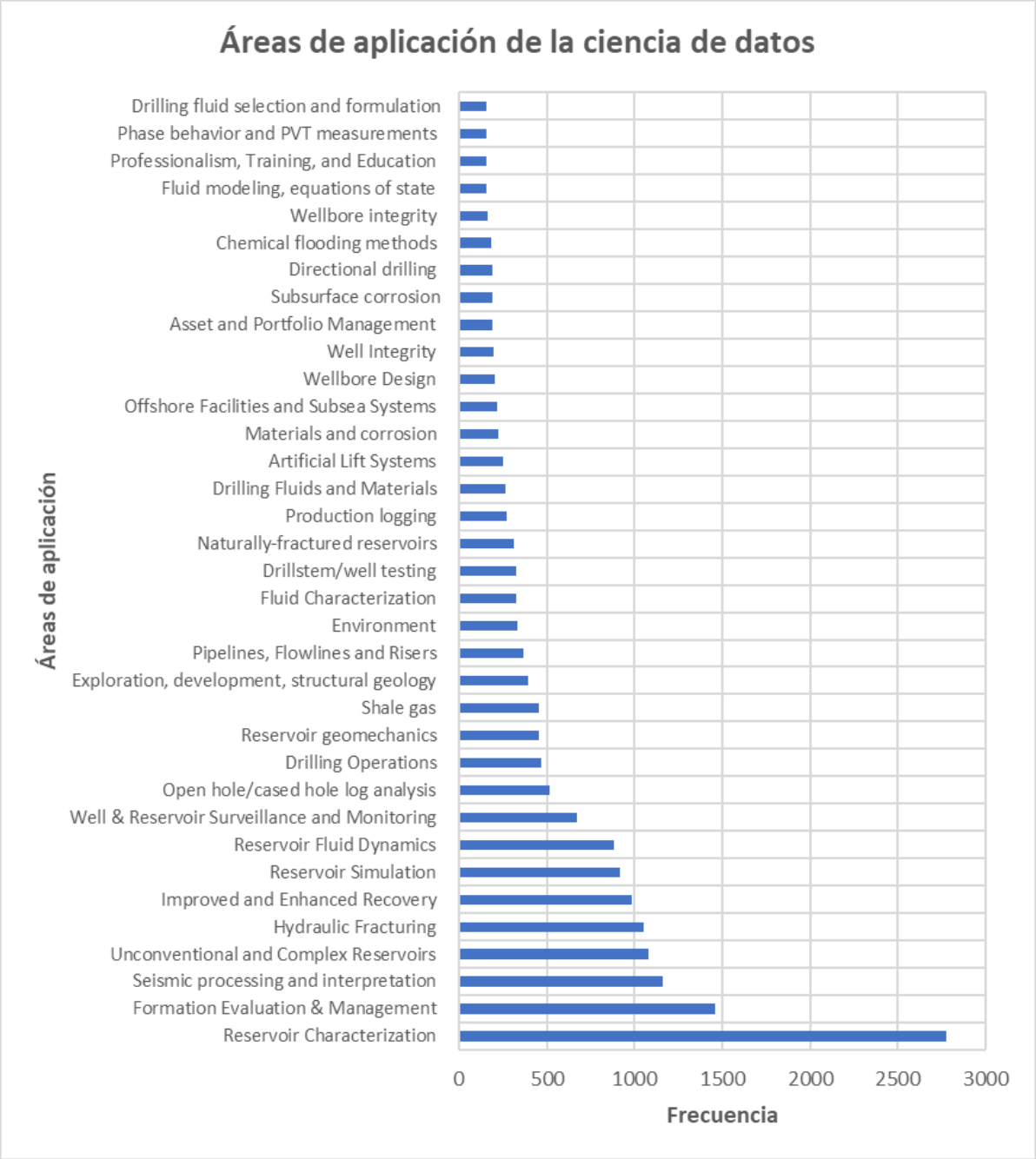
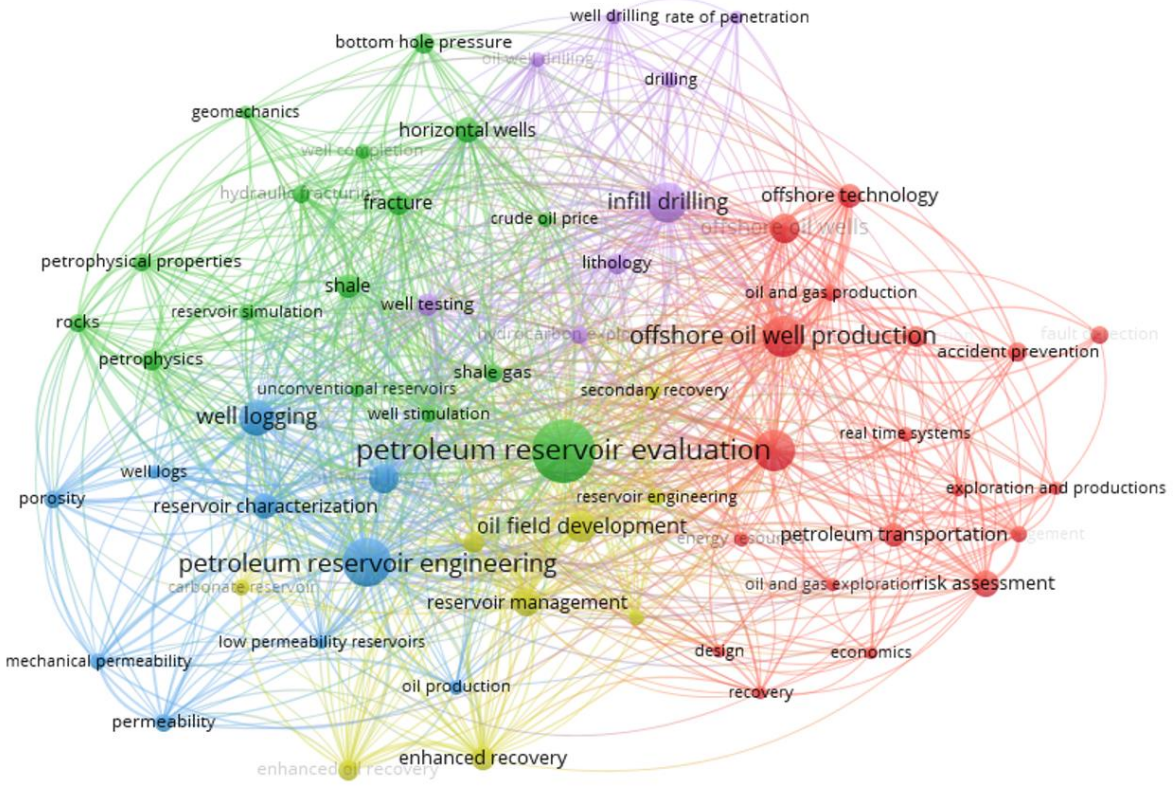


Figura 24

Análisis bibliométrico de SCOPUS (Science Direct)



Nota. La figura corresponde a una red bibliométrica generada mediante el software VOSviewer a partir de datos bibliográficos obtenidos de la base de datos SCOPUS. Cuanto mayor es el tamaño de los círculos, mayor es la frecuencia con que aparecen las palabras clave en los documentos considerados en el análisis bibliométrico.

En las Figuras 23 y 24 se puede evidenciar que el área o sector de la industria petrolera donde más se utilizan las técnicas de la ciencia de datos y la inteligencia artificial es en yacimientos, más específicamente en la caracterización de los mismos; también se aprecia un alto uso en el sector offshore, en las operaciones de perforación, en el recobro mejorado y en el fracturamiento hidráulico.

Siguiendo la idea anterior, se propondrán 4 ejemplos para el área de yacimientos relacionados con la caracterización, con la evaluación de formaciones, con la estimación de reservas y con el recobro mejorado. Con respecto al área de operaciones, se propondrán 3 ejemplos más, los cuales tendrán que ver con las operaciones de perforación, con el fracturamiento

hidráulico y con el comportamiento del pozo. Para finalizar, se plantearán 2 ejemplos para el área de administrativas y complementarias asociados con la geomecánica y con la evaluación económica de proyectos y la optimización de costos de los mismos.

3.4.1 Ejemplos para el área de yacimientos

3.4.1.1 Clasificación de litofacies y estimación de la permeabilidad.

Objetivo.

Con base en los registros de pozos y los datos de los núcleos, construir un modelo de permeabilidad para reducir la incertidumbre en la caracterización de yacimientos a través de redes neuronales probabilísticas y un modelo de regresión forzada generalizada.

Justificación.

En la industria de los hidrocarburos uno de los mayores retos es la caracterización de yacimientos, ya que estos no son homogéneos en ninguna de sus propiedades y al determinar dichas características en zonas específicas siempre estará presente la incertidumbre, por lo tanto, reducir esta incertidumbre no sólo les da una mayor precisión a los datos, sino que también permite reducir riesgos a la hora de realizar procedimientos con base en esta información (ya que es de mayor calidad). También, esta técnica permite clasificar las facies presentes en la formación, lo que permite obtener una visión más amplia del yacimiento además de poder realizar una estimación eficiente de las propiedades petrofísicas en intervalos donde no han extraído núcleos o donde no se pudieron realizar todos los registros requeridos.

Datos.

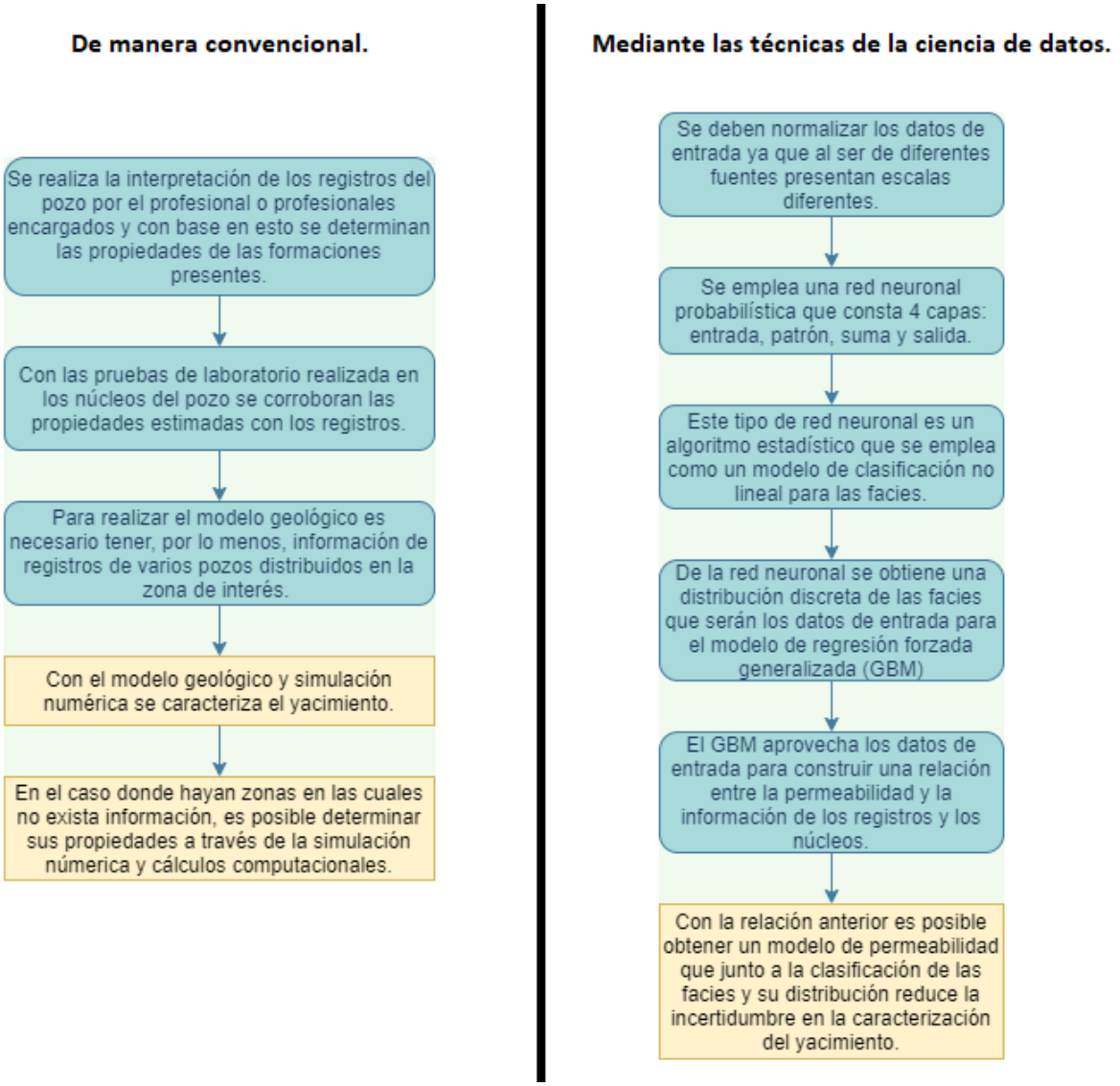
Los datos provienen de registros de pozo (neutron, gamma-ray y resistivo) y de mediciones directas de las propiedades de varios núcleos.

Metodología.

A continuación, en la Figura 25, se presenta la metodología implementada para un procedimiento convencional y para aquel basado en ciencia de datos.

Figura 25

Metodología para el modelado de la permeabilidad.



Resultados.

En el trabajo de (Al-Mudhafar, 2017) se analizan los resultados para 2 etapas. En la primera etapa, que es la clasificación de las facies distribuidas, la predicción puede ser correcta o incorrecta,

la red neuronal probabilística tuvo un 95,81% de tasa de éxito. En la segunda etapa, la permeabilidad predicha se evalúa con la observada, para lo cual el indicador R cuadrado presentó valor de 0,9953 lo que se traduce como una excelente correspondencia entre los datos reales y los datos generados por el modelo. Finalmente, se demuestra que el método es recomendable para caracterización de yacimientos heterogéneos que presentan intervalos sin datos de núcleos o registros.

3.4.1.2 Estimación de reservas.

Objetivo.

Calcular o estimar la cantidad de hidrocarburos producibles en un yacimiento no convencional a través de las técnicas de la ciencia de datos.

Justificación.

Como consecuencia de la demanda energética actual, los yacimientos no convencionales pasaron a estar en el punto de mira de la industria del petróleo debido a su gran contenido de hidrocarburos y a los avances tecnológicos que permiten su explotación. Una de las características más importantes de un yacimiento son sus reservas, que son la base del estudio económico previo a cualquier proyecto que se quiera realizar (perforación, estimulación, recobro mejorado, etc.), por lo que resulta crucial contar con un método de estimación confiable. Los modelos basados en la física (teóricos) han demostrado deficiencias para el diseño de operaciones en este tipo de formaciones; por otra parte, la simulación ha avanzado bastante en los últimos años pero presenta ciertas desventajas como el gran esfuerzo para llevarla a cabo, los elevados costos y la subjetividad de la misma hacen que sea imposible de implementar en todas las empresas o en todos los pozos, por esto, una técnica basada en ciencia de datos es la mejor elección ya que posee adaptabilidad,

precisión, riesgos y esfuerzos menores para implementarla y costos moderados que, en conjunto, beneficia a la industria.

Datos.

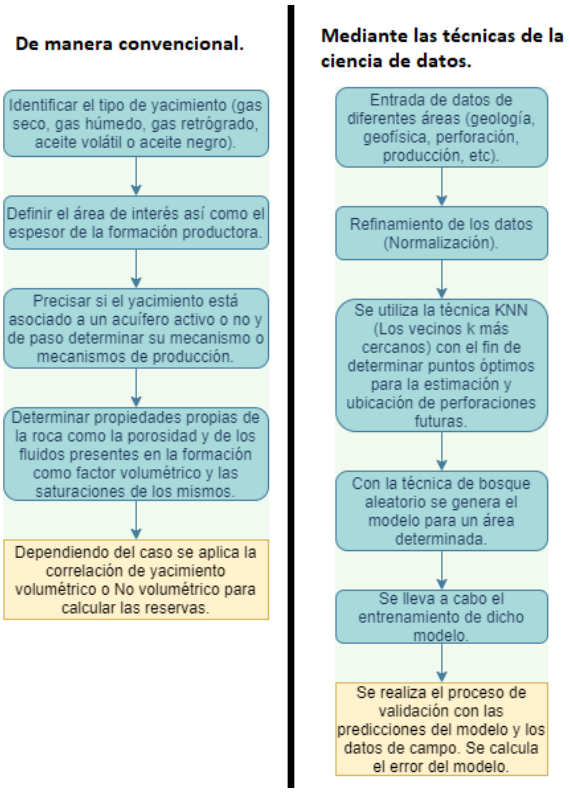
Los datos provienen de sísmica, registros de pozo, núcleos, perforación, completamiento y producción e incluye parámetros tales como las coordenadas de la posición del pozo, tipo de hidrocarburo producido, saturación, porosidad, espesor de la formación, módulo de Young, gradiente de presión, relación de Poisson, profundidad del pozo, gravedad del crudo, etc.

Metodología.

A continuación, en la Figura 26, se presenta la metodología para el método que se realiza de maneja convencional y el que se implementa con técnicas basadas en la ciencia de datos.

Figura 26

Metodología para estimación de reservas



Resultados.

El estudio de (Zhao y otros, 2020) se resume en 4 secciones: recopilación de los datos, refinamiento de datos (control de calidad), modelado de tendencias y modelado de regresión. Los resultados se presentan comparando el error al estimar las reservas con las técnicas empleadas en el mencionado estudio con un estudio previo realizado por los mismos autores en el año 2019 (Liang y Zhao, 2019) basado en la técnica de Big Data Analytics. Las estimaciones de la actual investigación mostraron reducciones en el error cuadrático de 26% para petróleo y 56% para gas en comparación con el proyecto de 2019 y, en términos generales, la estimación a partir del modelo basado en ciencias de datos fue más preciso que el resultado obtenido por el procedimiento convencional.

3.4.1.3 Interpretación de registros de pozo.

Objetivo.

Predecir la litología de una formación y la porosidad de la misma a partir de la interpretación de registros de pozo en hueco abierto mediante una red neuronal artificial.

Justificación.

La interpretación de registros de pozo es una labor de gran importancia dentro de la ingeniería de yacimientos, ya que permite determinar propiedades del yacimiento tales como la porosidad o la permeabilidad y también ayuda a identificar los fluidos presentes en la formación, tales como el aceite, el gas o el agua; lo anterior es vital para el posterior desarrollo del activo.

Esta tarea es colosal y requiere de profesionales con gran experiencia para interpretar los valores medidos en los registros y convertirlos en información útil. Adicionalmente, se requiere una gran cantidad de tiempo y los resultados, al estar sujetos a la interpretación de una persona, pueden ser subjetivos y erróneos. Considerando los motivos antes mencionados, se hace evidente

la necesidad de disponer de una alternativa más eficiente en términos de tiempo y costo y que no esté limitada por la subjetividad, por lo que el uso del machine learning se perfila como una opción a considerar.

Datos.

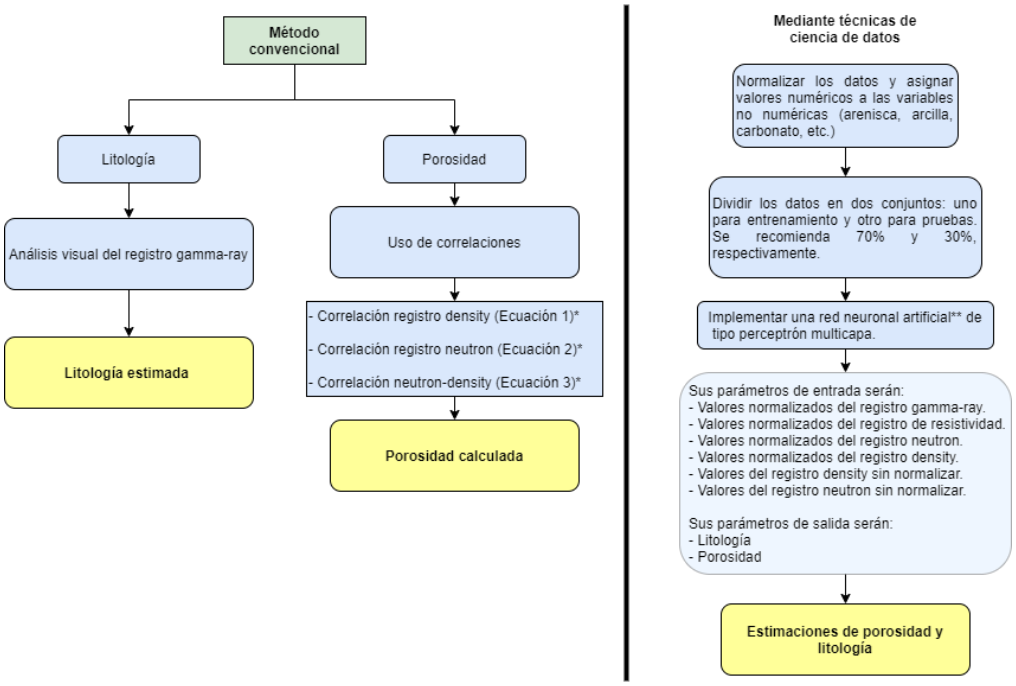
La información consiste en las medidas de los registros gamma-ray, neutron, density y resistividad. Es necesario normalizar los valores.

Metodología

En la Figura 27 se presentan las dos alternativas para abordar la interpretación de los registros de pozo: por la manera convencional y mediante la aplicación de técnicas de la ciencia de datos.

Figura 27

Metodología para la interpretación de registros de pozo



Nota. * Las ecuaciones indicadas se encuentra en el apéndice H. ** El proceso de implementar una red neuronal artificial se presenta en el capítulo 4.1.

Resultados.

Con el fin de comparar los resultados que se obtienen al realizar la interpretación de registros de pozo en hueco abierto mediante la manera convencional y aquellos obtenidos mediante técnicas de la ciencia de datos, se aprovechará el estudio realizado por (Gupta y Soumya, 2020).

En este estudio, los autores utilizan un tipo de red neuronal artificial conocida como perceptrón multicapa con 6 neuronas de entrada, siendo los parámetros ingresados los valores normalizados de los registros gamma-ray, density, neutron y resistivo y los valores sin normalizar de los registros density y neutrón. La red permitía predecir la litología de la formación y su porosidad y, según los autores, las predicciones de la litología fueron altamente precisas, mientras que la calidad de las predicciones de la porosidad dependía en gran medida de la precisión de los datos de los registros. Ellos concluyeron que la alternativa estudiada tuvo buena precisión y un potencial considerable en cuanto a reducir costos y tiempo en la interpretación de los registros.

3.4.1.4 Determinación del desempeño de la inyección de agua como método de recobro.

Objetivo.

Desarrollar un modelo predictivo del índice de desarrollo de un pozo, el cual se utilizará posteriormente para evaluar el desempeño de la inyección de agua.

Justificación.

Cuando la presión de un yacimiento con mecanismo de producción primaria comienza a reducirse o está cerca de hacerlo, es necesario una alternativa que permita conservar la energía de la formación para que se mantenga la producción, siendo la inyección de agua la alternativa empleada en la mayoría de los casos. Uno de los retos para cada campo es determinar el

rendimiento de esta técnica antes de ser implementada para definir si es rentable aplicarla o no y para ello se cuenta convencionalmente con el método de simulación numérica.

Aunque la simulación numérica produce resultados con aceptable precisión, posee la desventaja de requerir mucho tiempo, ser bastante compleja debido a todos los parámetros y apreciaciones que se deben tener en cuenta y depender del ajuste de los datos históricos. Con lo anterior en cuenta, se propone como alternativa el uso de inteligencia artificial enfocada en Big data para extraer información útil a partir de grandes volúmenes de datos y definir el rendimiento del proceso de inyección superando las desventajas que ya se mencionaron.

Datos.

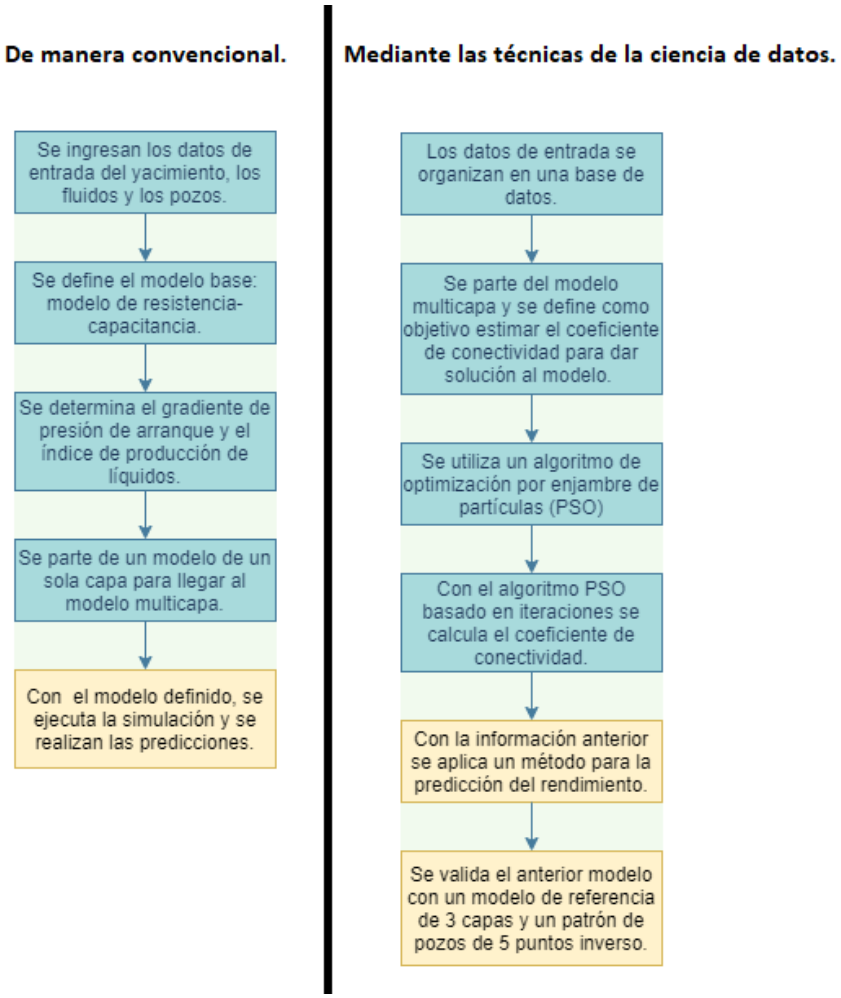
La información para desarrollar el modelo predictivo se compone de datos de inyección, de producción, de fluidos, geológicos y de presión.

Metodología.

A continuación, en la Figura 28, se presentan la metodología convencional y la metodología basada en técnicas de ciencia de datos empleadas para dar cumplimiento al objetivo planteado.

Figura 28

Metodología para evaluar el rendimiento de inyección de agua.



Resultados.

Se ha utilizado el estudio de (Chen et al., 2021) para presentar las ventajas de utilizar técnicas de la ciencia de datos en el análisis del desempeño de la inyección de agua con respecto al uso de la simulación numérica convencional. Los autores desarrollaron un modelo de conectividad dinámica entre pozos y a partir de tal modelo propusieron un método para la predicción del índice de producción dinámico. Ellos utilizaron su modelo varios campos del Mar de Bohai y obtuvieron resultados con alta precisión y en un tiempo muy inferior al que le tomaría a un simulador numérico.

3.4.2 Ejemplos para el área de operaciones

3.4.2.1 Predicción de la tasa de penetración (ROP) a partir de parámetros de perforación.

Objetivo.

Predecir la ROP usando el mínimo de parámetros posibles mediante una red neuronal artificial.

Justificación.

La tasa de penetración es la velocidad con la cual se rompe la roca por debajo de la broca y mide el avance o progreso de esta mientras se perfora una formación. Buena parte del presupuesto de un pozo se gasta en la fase de perforación, la cual depende, en gran medida, de la ROP. De lo anterior se intuye que una acertada estimación y posterior optimización de la ROP puede llevar a disminuir los costos de la perforación del pozo.

Datos.

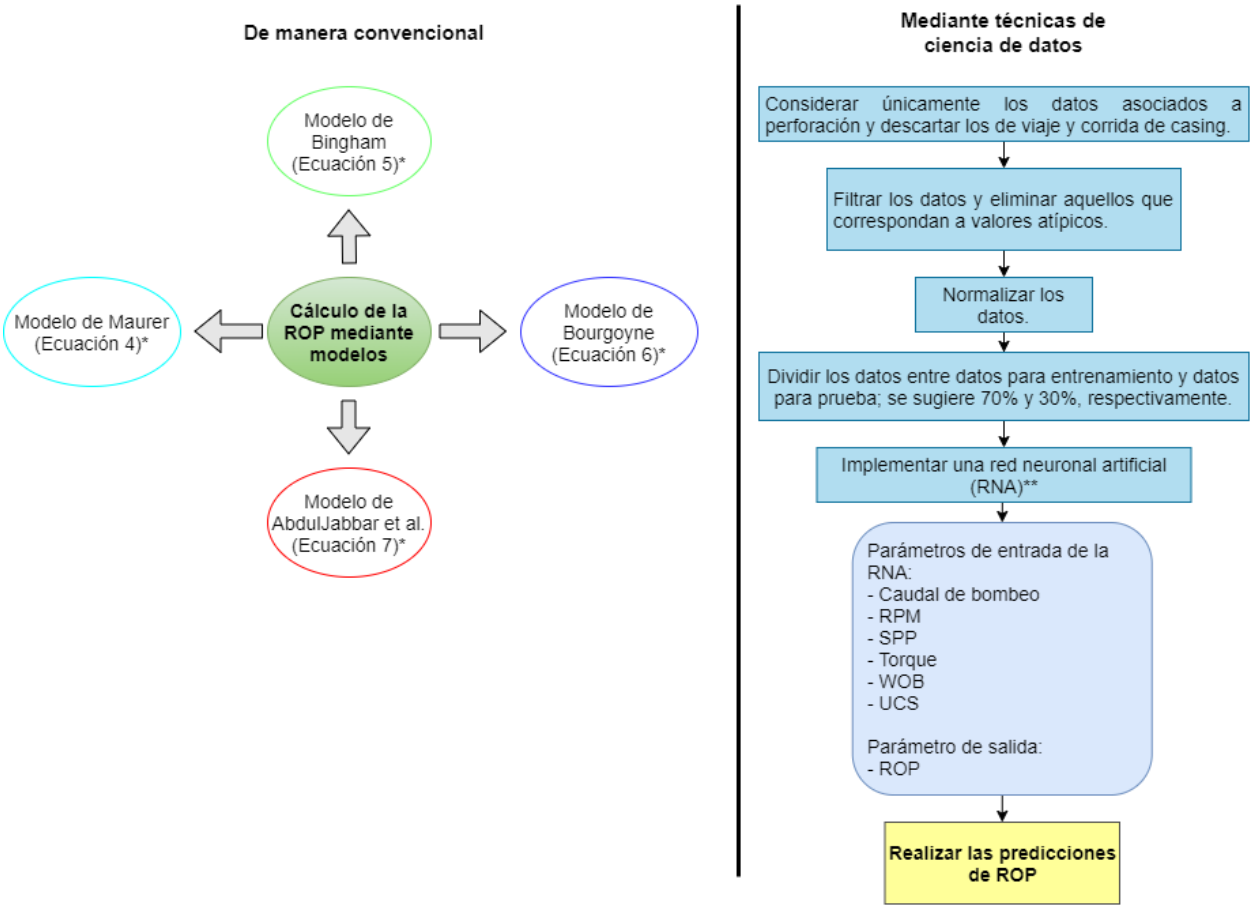
El conjunto de datos se compone de los siguientes parámetros: ROP, caudal de bombeo, rotaciones por minuto (RPM) de la broca, presión en la tubería vertical (Standpipe Pressure, SPP), torque, peso sobre la broca (WOB) y la resistencia a la compresión no confinada (UCS).

Metodología.

La Figura 29 ilustra los dos enfoques que pueden emplearse para determinar la ROP: el enfoque convencional, que consiste de modelos matemáticos, y el enfoque basado en el uso de técnicas de la ciencia de datos.

Figura 29

Metodología para determinar la tasa de penetración (ROP)



Nota. *Las ecuaciones indicadas se encuentran en el Apéndice H. **El procedimiento para implementar una red neuronal artificial se explica en la sección 4.1.

Resultados.

(Al-AbdulJabbar et al., 2018) implementaron una red neuronal artificial prealimentada que recibía como datos de entrada los 6 parámetros indicados en la Figura 29, tenía una capa oculta con 12 neuronas y su salida era la ROP. Ellos usaron la función de entrenamiento Levenberg-Marquart y la función de activación lineal. La red fue desarrollada con 1528 puntos de datos, correspondientes a un pozo denominado “pozo A”, de los cuales el 70% se usó en el entrenamiento y el 30% en pruebas. Posteriormente, la red fue probada usando los datos del “pozo B”, el cual tenía 2157 puntos de datos, y los datos del “pozo C”, con 849 puntos de datos.

Como resultado, encontraron que el error porcentual absoluto promedio (AAPE) fue de 9.8%, 7.4% y 8.7% para los pozos A, B y C, respectivamente. Además, compararon los errores de la RNA con los errores obtenidos mediante el modelo de Maurer (36.5%), el modelo de Bingham (17.1%), el modelo de Bourgoyne (12%) y el modelo de AbdulJabbar et al. (10.9%) y concluyeron que la predicción de la ROP mediante una RNA tiene mayor precisión que los modelos convencionales y, además, la RNA tiene la ventaja de predecir la ROP de otros pozos basándose en los datos de un único pozo.

3.4.2.2 Diseño de fracturamiento hidráulico.

Objetivo.

Desarrollar un modelo basado en inteligencia artificial que permita medir el impacto de diversos parámetros involucrados en el fracturamiento hidráulico y posibilite optimizar dicho proceso. El impacto de los parámetros se mide al estimar la producción de aceite o gas de acuerdo con el valor de los parámetros ingresados al modelo generado.

Justificación.

El fracturamiento hidráulico es una práctica que se emplea para estimulación de pozos, con lo cual se consigue aumentar la producción y el recobro. Por otra parte, también es una técnica fundamental en la explotación de yacimientos de shale, donde las permeabilidades extremadamente bajas características de estos yacimientos requieren la creación de fracturas artificiales para poder producir los fluidos allí presentes.

El diseño del fracturamiento hidráulico es una tarea ardua y que normalmente requiere del uso de simuladores numéricos para evaluar los diferentes escenarios, lo que implica un importante gasto de tiempo y la necesidad de personal con suficiente experiencia para que pueda interpretar

acertadamente los resultados de las simulaciones. El uso de inteligencia artificial y minería de datos se presenta como una alternativa para para hacer análisis basados en los datos y desarrollar modelos que asistan en el diseño del fracturamiento hidráulico.

Datos.

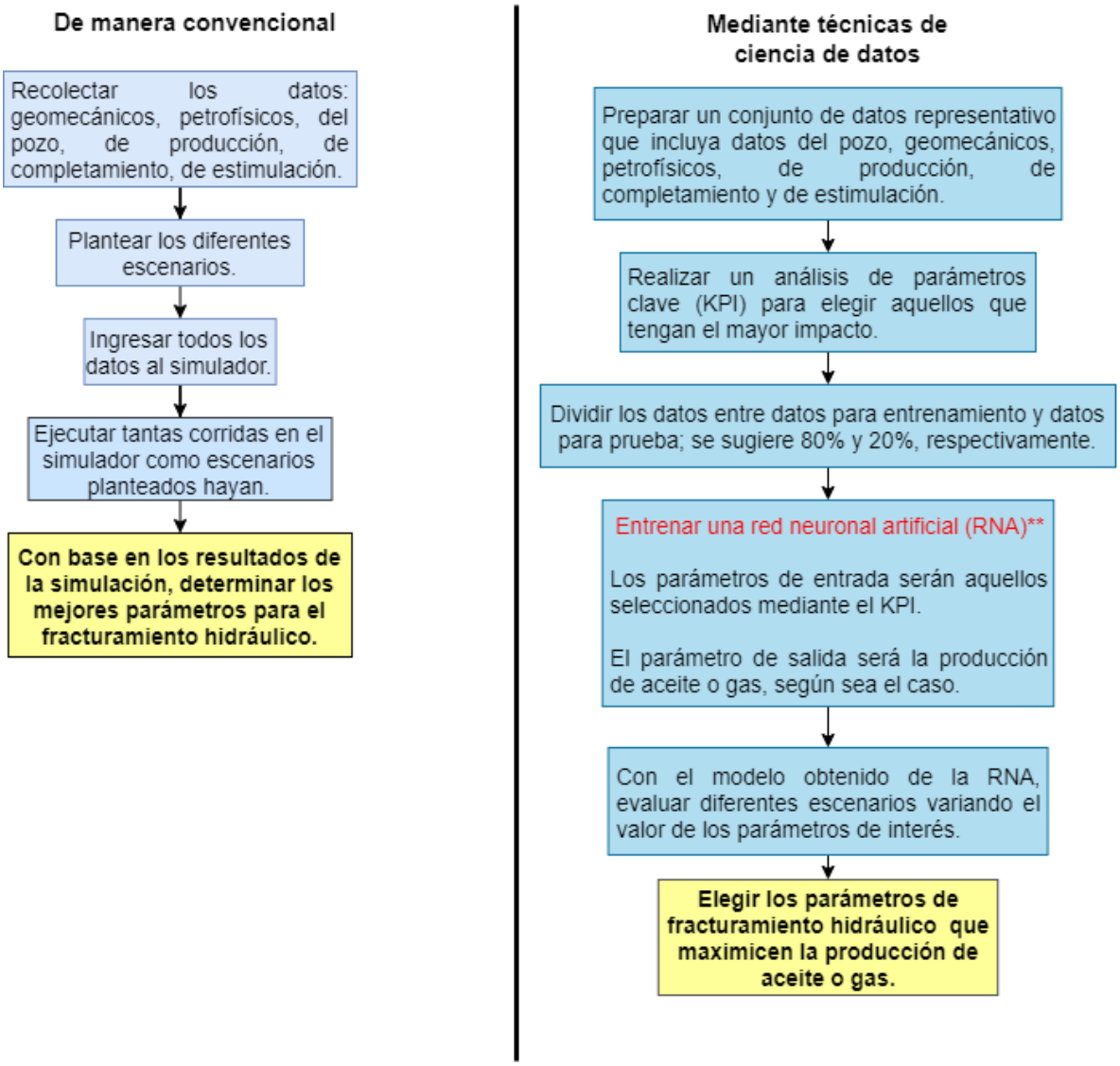
Los datos corresponden a información del pozo (coordenadas, profundidad, azimut, etc.), características del yacimiento (porosidad, permeabilidad, espesor neto, etc.), propiedades geomecánicas, propiedades del completamiento, propiedades operacionales (gas producido acumulado, aceite producido acumulado, presión en cabeza del pozo) y propiedades del tratamiento de estimulación (presión de inyección, volumen de lechada inyectada, volumen de propante, etc.).

Metodología.

En la Figura 30 se ilustra metodológicamente las maneras como se lleva a cabo la evaluación de diferentes escenarios de fracturamiento hidráulico con el fin de elegir los parámetros que hagan más eficiente el proceso, tanto de la manera convencional como usando técnicas de la ciencia de datos.

Figura 30

Optimización del diseño de fracturamiento hidráulico



Nota. ** El procedimiento para implementar una red neuronal artificial se explica en la sección 4.1.

Resultados.

Con la intención de ofrecer una perspectiva sobre los beneficios de utilizar técnicas de la ciencia de datos en el procedimiento de diseño y optimización del fracturamiento hidráulico, se analizó el trabajo realizado por (He et al., 2019). Los mencionados autores desarrollaron un modelo para evaluar el impacto de diversos parámetros de estimulación y completamiento con el fin de

optimizar el proceso de fracturamiento hidráulico. El modelo fue desarrollado mediante una red neuronal artificial con datos tomados de 150 pozos de gas de la formación de shale Marcellus.

El coeficiente de correlación de entrenamiento de la red neuronal fue de 0.93699, lo cual es un valor aceptable e implica que el modelo obtenido con la red puede ser utilizado para evaluar diferentes escenarios de fracturamiento hidráulico y es de esperar resultados con buena precisión; además, el tiempo que le toma al modelo evaluar cada escenario es inferior a 1 segundo, lo cual le da una enorme ventaja con respecto a los simuladores numéricos.

3.4.2.3 Predicción del comportamiento de afluencia (IPR) en pozos verticales de aceite con empuje por gas en solución.

Objetivo.

Predecir de manera rápida y confiable el IPR para yacimientos con empuje por gas en solución.

Justificación.

La predicción del IPR de un pozo de aceite es muy importante para determinar el esquema óptimo de producción y para el diseño de los equipos de producción y el sistema de levantamiento artificial.

Datos

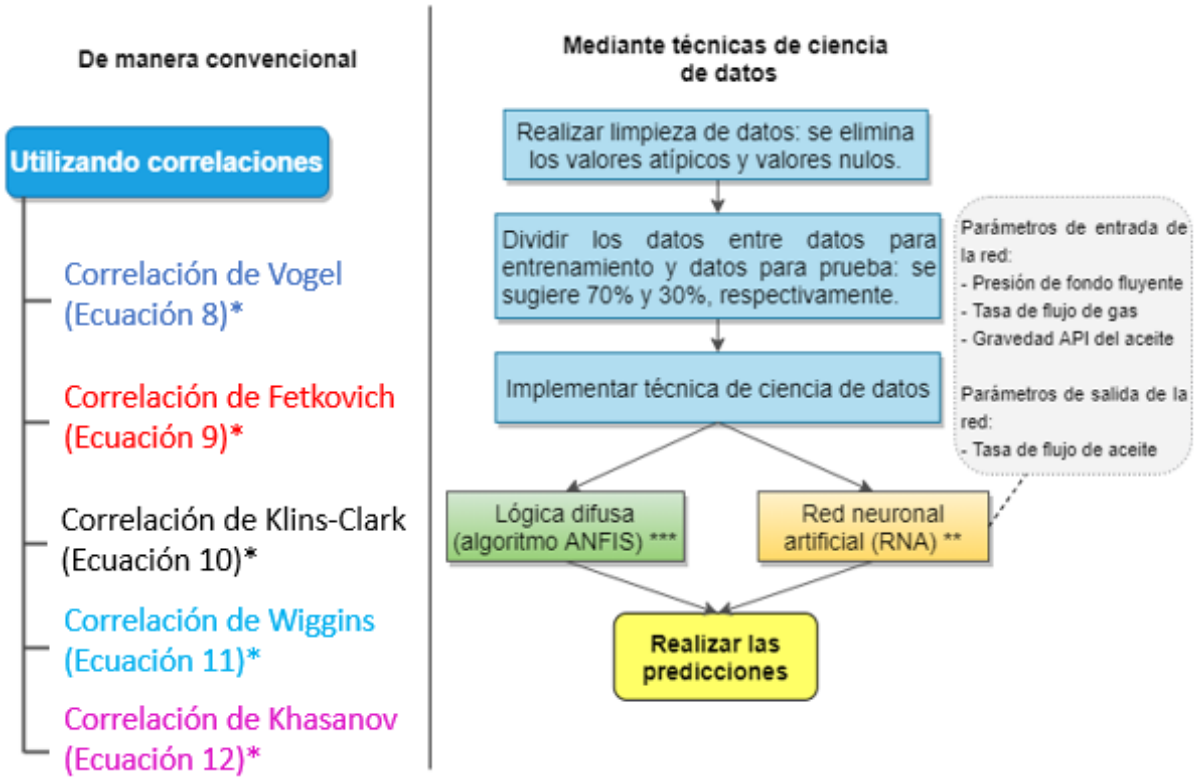
El conjunto de datos está compuesto por valores de los siguientes parámetros: tasas de flujo de aceite, agua y gas, presión del yacimiento, gravedad API del aceite, presión de fondo fluyente (P_{wf}) y su correspondiente tasa de flujo de aceite.

Metodología

La Figura 31 ilustra una comparación entre los procedimientos para determinar el comportamiento de afluencia del pozo (IPR), siendo el primero el método o manera convencional y el segundo, una alternativa basada en técnicas de la ciencia de datos.

Figura 31

Metodología para determinar el comportamiento de afluencia (IPR)



Nota. *Las ecuaciones indicadas se encuentran en el Apéndice H. **El procedimiento para implementar una red neuronal artificial se explica en la sección 4.1. *** El procedimiento para implementar el algoritmo ANFIS se explica en la sección 4.6

Resultados

En el estudio realizado por (Basfar et al., 2018), los autores implementaron una red neuronal artificial (RNA) de tipo perceptrón multicapa con dos capas ocultas que recibía como datos de entrada la presión de fondo fluyente, la tasa de flujo de gas y la densidad API del aceite

y cuya salida correspondía a la tasa de flujo del aceite. Además, también implementaron un modelo de lógica difusa basado en el algoritmo ANFIS utilizando la función Genfis2, 500 *epochs* (iteraciones) de entrenamiento y un radio de agrupamiento de 0.04.

Los autores evaluaron la calidad de la predicción mediante el error relativo porcentual promedio (APRE). Como resultado obtuvieron que, de las correlaciones empíricas, la más precisa fue la de Fetkovich, con un APRE de 13.5%; por otra parte, con el modelo basado en redes neuronales artificiales el APRE fue de 12.67% y con el modelo basado en lógica difusa se obtuvo un APRE de 1.22%. Basándose en lo anterior, se concluyó que el modelo basado en lógica difusa es altamente preciso y supera ampliamente a las correlaciones empíricas existentes.

3.4.3 Ejemplos para el área de administración y complementarias

3.4.3.1 Predicción de parámetros de falla de roca.

Objetivo.

Predecir o estimar la cohesión y el coeficiente de fricción interna de una roca para determinar su criterio de falla a través de técnicas de la ciencia de datos: red neuronal artificial y regresión de vectores de soporte (SVR).

Justificación.

Mantener la estabilidad del pozo durante las operaciones de perforación es una tarea crucial puesto que en caso de que se den eventos de inestabilidad, como pueden ser los *breakouts* y las fracturas inducidas, se pueden presentar pegas de tubería, desvíos de la trayectoria del pozo y pérdidas de circulación; tales problemáticas se traducen en pérdidas económicas, aumento del tiempo de perforación y daños en el pozo.

Para evaluar la estabilidad del pozo se emplea un modelo constitutivo (con el que se estiman los esfuerzos alrededor del pozo) en conjunto con un criterio de falla que es utilizado para predecir la máxima resistencia de las rocas de la formación. Uno de los criterios de falla más utilizados es el criterio de Mohr-Coulomb, el cual se fundamenta, especialmente, en los parámetros de cohesión y coeficiente de fricción interna.

La cohesión y el coeficiente de fricción interna son determinados mediante pruebas de laboratorio. Las mediciones de laboratorio de estos parámetros son altamente precisas y confiables, sin embargo, también son costosas y toman bastante tiempo. En tal sentido, resulta atractivo considerar la alternativa de emplear técnicas de la ciencia de datos para estimar los mencionados parámetros, toda vez que podría suponer una opción rápida y de bajo costo.

Datos.

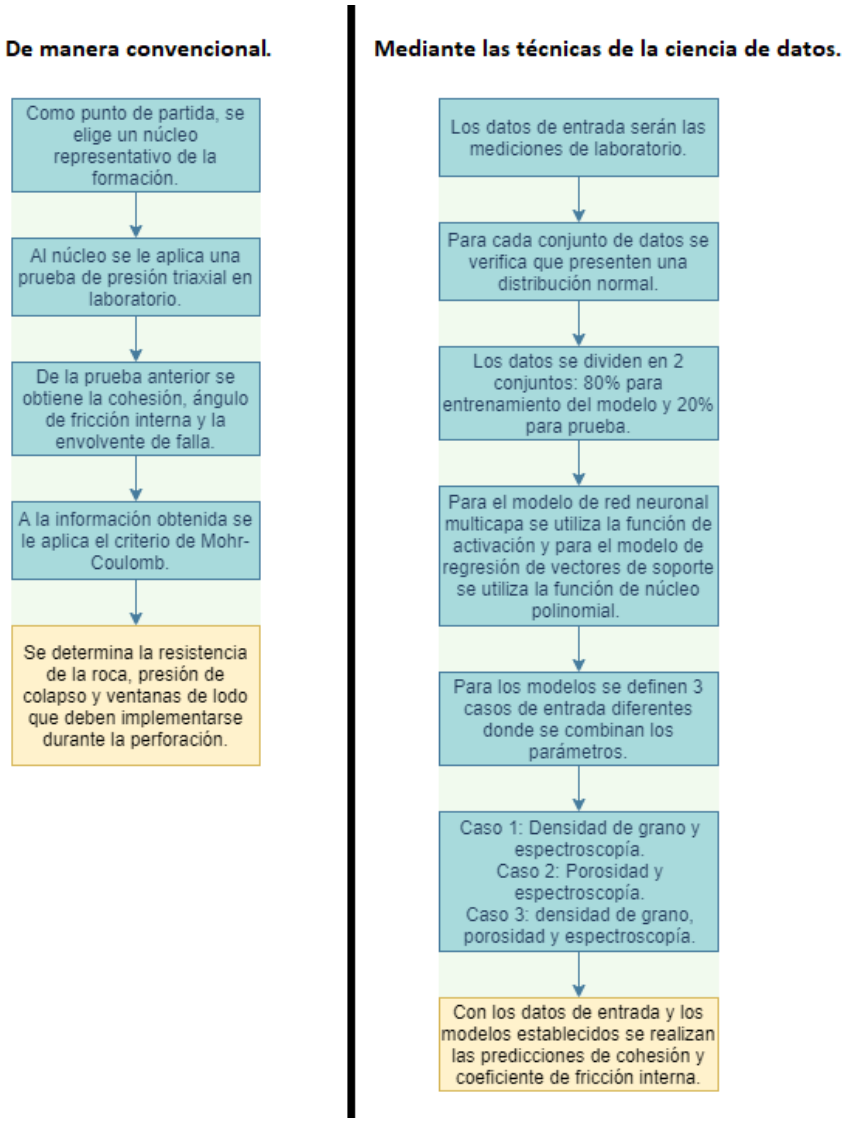
La información proviene de muestras de núcleos a los que, a través de pruebas de laboratorio, se les extrae datos de densidad de grano, porosidad, densidad aparente, espectroscopía elemental, coeficiente de fricción interna y cohesión (estos 2 últimos pueden ser estimados a partir de la envolvente de falla de Mohr-Coulomb).

Metodología.

En la Figura 32 se presenta la metodología abordada para la predicción de parámetros de falla de rocas de manera convencional y a través de las técnicas de la ciencia de datos.

Figura 32

Metodología para la estimación de parámetros de falla de roca



Resultados.

Los autores (Negara et al., 2018) utilizaron las técnicas de red neuronal artificial (específicamente, perceptrón multicapa) y regresión de vectores de soporte (SVR) para generar modelos con los cuales estimar la cohesión y el coeficiente de fricción interna. Ellos lograron determinar que, al utilizar la densidad de grano, la porosidad y la espectroscopía elemental como

parámetros de entrada, tanto para entrenar los modelos de red neuronal artificial y SVR, se obtienen los mejores resultados en la estimación de la cohesión y el coeficiente de fricción interna.

De su estudio, pudieron concluir que el uso de la red neuronal artificial permite obtener resultados más precisos que aquellos obtenidos mediante SVR. Al momento de predecir el coeficiente de fricción interna, obtuvieron un coeficiente de determinación R^2 de 0.951 y un error porcentual absoluto promedio (MAPE) de 9.185%. Por otra parte, durante la predicción de la cohesión se obtuvo un R^2 de 0.888 y un MAPE de 12.501%. A partir de tales resultados se concluye que el uso de redes neuronales artificiales y la regresión de vectores de soporte representa una alternativa rápida, económica y con buena precisión para la estimación indirecta de los parámetros de falla de la roca.

3.4.3.2 Evaluación económica y optimización de costos.

Objetivo.

Con base en un modelo predictivo de aprendizaje automático y teniendo en cuenta los ingresos y egresos a nivel de campo, plataforma o pozo realizar una planificación presupuestaria óptima para reducir gastos.

Justificación.

La industria de los hidrocarburos depende en gran medida del precio del barril en el mercado mundial, y este varía dependiendo de distintos factores, algunos de ellos directamente relacionados con las características del hidrocarburo como tal, como puede ser el tipo de crudo al que se haga referencia, aunque también existen otros factores indirectos que juegan un papel muy importante y tienen gran impacto, como la geopolítica. En ocasiones, se puede llegar a disminuir

en gran medida el precio del barril, lo que conlleva a que las empresas se vean obligadas a aplicar medidas radicales con el objetivo de mitigar el daño financiero.

La evaluación de los costos y el análisis financiero es un componente vital para una empresa, ya que define cómo se deben realizar las operaciones en el presente y en el futuro inmediato. Es indispensable que estos procedimientos sean precisos, rápidos, confiables y detallados, puesto que de ello depende la economía de la empresa. Teniendo en cuenta lo anterior, la implementación de las técnicas de la ciencia de datos en estos procesos puede ofrecer grandes ventajas y proporcionar mejores resultados que los métodos convencionales, lo que se traduce como un gran beneficio para la industria.

Datos.

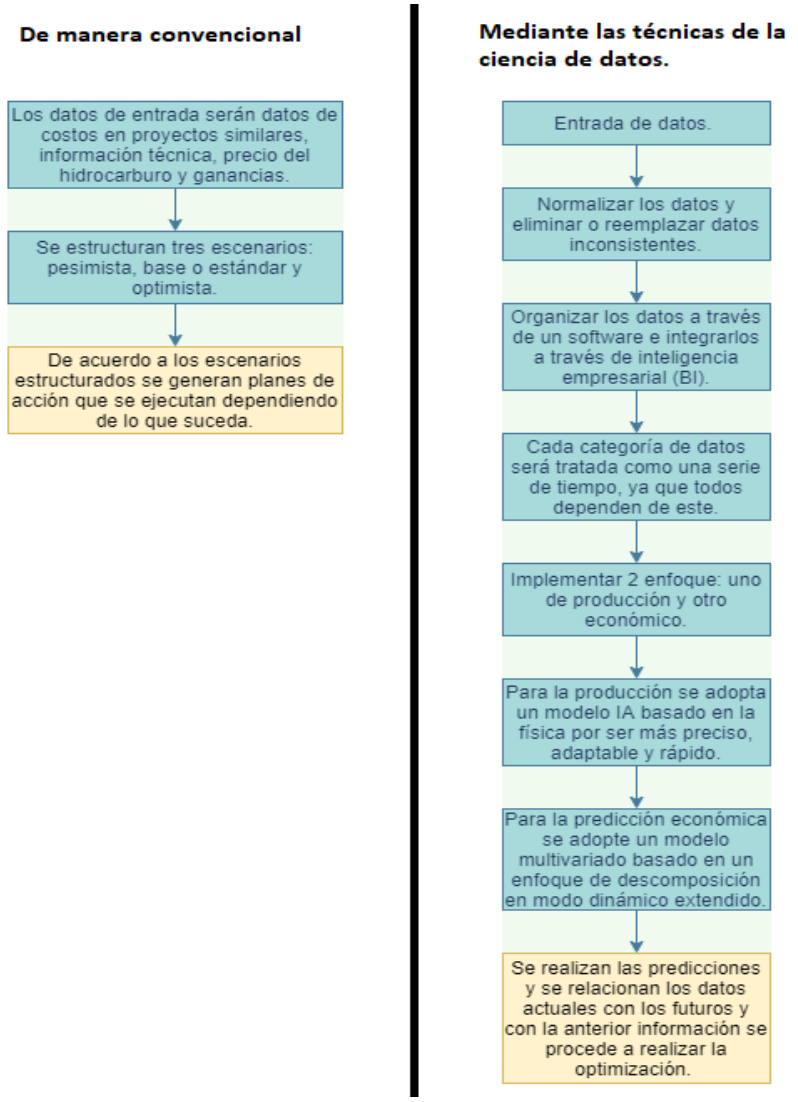
Se utilizaron los datos de un campo compuesto de 600 pozos de petróleo y gas (producción convencional) y 50 instalaciones centrales de procesamiento. Cabe resaltar que cada empresa tiene una forma diferente de organizar sus datos económicos y que, en general, se busca integrar distintos tipos de información de diferentes fuentes.

Metodología.

En la Figura 33 se presenta la metodología empleada en el proceso.

Figura 33

Metodología para la optimización de costos



Nota: Debido a que cada empresa define su manera de llevar a cabo la optimización de costos de forma convencional, en ese caso, se aborda la metodología de manera general.

Resultados.

Los autores (Klie et al., 2020) obtuvieron en el presente trabajo un ahorro de más de 300000 $US mensuales como resultado que equivale a aproximadamente el 30% de los gastos en campo. Todo lo anterior a través de procesos basados en inteligencia artificial que permiten flujos de trabajo automatizados, formas rápidas y robustas de generar decisiones económicas y la capacidad de visualizar oportunidades para optimizar los gastos de operación.

4. Metodología para la implementación de las técnicas y algoritmos de la ciencia de datos a los procedimientos seleccionados

En este capítulo se presenta, de manera conceptual, el proceso de implementación de los diferentes algoritmos y técnicas de la ciencia de datos que pueden ser aplicados a los procedimientos de ingeniería que han sido indicados en la Tabla 10 (Sección 3.3). Como puede notarse en la mencionada tabla, los diferentes procedimientos pueden ser abordados utilizando diversas técnicas de la ciencia de datos, sin embargo, predomina el uso de las redes neuronales artificiales y del Machine Learning y sus variados algoritmos.

4.1 Redes Neuronales Artificiales

Una red neuronal artificial es una técnica de *Machine Learning* que simula el comportamiento del sistema nervioso humano para el aprendizaje. Este se conforma de unidades computacionales llamadas neuronas, las cuales se conectan unas a otras a través de *pesos* que afectan las funciones calculadas en la respectiva unidad. Una red neuronal sirve para la resolución de problemas complejos, tal como el reconocimiento de patrones, imágenes, predicciones a partir de secuencias, entre otros (Aggarwal, 2018).

Los pasos para la implementación de una red neuronal básica son los siguientes (Tulleken, 2009):

1. Seleccionar el tipo de red neuronal: existen diversos tipos con diferentes fines, por ejemplo, está el perceptrón multicapa que es útil para clasificación y reconocimiento de patrones; otro muy conocido es la red neuronal recurrente, diseñada para tratar datos secuenciales tales como textos, series de tiempo, secuencias biológicas, entre otros. También vale la pena nombrar las redes neuronales convolucionales cuyo enfoque es

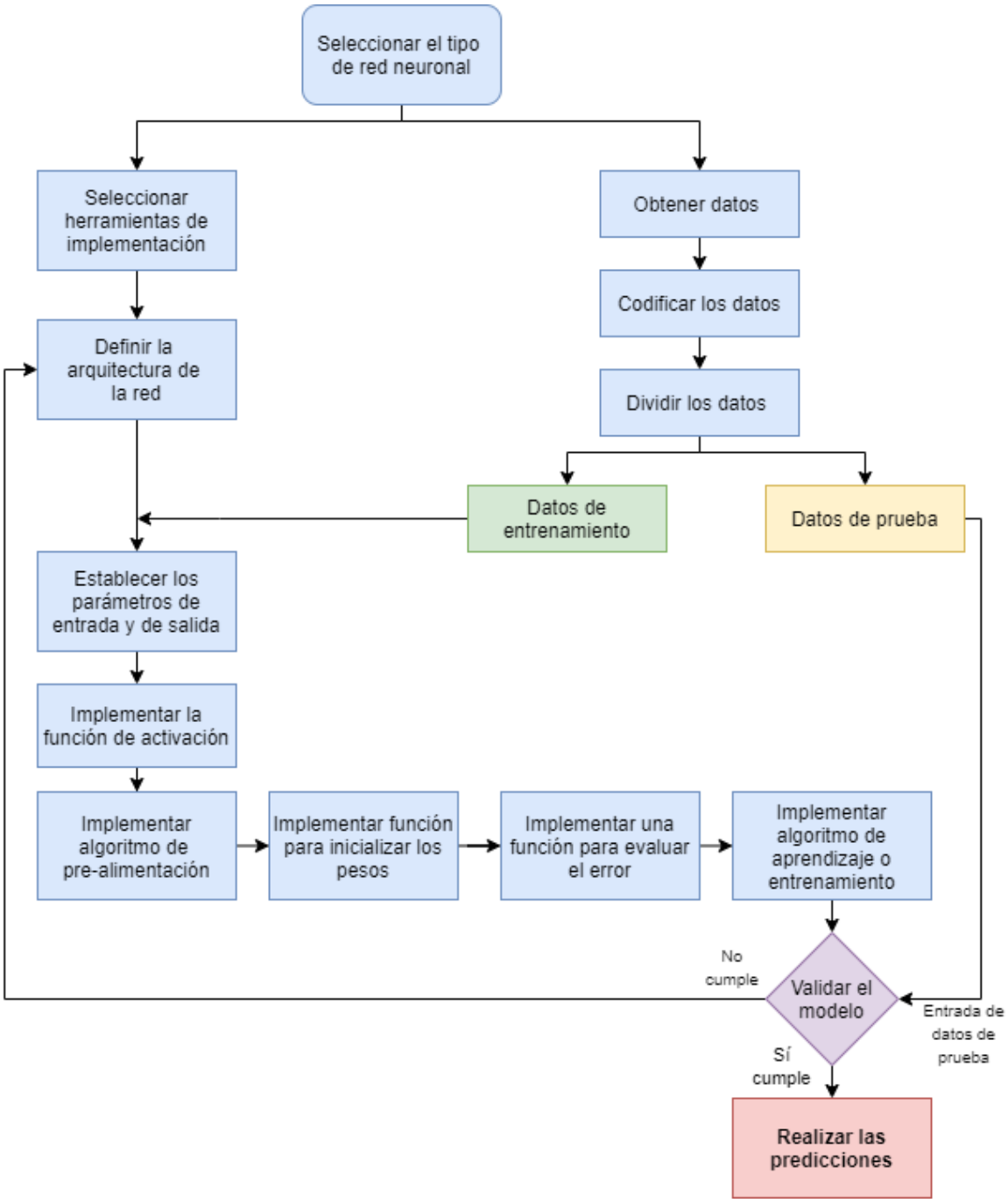
utilizado para computación visual por su alto rendimiento en el reconocimiento de imágenes.

2. Seleccionar las herramientas de implementación: existen diversos lenguajes y librerías que se pueden utilizar, por lo que hay que seleccionar las que se ajusten al problema a resolver y dependerán del tipo de red que se construirá.
3. Definir la arquitectura de la red: se debe establecer el número de capas ocultas y el número de neuronas iniciales en cada capa oculta.
4. Establecer los parámetros de entrada y de salida: con base en el problema a resolver y los datos disponibles, se establecen los datos de entrada a la red (parámetros de entrada) y el o los resultados que se obtienen (parámetros de salida).
5. Obtener los datos: este dependerá del tipo de problema a resolver, los datos pueden ser imágenes, textos, números, clases, entre otros. Es importante contar con una buena cantidad de datos significativos y correctamente depurados.
6. Codificar los datos: no importa el formato inicial de los datos, es importante convertirlos en una representación numérica para que puedan ser utilizados en la red neuronal; comúnmente se utilizan vectores o matrices para la representación.
7. Dividir los datos: se deben dividir entre datos de entrenamiento (es el conjunto con la mayor cantidad, generalmente, 70%) y datos de prueba (conjunto con menor cantidad de datos, generalmente, 30%).
8. Implementar la función de activación: esta devuelve una salida que será generada por la neurona dada una entrada o conjunto de entradas.
9. Implementar algoritmo de pre-alimentación: este paso es aplicable en redes que no tienen bucles de retroalimentación, tal como el perceptrón multicapa, y se encarga de propagar la información comenzando desde la capa de entrada, pasando por todas las capas ocultas y generando una salida.
10. Implementar una función para inicializar los pesos: la forma más común de hacerlo consiste en recibir un peso máximo, un alto y un ancho y retornar una matriz con pesos asignados aleatoriamente según el máximo asignado.

11. Implementar una función para evaluar el error: el error cometido por la red neuronal depende de los pesos y *bias* de las neuronas. Existen diversas maneras de calcularlo, pero una de las más comunes es el error cuadrático medio.
12. Implementar el algoritmo de aprendizaje o entrenamiento: este dependerá del tipo de red seleccionada. Consiste en un algoritmo iterativo que ajusta cada uno de los pesos de las entradas de todas las neuronas para que las respuestas de la capa de salida se ajusten lo más posible a los datos correctos. Un algoritmo ampliamente utilizado es el de propagación hacia atrás o *backpropagation,* el cual compara la salida de la red con el objetivo a alcanzar para producir un error, luego este error se propaga hacia atrás iterativamente hasta que converja la salida de dicha red. Durante la ejecución, los pesos y *bias* de las neuronas se van actualizando.
13. Validar el modelo: una vez entrenado el modelo, se realizan pruebas con los datos de prueba para verificar que las respuestas obtenidas son las esperadas: si lo son, se procede al último paso; si no, se modifica la arquitectura de la red y, opcionalmente, se prueba cambiando la función de activación y el algoritmo de aprendizaje de la red.
14. Realizar las predicciones.

Figura 34

Proceso de implementación de una red neuronal artificial



4.2 Big Data Analytics

Es un proceso mediante el cual se analizan grandes volúmenes de datos que puedan ser procesados y transformados en información relevante que permita dar respuesta a interrogantes

según los objetivos que se desean alcanzar. Los pasos fundamentales a seguir en este procedimiento son los siguientes (Zhang, 2017):

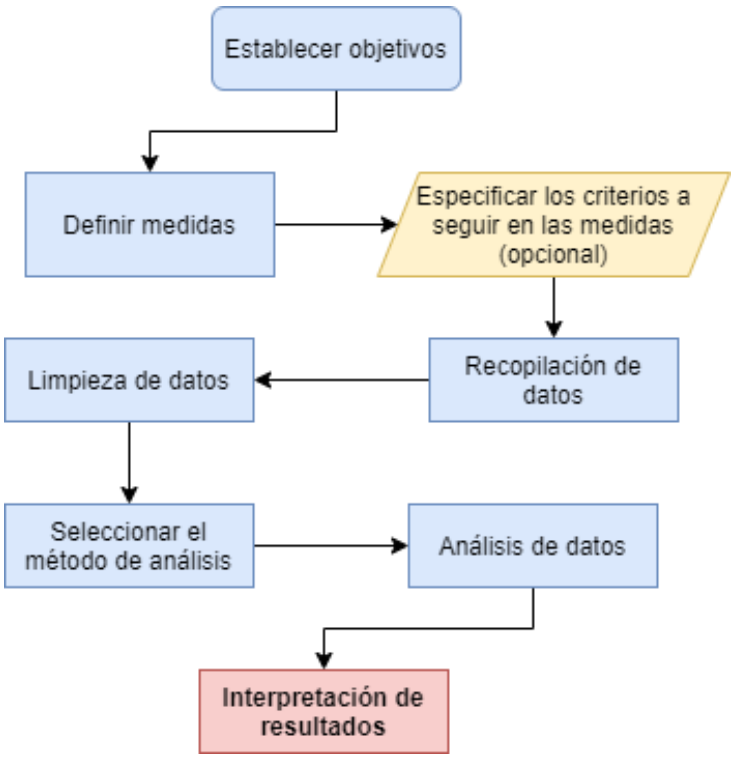
1. Establecer los objetivos: se plantean las preguntas e incógnitas que se desean responder mediante los datos; estas dependerán de las necesidades del estudio o problema a resolver.
2. Definir las medidas a realizar: dependerán de las preguntas planteadas en el paso anterior, una vez conocido qué es lo que se desea responder, se debe elegir cuál es el dato que responderá dicha incógnita; este podría ser, por ejemplo, tiempo, cantidad monetaria, cantidad de objetos o personas, entre otros.
3. Especificar los criterios a seguir en las medidas: se deben especificar las unidades de medidas que se manejarán, así como también considerar los factores que se incluyen o no en el cálculo de los datos, tal como bonificaciones, descuentos, entre otros.
4. Recopilación de datos: se debe recolectar los conjuntos de datos a utilizar de acuerdo con las necesidades, medidas y criterios establecidos. Se recomienda comprobar si existen datos disponibles que ya den respuesta a las preguntas establecidas para así evitar datos duplicados. Luego, se deben implementar métodos para obtener los datos faltantes, tales como encuestas, pruebas de laboratorio, estudios de observación o, incluso, buscar en registros pasados. Este paso debe hacerse de la manera más cuidadosa y organizada posible evitando la redundancia y ambigüedad en los datos.
5. Limpieza de datos: en este paso se busca eliminar cualquier dato erróneo, duplicado, incompleto, redundante o ambiguo en el conjunto de datos. Para esto se debe tener en cuenta que todos los datos deben ser específicos y cuantificables, además, cada grupo de datos debe tener un tipo especificado, ya sea número, texto, porcentajes, fechas, entre otros, para que sean datos que la computadora pueda entender y calcular.
6. Seleccionar el método de análisis que se aplicará: una vez teniendo los conjuntos de datos limpios y listos para el análisis, se selecciona el método que se utilizará para analizarlos. Existen diversos métodos dependiendo del objetivo, tales como minería de datos, inteligencia de negocios, visualización de datos, entre otros.
7. Análisis de datos: sea cual sea el método que se haya elegido para el análisis, el objetivo final será transformar los conjuntos de datos en información útil y entendible que permita a

los analistas extraer las respuestas planteadas al inicio. Un ejemplo muy común suelen ser tablas resumidas o gráficas, en el caso de la visualización de datos.

8. Interpretación de resultados: una vez realizado el análisis, se procede a plantear las conclusiones y extraer las respuestas a las preguntas planteadas. Si los datos logran responder dichas preguntas, la investigación se considera completa y los datos, definitivos.

Figura 35

Proceso de implementación de Big Data Analytics



4.3 Algoritmo de árbol de decisión (Decision Tree)

Un árbol de decisión es un algoritmo de *machine learning* que se utiliza en problemas de clasificación y decisión. Consiste, básicamente, en crear una estructura jerárquica y secuencial que modela todos los posibles caminos existentes. Este modelo jerárquico está definido por una serie de preguntas, donde cada una conducirá a cierta etiqueta o valor una vez que se haya aplicado una

observación, hasta llegar a una etiqueta final que representa la respuesta a la pregunta o problema inicial; la idea es tomar la mejor decisión entre todas las posibles. Existen dos tipos de árboles de decisión: por una parte, están los de clasificación, los cuales utilizan valores discretos tales como variables etiquetadas; por otro lado, están los árboles de regresión, que utilizan valores continuos, típicamente numéricos (Cooper, 2018).

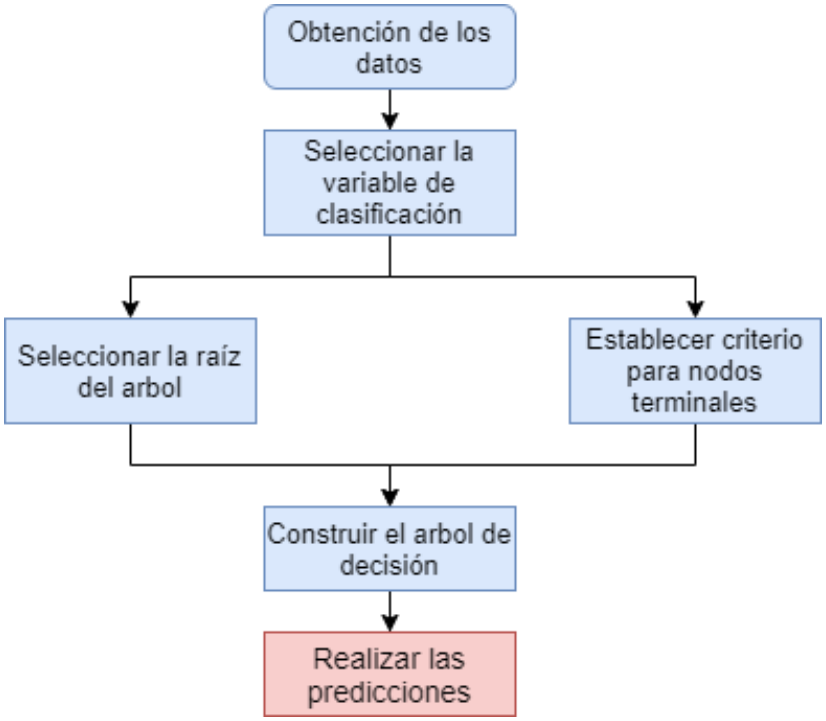
A continuación, se muestran los pasos para la creación de un árbol de decisión:

1. Obtención de los datos: tal como en cualquier método de machine learning, se debe partir de un conjunto de datos limpio y depurado. Estos conjuntos suelen ser una agrupación de características o atributos que describen a un determinado objeto. Estos datos también pueden ser numéricos, tales como precios, medidas dimensionales, peso, tiempo, entre otros números asociados con un objeto o clase en particular.
2. Seleccionar la variable de clasificación: cuando se construye el árbol, hay varias maneras de obtener la mejor división para cada nodo, por ejemplo, se puede considerar el Error de Clasificación, el índice Gini o la Entropía. El valor de esta variable para cada nodo es el que determinará de qué manera se dividirán los caminos del mismo.
3. Seleccionar la raíz del árbol: esto puede hacerse utilizando la variable de clasificación sobre todo el conjunto inicial para obtener la primera división.
4. Crear nodos terminales: es importante decidir qué tan profundo será el árbol; esto dependerá de la cantidad de datos en el conjunto: usualmente, cuando la cantidad de datos es muy grande, se utilizan criterios para detener la creación del árbol en cierto punto y que así no se convierta en una estructura de datos demasiado pesada. Dichos criterios pueden ser la profundidad del árbol (cantidad de niveles del mismo) o la cantidad de nodos totales. De esta forma, cada vez que se va a crear un subconjunto hay que evaluar el criterio elegido para decidir si se sigue dividiendo o si se ha llegado a un nodo terminal. Cuando ya no es posible seguir dividiendo el árbol, ya sea por profundidad máxima o cantidad de nodos máximo, se debe tener un criterio para seleccionar el dato que quedará como predicción final, usualmente se selecciona el dato más común entre las opciones posibles.

5. Construir el árbol de decisión: esto se puede hacer mediante la aplicación de un algoritmo recursivo donde, dado un conjunto de registros de entrenamiento de un nodo, si pertenecen a la misma clase, se considera un nodo terminal, pero si pertenecen a varias clases, se dividen los datos en subconjuntos más pequeños en función de la variable de clasificación seleccionada. El proceso se repite hasta que se hayan conseguido todos los nodos terminales.
6. Realizar las predicciones: estas pueden hacerse implementando un algoritmo recursivo que reciba una entrada y haga un recorrido del árbol evaluando cada uno de los nodos para elegir el camino adecuando dependiendo de los datos de entrada, y así sucesivamente hasta llegar a un nodo terminal, el cual tendrá la predicción final.

Figura 36

Metodología para la implementación del algoritmo de árbol de decisión



4.4 Algoritmos genéticos

Los algoritmos genéticos constituyen un método de optimización que imitan el proceso de la evolución natural al modificar una población de soluciones individuales. El método selecciona aleatoriamente algunos individuos de la población inicial para que sean “padres” y los usa para producir “hijos” para la siguiente generación. Con el paso de varias generaciones, se llega a una solución óptima debido a que en cada generación los individuos se van haciendo mejores (evolucionan) (Mokhatab et al., 2014).

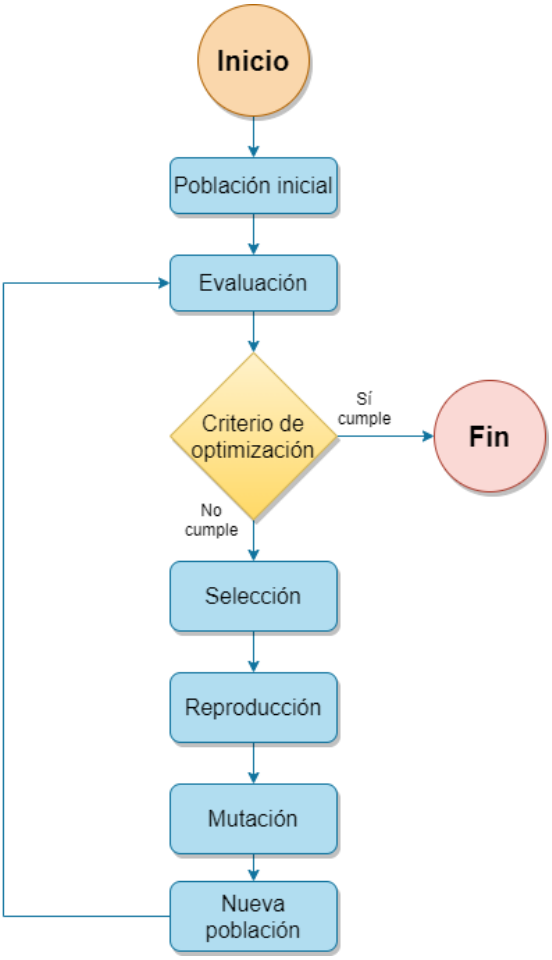
La implementación de un algoritmo genético tiene los siguientes pasos:

1. Generación de la población inicial: se genera el conjunto de individuos (soluciones), generalmente, en forma aleatoria.
2. Evaluación de los individuos: se aplica la función fitness para evaluar los individuos.
3. Criterio de optimización: se aplica un criterio de optimización que determina si el individuo, es decir, la solución, cumple con las condiciones deseadas. Si cumple, se detiene el proceso; si no, se procede al paso siguiente.
4. Selección de los individuos: los mejores individuos, según el paso anterior, son seleccionados.
5. Reproducción: se cruzan los individuos seleccionados mediante la función de cruce (crossover), lo cual genera nuevos individuos con características ligeramente diferentes basadas en las características de sus padres.
6. Mutación: se realizan pequeños cambios en las características de algunos de los nuevos individuos.
7. Nueva población: los individuos mutados constituyen un nuevo conjunto de soluciones que son, por lo general, mejores que las anteriores.

8. Se procede al paso 2.

Figura 37

Proceso de implementación de un algoritmo genético



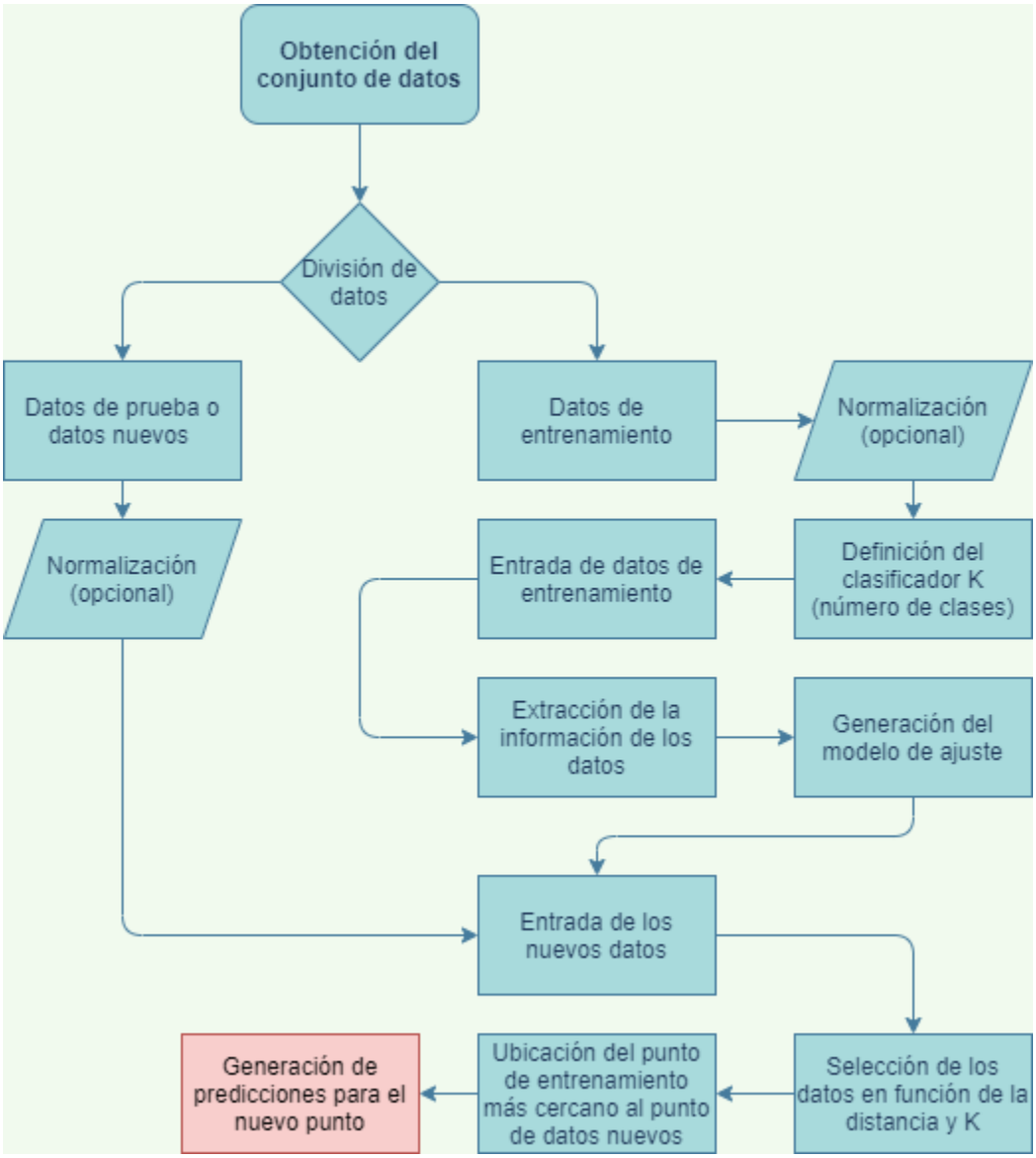
4.5 Algoritmo KNN (K-Nearest Neighbors)

Este algoritmo usa datos de entrenamiento para generar un modelo que hará una predicción para un nuevo conjunto de datos; esto lo hace encontrando un punto de entrenamiento (referencia) que esté más cerca del nuevo punto de datos. A continuación, se presentará la serie de pasos que sigue el algoritmo KNN (Müller y Guido, 2016):

1. Obtención de los datos: se recibe la información que va a ser tratada.
2. División de datos: generalmente, la información es agrupada en 2 categorías que son entrenamiento y prueba. A partir de los datos de entrenamiento se estructurará el modelo y con los datos de prueba se verificará su adecuado funcionamiento.
3. Normalización de datos: se puede aplicar si así se requiere. Lo que se busca es que todos los datos se integren fácilmente al modelo aplicándoles una serie de reglas.
4. Definición del clasificador K: se define como el número de clases o etiquetas donde se agrupan los datos; el usuario determina su valor.
5. Entrada de datos de entrenamiento: se ingresa el conjunto de datos con el objetivo de comenzar a estructurar el modelo.
6. Extracción de la información de los datos: se separan las principales características, tanto individuales como grupales, de los datos a analizar.
7. Generación del modelo de ajuste: se hacen las respectivas correcciones con la información de entrada y salida con el objetivo de calibrar el modelo para la entrada de nuevos datos.
8. Entrada de los nuevos datos: se ingresa la información de prueba o validación.
9. Selección de los datos: la selección de los mismos se hace con base en la distancia entre los datos analizados (K-vecinos) y el dato de referencia (entrenamiento), también se debe tener en cuenta el valor de K.
10. Ubicación del punto de entrenamiento más cercano al punto de datos nuevos: para llevar a cabo el proceso de predicción, es necesario ubicar el punto de entrenamiento que esté más cercano al nuevo punto de datos.
11. Generación de predicciones para el nuevo punto: se obtienen los resultados para los nuevos datos.

Figura 38

Proceso de implementación del algoritmo KNN (K-Nearest Neighbors)



4.6 Algoritmo ANFIS *(*Adaptive Neuro-Fuzzy Inference System)

La estructura de ANFIS se basa en el aprovechamiento de 2 grupos de parámetros: premisa y consecuencia. El objetivo de este algoritmo es encontrar dichos parámetros para obtener una correlación de los datos y a partir de ella generar los resultados.

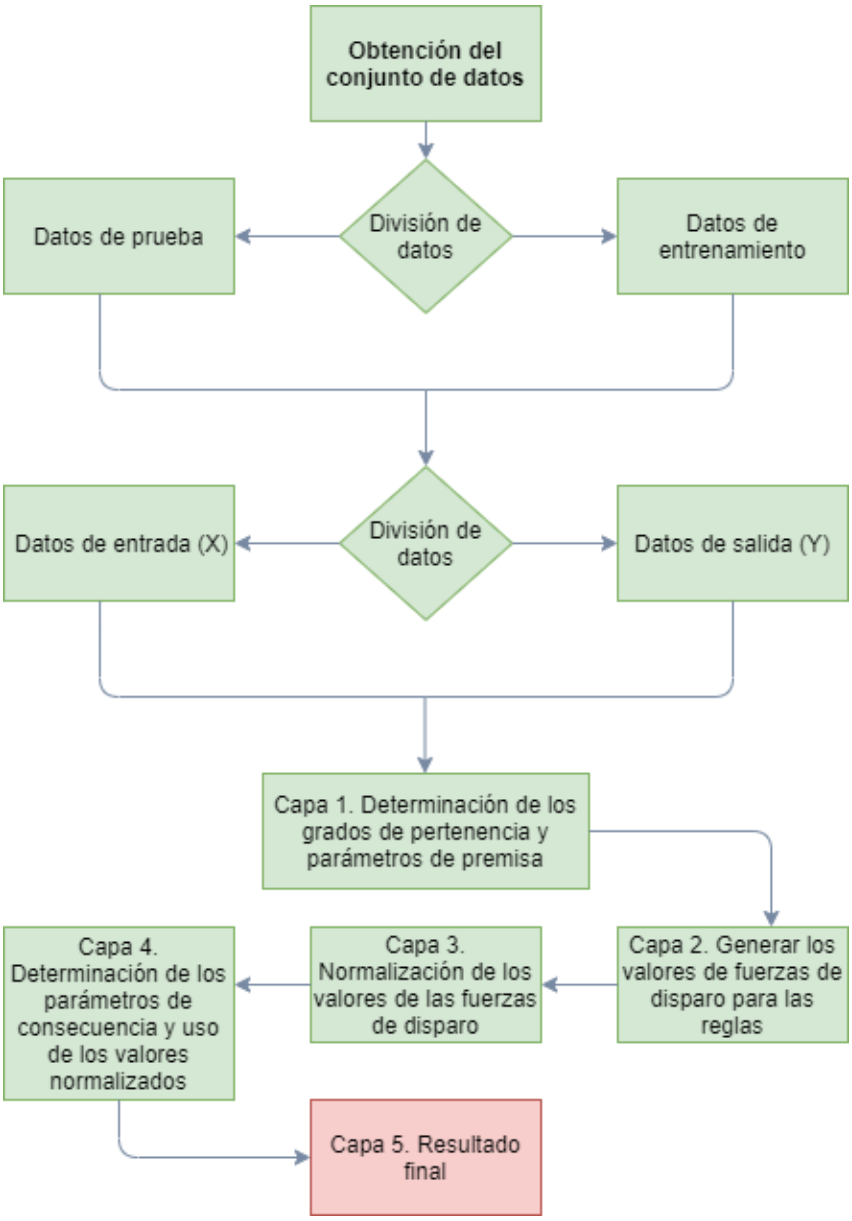
La secuencia que sigue este proceso es la siguiente (Karaboga y Kaya, 2018):

1. Obtención del conjunto de datos.

2. Primera división de los datos: en esta primera separación, los datos se agruparán en dos conjuntos: datos de entrenamiento y datos de prueba.
3. Segunda división de los datos: en la segunda separación que se hará al conjunto de entrenamiento y prueba por igual, los datos se separan en datos de entrada (X) y salida (Y).
4. Capa 1: también denominada capa de "fuzzyficación", se usa para determinar los grados de pertenencia a partir de los valores de entrada, pero para ello es necesario establecer los parámetros de premisa, que son los que determinan la forma de la función de pertenencia.
5. Capa 2: también denominada capa de regla, se usa para determinar las fuerzas de disparo para cada regla a partir de los valores de pertenencia calculados en la capa anterior.
6. Capa 3: también denominada capa de normalización, tiene como objetivo obtener un valor normalizado de las fuerzas de disparo calculadas en la capa anterior.
7. Capa 4: también denominada capa de "desfuzzyficación", aprovecha los valores normalizados de la capa anterior y los parámetros de consecuencia (que son 1 más que los de premisa, es decir, si los de premisa son 3 para cada regla, los de consecuencia deben ser 4) para obtener los valores ponderados de las reglas.
8. Capa 5: también denominada capa de suma, es la que finalmente obtiene los resultados sumando las salidas obtenidas para cada regla en la capa anterior.

Figura 39

Proceso de implementación del algoritmo ANFIS (Adaptive Neuro-Fuzzy Inference System)



4.7 Algoritmo SVM (Support Vector Machine)

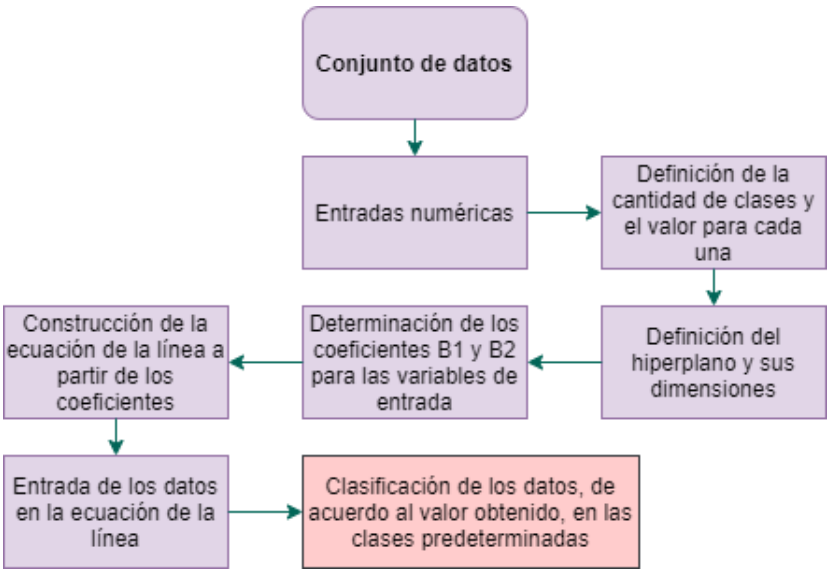
Es un algoritmo de aprendizaje automático que puede ser usado para clasificación, generando un modelo que prediga la clase de un dato o grupo de datos. También puede ser utilizado para procedimientos de regresión, aunque en este caso se le llama algoritmo “Support Vector Regression” (Regresión de vectores de soporte) y su implementación es similar.

A continuación, se muestran las fases de este algoritmo (Brownlee, 2017):

1. Obtención del conjunto de datos: se obtiene la información que se desea procesar.
2. Definición de las entradas numéricas: este modelo sólo acepta valores numéricos, en caso de ser una entrada categórica, esta debe tener un valor asociado para ser procesada.
3. Definición de la cantidad de clases: el usuario debe especificar cuántas clases tendrá el algoritmo y asociar un valor a cada una de ellas.
4. Determinación de los coeficientes B1 y B2: estos coeficientes son parte de la ecuación de una línea recta y se determinan a partir de los datos de entrada.
5. Construcción de la ecuación de la línea: a partir de los coeficientes obtenidos en el paso anterior, se determina la ecuación de la recta que es la que permite estimar la distancia que hay entre las clases.
6. Entrada de los datos en la ecuación de la línea: con los coeficientes y los datos de entrada se obtiene un valor que hace referencia a uno de los valores asociados a las clases o categorías.
7. Clasificación de los datos: finalmente, cada dato es clasificado de acuerdo al valor obtenido y a los valores asociados a cada clase.

Figura 40

Proceso de implementación del algoritmo SVM (Support Vector Machine)



4.8 Algoritmo de regresión logística

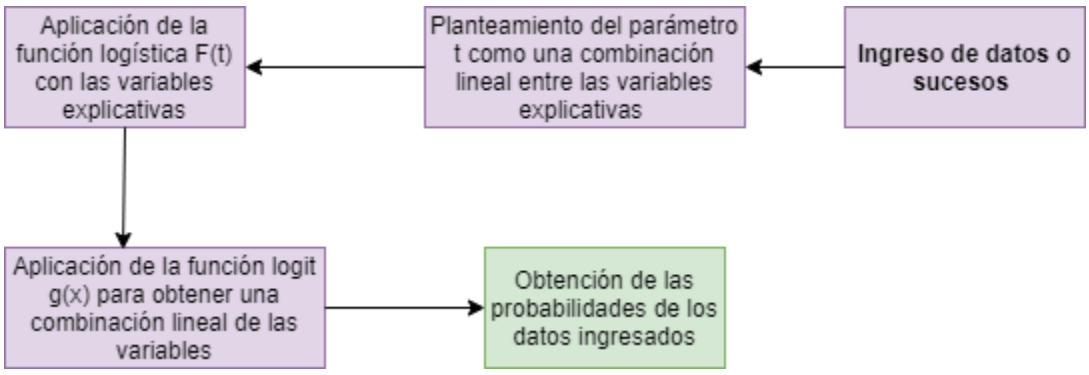
Se basa en obtener una variable de respuesta respecto a si un evento va a suceder o no. Tiene un enfoque predictivo asociado con probabilidad, además de ser un método aplicable a datos que no presentan una distribución normal.

De acuerdo con (Hackeling, 2014), los pasos de este algoritmo son:

1. Ingreso de datos: se ingresan los eventos a ser analizados.
2. Planteamiento del parámetro t: este parámetro es una combinación lineal de las variables explicativas (β_0, β_x). Al calcular este parámetro se genera una normalización de los datos que garantiza que todos los cálculos de probabilidad realizados permanezcan en el intervalo de [0,1].
3. Aplicación de la función logística F(t): al ingresar el parámetro t del paso anterior es posible obtener los valores de probabilidad asociado a los datos de entrada.
4. Aplicación de la función logit g(x): con el objetivo de obtener de vuelta una combinación lineal de las variables explicativas, se aplica esta función que es el inverso de la función logística.
5. Obtención de las probabilidades de los datos ingresados: finalmente, se obtienen de manera completa los datos ingresados relacionados con su respectiva probabilidad y las variables explicativas asociadas.

Figura 41

Proceso de implementación del algoritmo de Regresión Logística



4.9 Algoritmo de bosque aleatorio

Este algoritmo es una alternativa a la tendencia que presentan los árboles de decisión de sobreajustarse a los datos de entrenamiento y reducir la precisión de los resultados obtenidos. El bosque aleatorio es un conjunto de árboles de decisión que poseen ligeras diferencias entre sí y cada uno tiene una aleatoriedad asociada.

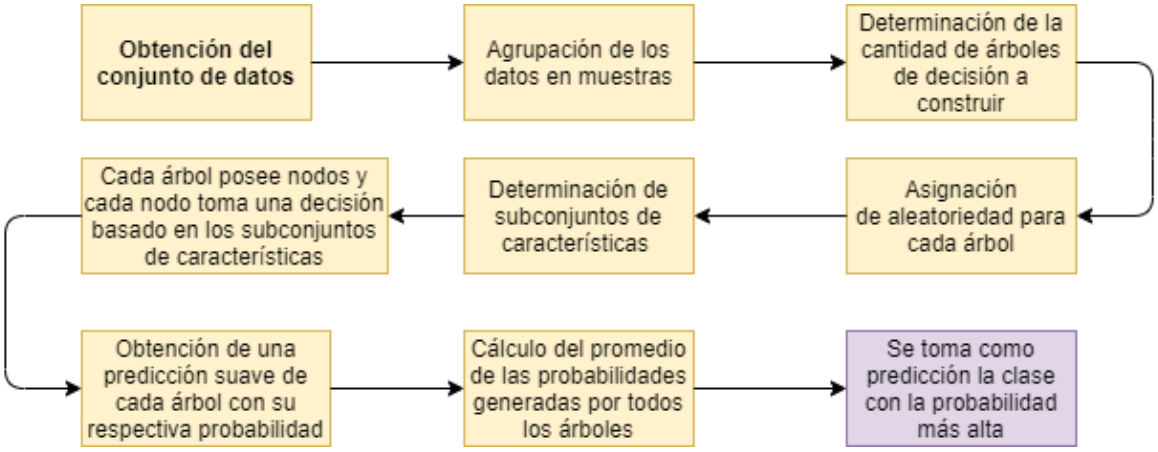
Para estructurar un bosque aleatorio, se deben tener en cuenta los siguientes pasos (Müller y Guido, 2016):

1. Obtención del conjunto de datos.
2. Agrupación de los datos en muestras: los datos se integrarán en diferentes conjuntos denominados “muestras”. En este caso, a pesar de que 2 muestras tengan los mismos datos, el orden de los datos dentro de las mismas las hará diferentes.
3. Determinación de la cantidad de árboles a construir: el usuario definirá cuántos árboles tendrá su bosque dependiendo del problema que esté abordando.
4. Asignación de aleatoriedad para cada árbol: cada árbol tendrá su estructura definida y su propia tendencia a elegir o no ciertas muestras, inclusive algunos usarán una misma muestra en repetidas ocasiones.
5. Determinación de subconjuntos de características: cada árbol tomará su propio subconjunto con las características que él prefiera. Gracias a la aleatoriedad asociada, los árboles tendrán diferentes subconjuntos y sus resultados también serán distintos.
6. Cada árbol posee nodos: estas partes se encargan de tomar decisiones, cada uno toma una decisión basada en el subconjunto de características elegido por el árbol. Un árbol posee varios nodos, por ende, cada árbol tendrá su propia gama de decisiones.
7. Obtención de una predicción suave: hace referencia a una predicción que tiene su propia probabilidad asociada de manifestarse.
8. Cálculo del promedio de las probabilidades: debido a que cada árbol genera su propia predicción suave, es necesario calcular un promedio para tener un valor más preciso y que también tenga en cuenta el aporte de todos los árboles.

9. Resultados: finalmente, se tomará como respuesta la predicción con la probabilidad promedio más alta.

Figura 42

Proceso de implementación del algoritmo de Bosque Aleatorio (Random Forest)



4.10 Algoritmo PCA (Principal Component Analysis)

El enfoque del PCA va dirigido a la reducción de la dimensionalidad, es decir, transformar las funciones en un espacio de menor dimensión para posteriormente ser procesadas.

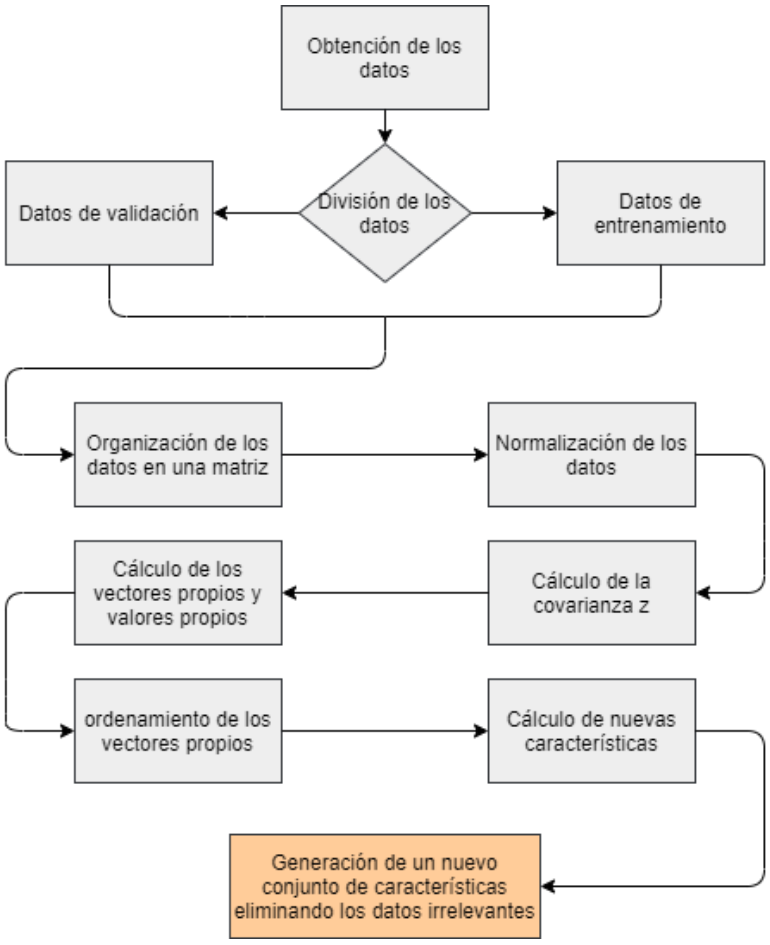
Para llevar a cabo este método se debe seguir la siguiente estructura (Tharwat, 2016):

1. Obtención de los datos.
2. División de los datos: se agruparán los datos en 2 grupos que son entrenamiento y validación. A ambos se le aplicarán los mismos procedimientos.
3. Organización de los datos en una matriz: para un mejor procesamiento, la información debe ser tabulada en una matriz, generalmente de 2 columnas.
4. Normalización de datos: proceso necesario para evitar datos redundantes o irrelevantes.
5. Cálculo de la covarianza: se estructura una matriz simétrica de covarianza donde se representa la desviación de cada variable del valor promedio.

6. Cálculo de los vectores propios y los valores propios: la matriz de covarianza estructurada en el inciso anterior es resuelta a través del cálculo de los vectores propios y valores propios.
7. Ordenamiento de los vectores propios: los vectores propios deben organizarse de manera matricial.
8. Cálculo de nuevas características: se efectúa a través de la creación de nuevas proyecciones.
9. Generación de un nuevo conjunto de características: finalmente, en el conjunto de las nuevas características o proyecciones se eliminan los datos irrelevantes y se reduce la robustez de las funciones.

Figura 43

Proceso de implementación del algoritmo PCA (Principal Component Analysis)



4.11 Algoritmo K-means

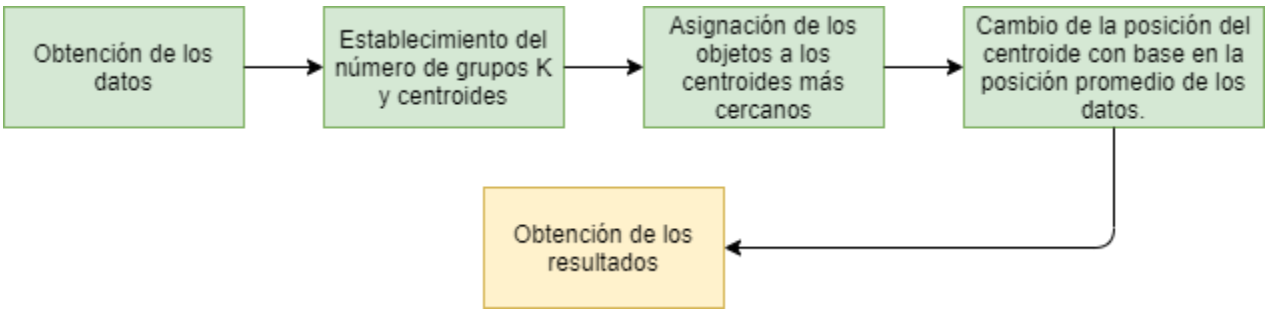
Es un algoritmo de clasificación o agrupamiento que consiste en dividir los datos en K grupos (previamente definidos) y procesarlos.

A continuación, se presentan los pasos a seguir (Pascual et al., 2007):

1. Obtención de los datos.
2. Establecimiento del número de grupos K y centroides: para cada grupo se define un centroide.
3. Asignación de los objetos a los centroides más cercanos: los centroides sirven como referencia para ubicar los puntos en la clase de acuerdo con el que esté más cercano.
4. Cambio de posición del centroide: es un proceso iterativo, es decir, se repite el proceso de cambio de posición de centroide hasta que ya no haya un cambio en los grupos al realizar la iteración.
5. Generación de los resultados: se muestran los datos clasificados.

Figura 44

Proceso de implementación del algoritmo K-means



5. Casos documentados de la aplicación de las técnicas de la ciencia de datos en la industria petrolera

De acuerdo con lo expuesto en capítulos anteriores del presente trabajo de investigación, se ha evidenciado que el uso de técnicas de la ciencia de datos supone importantes beneficios a aquellas industrias o sectores donde son aplicadas. Dentro de los mencionados beneficios, los más destacables son: la reducción del tiempo invertido, la disminución de costos, una mayor precisión en los cálculos, la posibilidad de utilizar paquetes muy grandes de información y que provienen de distintas fuentes y la automatización de diversas metodologías que se lleven a cabo, entre otros.

Este enfoque de la informática es relativamente nuevo, sin embargo, debido al gran campo de aplicación que maneja, los investigadores lo han empleado en distintos procesos, que van desde el diagnóstico y la simulación hasta la optimización de equipos y prácticas, con el fin de mejorar todos los sistemas o medios que provean o requieran información en su funcionamiento.

En este último capítulo se presentarán los resultados de varios estudios realizados por diferentes autores sobre la aplicación de técnicas de la ciencia de datos en diversas áreas de la industria petrolera, tales como la exploración, perforación y la producción.

5.1 Análisis del impacto de las propiedades reológicas del lodo en la tasa de penetración mediante minería de datos

(Al-Hameedi, Alkinani, Dunn-Norman, Flori, et al., 2019) realizaron un estudio enfocado en la tasa de penetración (ROP), que es uno de los factores más importantes en la perforación y, por ende, tiene un impacto directo en los costos de la misma. La ROP se puede definir como la distancia que avanza la broca mientras penetra la formación por unidad de tiempo o como “el volumen removido de roca por unidad de área por unidad de tiempo”. Por lo tanto, este parámetro es una medida del avance del taladro hasta llegar al yacimiento y a su vez tiene un efecto

significativo en los tiempos no productivos. Es importante resaltar que hay varios factores que afectan la ROP, entre los cuales destacan las propiedades reológicas del fluido de perforación que, por su parte, tienen la posibilidad de ser modificadas o controladas con la finalidad de precisar su impacto en la tasa de penetración y poder optimizarla, todo esto a través del uso de la técnica de minería de datos, que es el objetivo del presente trabajo.

Para lograr dicho fin fue necesario realizar un análisis de sensibilidad y estadístico para un conjunto de datos considerable (información de más de mil pozos perforados en el área de Basora, en Irak). Estos campos representan aproximadamente el 83% de la producción total del país, convirtiendo esta zona en un lugar de alto interés económico donde cualquier método que se aplique para resolver un problema o mejorar un proceso representa una ganancia económica notable. En la Tabla 11, se presentan los pozos que se tuvieron en cuenta para este estudio:

Tabla 11

Número de pozos de cada campo petrolero para este estudio

Nombre del Campo	Número de Pozos Usados
Rumaila Sur	300
Rumaila Norte	300
Nahur Umar	30
Luhais	75
Tuba	80
Sindbad	20
Ratawi	50
Zubair	150
Qurna Oeste 1	100
Qurna Oeste 2	50
Pozos totales	**1155**

Nota. Tomado de (Al-Hameedi, Alkinani, Dunn-Norman, Flori, et al., 2019).

Los parámetros que influyen en la ROP son el contenido de sólidos (SC), el peso del lodo (MW), el caudal (Q), la viscosidad plástica (PV), *strokes* por minuto (SPM), punto de fluencia (Yp), fuerza de gel, densidad equivalente de circulación (ECD), peso sobre la broca (WOB) y tamaño y geometría de las boquillas. Estas variables tienen la ventaja de poder ser modificadas antes de la perforación, pero presentan la desventaja de que no se puede alterar una propiedad sin afectar las demás.

En la presente investigación se recopilaron datos de informes diarios de perforación (DDR), registros de lodo, información geológica y de completamiento. Además, toda esta recopilación se combinó en un solo conjunto para realizar el respectivo procesamiento previo.

En la Tabla 12 se presenta la combinación de los datos utilizados en los 1155 pozos tabulados en intervalos con su correspondiente desviación estándar:

Tabla 12

Propiedades de los datos utilizados en este estudio

Parámetro	Mínimo	Máximo	Desviación Estándar
PV (cP)	6	29	3,41
Yp (Lb/100 pies)	11	30	4,45
SC (%)	2	10	1,86
ROP (m/h)	2	13	2,49

Nota. Tomado de (Al-Hameedi, Alkinani, Dunn-Norman, Flori, et al., 2019).

El procedimiento basado en Data Mining que se lleva a cabo es:

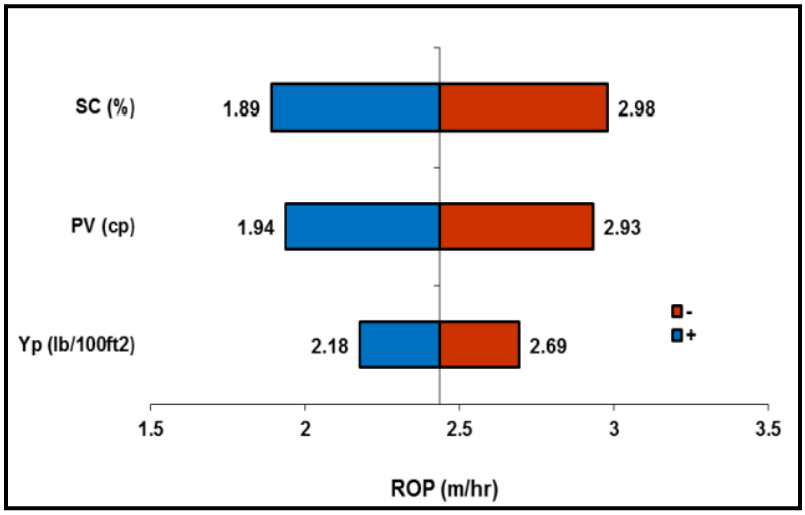
1. Se realiza un análisis descriptivo con el fin de entender el efecto del contenido de sólidos, la viscosidad plástica y el punto de fluencia (yield point).
2. Se eliminan los valores atípicos (valores por fuera del intervalo evidenciado en la tabla 12) con el uso de diagramas de caja.
3. Se determina si la distribución de los datos es normal o no.

4. Si los datos no presentan una distribución normal es necesario convertirlos o normalizarlos.
5. En el caso que no sea posible convertir los datos a una distribución normal, se deben aplicar las correlaciones de Kendall, Spearman o Hoeffding (presentan limitaciones).
6. Se utiliza el método de correlación de Pearson para distribuciones normales para calcular el coeficiente de correlación (CC) con el fin de determinar la relación o dependencia de cada parámetro (SC, Yp & PV) con la ROP.

Después del anterior tratamiento, realizaron el respectivo análisis de sensibilidad y obtuvieron la gráfica mostrada en la Figura 45:

Figura 45

Análisis de sensibilidad para cada parámetro



Nota. En la ilustración se aprecia que el parámetro que tiene mayor impacto en la ROP es el contenido de sólidos (SC). Tomado de (Al-Hameedi, Alkinani, Dunn-Norman, Flori, et al., 2019).

Los autores concluyeron que las propiedades reológicas del lodo tienen un impacto negativo en la ROP (ya que la disminuyen y lo que generalmente se busca es aumentarla o mantenerla), donde el contenido de sólidos tiene la mayor influencia. La viscosidad plástica y el

punto de fluencia también presentan una relación inversa con la tasa de penetración, pero con un menor impacto. Finalmente, se recomienda monitorear continuamente estos parámetros y mantener el pozo limpio para que el contenido de sólidos permanezca en el valor mínimo.

5.2 Interpretación de registros de pozos mediante redes neuronales de aprendizaje profundo

Los registros de pozos son una herramienta que permite evaluar formaciones y yacimientos para reducir la incertidumbre respecto a las propiedades de la roca y a los fluidos como petróleo, gas y agua. Con base en lo anterior, se puede esperar un volumen de datos considerable, además de que su interpretación se vuelve una tarea bastante compleja que necesita de ingenieros altamente experimentados para convertirlos en información valiosa que, utilizada de manera adecuada, puede mejorar, optimizar o reducir gastos en procesos de perforación, producción, completamiento, mantenimiento de pozos, entre otros.

En el presente trabajo, (Gupta y Soumya, 2020) plantean el uso de la inteligencia artificial basada en la aplicación de una red neuronal artificial para proporcionar un método más preciso, económico y eficiente en el área de la interpretación de registros. Una ventaja de esta técnica es la posibilidad de encontrar una solución a los problemas de análisis de registros, aunque algunos datos no estén disponibles inicialmente.

Las redes neuronales son sistemas de reconocimiento, transmisión y predicción de información basados en el funcionamiento del cerebro humano, por lo tanto, es necesario un proceso de retroalimentación o entrenamiento para que estas funcionen de manera adecuada y no se presenten errores inaceptables.

Para la aplicación de esta técnica se debe realizar el siguiente procedimiento:

1. Normalizar los datos de entrada: debido a que esta información viene de diferentes tipos de registros, es necesario normalizarlos para que la red pueda integrarlos y

analizarlos como un conjunto. En el presente estudio se desea identificar la litología de las formaciones examinadas (datos no numéricos), por lo tanto, es necesario definir un valor para cada clasificación dentro de la red (clases de numeración), por ejemplo, para la arenisca con contenido de agua se le asigna el valor de 1, al esquisto el valor de 2, etc. Los registros utilizados fueron: gamma-ray, de resistividad, de neutrones y de densidad. Para el cálculo de porosidad no sólo se emplearon los datos de los 4 registros normalizados, además fueron necesarios valores reales de las técnicas de neutrones y densidad.

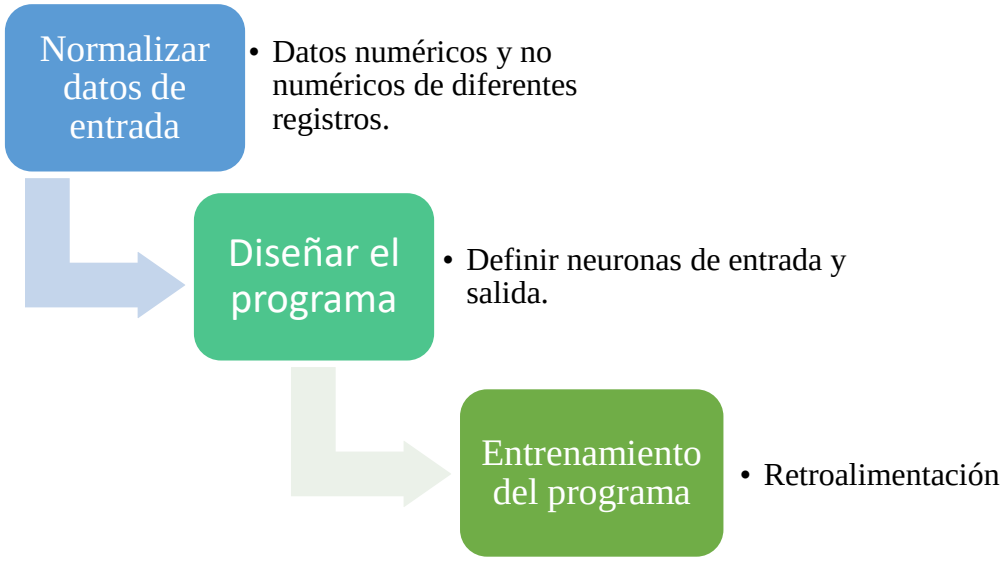
2. Diseñar o definir el programa: este programa puede descifrar patrones o relaciones y, además, modelar funciones complejas. También, se deben definir las neuronas de entrada que son los valores de los registros gamma-ray (I1), resistividad (I2), neutrones (I3), densidad (I4), y los valores reales de neutrones (I5) y densidad (I6); así mismo, las neuronas de salida que son litología (O1) y porosidad (O2).
3. Igualmente, se deben definir los pesos para la información de entrada. En el caso de la litología, que no depende de los datos reales de los registros neutrones y densidad, de esa manera el peso asignado para I5 e I6 debe ser cero. Para el caso de la porosidad, que sólo depende de I5 e I6, el peso asignado para I1, I2, I3 e I4 debe ser cero.
4. Entrenamiento del programa: como se especificó previamente, es necesario retroalimentar el programa y para ello es preciso dividir aleatoriamente el conjunto de datos (70% para entrenamiento y 30% para la prueba). También es importante definir tres factores de formación que son tasa de aprendizaje, número de iteraciones y porcentaje de error. Cuanto más grande sea el volumen de datos de entrenamiento, mayor será la calidad de los resultados. Para evitar errores de memorización de

soluciones por parte del programa, se debe reservar una parte de los datos para usarse como conjunto de validación.

En este trabajo, el error aceptable de litología fue de 0,1% y el de porosidad fue 0,01%.

Figura 46

Proceso de las redes neuronales en la interpretación de registro de pozos



Nota. Adaptado de (Gupta y Soumya, 2020).

Los autores concluyeron que los resultados obtenidos fueron de alta precisión, lo que se manifiestan como una mejor predicción de la litología, con el valor agregado de una reducción del tiempo invertido en todo el proceso. Adicionalmente, el cálculo de la porosidad depende de la calidad de los registros. Gracias a esto, se cumple el objetivo del proyecto que es la disminución del error humano y la optimización del tiempo. Se proyecta para un futuro cercano que se puedan estimar más propiedades con el uso de más información, por ejemplo, la utilización del registro SP para calcular volúmenes.

5.3 Predicción de pérdidas de volumen en formaciones naturalmente fracturadas utilizando inteligencia artificial

(Al-Hameedi, Alkinani, Dunn-Norman, Al-Alwani, et al., 2019) definen las pérdidas de circulación como la cantidad de lodo que se introduce en una o más formaciones geológicas sin posibilidad de retorno, dando como resultado una disminución del volumen de circulación, lo cual se asocia con problemas financieros, riesgo de patadas o reventones, daño a la formación, problemas de colapso, imprecisiones en los registros, aumento del tiempo no productivo, entre otros.

Los métodos que se pueden aplicar contra estos problemas de pérdidas se clasifican en preventivos, correctivos y avances de tecnología de perforación. Es indispensable controlar o minimizar esta problemática y, generalmente, se logra a través del ajuste de parámetros de perforación y propiedades de los fluidos, es decir, se maneja un enfoque predictivo para lo cual la mejor opción es la creación de modelos a través de la inteligencia artificial (Machine Learning).

Para la aplicación de esta técnica se generarán relaciones matemáticas cuya finalidad es pronosticar las pérdidas de fluido de perforación, la densidad de circulación equivalente (ECD) y la tasa de penetración (ROP). Con el fin de lograr estimaciones precisas de los parámetros mencionados, los datos a utilizar para este trabajo provienen de 500 pozos perforados en el campo Rumalia, en Irak, y se presentan en la Tabla 13:

Tabla 13

Datos de entrada

Parámetro	Mín	Máx	Desviación estándar
Peso del lodo (MW) (g/cc)	1,12	1,17	0,01
Densidad de circulación equivalente (ECD) (g/cc)	1,13	1,19	0,01
Viscosidad plástica (PV)	9	21	3,19

Parámetro	Mín	Máx	Desviación estándar
Punto de fluencia (Yp) (Lb/100 ft2)	13	32	5,33
Tasa de penetración (ROP) (m/h)	3	17	3,31
Revoluciones por minuto (RPM)	60	110	11,34
Peso de la broca (WOB) (Ton)	5	22	3,55
Caudal (Q) (L/min)	1760	2816	271,70
ΔP pérdidas (Kg/cm2)	1,6636	9,029	1,46
Boquillas TFA (pulg2)	0,45	10,6	2,35
Pérdidas de lodo (m3/h)	0	88	24,56

Nota. Tomado de (Al-Hameedi, Alkinani, Dunn-Norman, Al-Alwani, et al., 2019)

A continuación, se describe el proceso de modelamiento basado en Machine Learning avanzado que utilizaron los autores del presente estudio:

1. Metodología de regresión: en este inciso se plantea el uso del método de mínimos cuadrados parciales (PLSR), un tipo de regresión especial que se utiliza para crear modelos aprovechando los datos de entrada y los resultados de salida, todo esto a través de la generación de factores (componentes principales).
2. Elegir los factores numéricos: para asegurar un buen desempeño en el modelo, es necesario encontrar el número ideal de factores, por lo tanto, se definieron 2 criterios. El primero es la raíz media de la suma de cuadrados residual predicha (PRENSA), que debe ser minimizada, y el segundo es la existencia de una relación entre las variables “X” e “Y” manifestada en cada gráfico de puntuación. Ambos criterios se deben cumplir para la correcta estimación del número de factores.
3. Elegir las variables: es posible realizar esta acción gracias al tipo de regresión implementado (LPSR) que permite llevarlo a cabo en función de la importancia de la variable de proyección (VIP), la cual se puede calcular para cada parámetro y debe ser superior a 0,8 (en caso contrario la variable será eliminada).

4. Análisis de sensibilidad: se aplica con el objetivo de cuantificar y entender el efecto de cada parámetro en el modelo. Al modificar una variable, se analiza el cambio generado en los resultados; para esto se decidió aumentar o disminuir cada parámetro en un 10%.

(Al-Hameedi, Alkinani, Dunn-Norman, Al-Alwani, et al., 2019) obtuvieron resultados satisfactorios gracias a los 3 modelos predictivos que se desarrollaron, los cuales, a su vez, fueron validados con datos nuevos de 30 pozos demostrando una excelente correspondencia con los datos reales para pérdidas de lodo, ECD y ROP. También se hizo una comparación con otros modelos que ya habían sido desarrollados donde los resultados generados por inteligencia artificial. Con el análisis de sensibilidad fue posible determinar que, para minimizar las pérdidas de fluido de perforación, es imprescindible modificar el peso del lodo (MW). La tasa de penetración, por su parte, depende de peso sobre la broca (WOB) y de la viscosidad plástica (PV), además de que las boquillas (TFA) tienen una influencia negativa sobre la ROP.

5.4 Detección temprana de patadas de pozo mediante minería de datos

Durante la perforación, uno de los peores escenarios que se pueden presentar son los reventones, que se definen como el flujo incontrolado dentro del pozo que puede ir hacia superficie o hacia una formación expuesta con menor presión. Debido a su magnitud, este problema debe ser tratado con un enfoque preventivo, el cual se basa en la detección de las patadas (aumentos anormales de presión de los fluidos de formación) para seguir los protocolos establecidos y recuperar el control del pozo. Entre más precisas y rápidas sean las predicciones de estos amagos de reventón, mayor será el tiempo que tendrán los trabajadores para controlar esta anomalía, lo cual es una ventaja para mantener la perforación bajo control logrando un desempeño seguro y óptimo.

Para llevar a cabo dicho proceso predictivo, (Alouhali et al., 2018) proponen la técnica de minería de datos para tratar el enorme conjunto de datos recopilados y generados en tiempo real que, además, provienen de distintos lugares del pozo, dependiendo de la ubicación de los sensores instalados. Para la aplicación exitosa de la minería de datos, es necesario refinar la información, es decir, clasificar los sucesos (instancias) registrados en “Kick” o “No Kick” con el objetivo de generar un conjunto de datos de entrenamiento. Se parte de una agrupación de más de un millón de instancias y después de dicho refinamiento se reduce a aproximadamente 120000, con el beneficio de que se eliminaron los valores atípicos y se mejoró la productividad de la técnica.

Con base en lo anterior, se procede a aplicar la minería de datos de la siguiente manera:

1. Lo primero es definir los criterios de evaluación. Para este trabajo se seleccionaron 5 modelos, los cuales son: árbol de decisión, K-Nearest Neighbors (KNN), algoritmo de optimización mínima secuencial (SMO), red neuronal artificial (ANN) y la red bayesiana. Con el fin de visualizar una comparación más profunda, se delimitaron 4 parámetros:
    a. Verdadero positivo (TP): instancias clasificadas correctamente como “Kick” (patada).
    b. Falso positivo (FP): instancias clasificadas incorrectamente como “Kick” (patada).
    c. Verdadero negativo (TN): instancias clasificadas correctamente como “No Kick” (no patada).
    d. Falso negativo (FN): instancias clasificadas incorrectamente como “No Kick” (no patada).

2. Posteriormente se debe cuantificar el rendimiento y para ello se establecen las siguientes métricas:

    - Precisión: se basa en la relación de los valores positivos reales entre todos los valores analizados.
    - Recuerdo: se centra en las instancias relevantes y es definido como los sucesos positivos reales entre todos los sucesos revisados.
    - Medida o puntuación F: es un valor que permite relacionar y unificar las 2 métricas anteriores.
    - Coeficiente de correlación de Mathews (MCC): a partir de una matriz donde se tienen en cuenta los 4 parámetros previamente mencionados, se calcula una medida entre lo real y lo estimado, donde 1 significa coincidencia perfecta, -1 denota que no hay coincidencia y cero se traduce como predicción aleatoria.
    - Área ROC: aunque puede presentar imprecisiones para clases desequilibradas, es un buen indicador de precisión para un modelo, sobre todo si el valor de esta métrica es cercano a 1. El cálculo se efectúa a partir del área bajo la curva de la característica del operador del receptor (ROC).
    - Área PRC: la curva de recuperación de precisión (PRC) maneja la misma metodología que la métrica anterior, pero con diferencias tales como el enfoque y el hecho de que no presenta desventaja alguna frente a clases desequilibradas.
    - Estadístico Kappa: este valor se define como la comparación entre la medida esperada (obtenida estocásticamente) y la medida observada.

En la Tabla 14 se resume los resultados obtenidos para los 5 modelos:

Tabla 14

Valores de las métricas para cada modelo analizado

Métrica	Valor M1	Valor M2	Valor M3	Valor M4	Valor M5
Precisión	0,989	0,992	0,767	0,754	0,988
Recuerdo	0,972	0,987	0,919	0,51	0,852
Medida F	0,98	0,99	0,836	0,609	0,915
MCC	0,98	0,989	0,838	0,617	0,917
Área ROC	0,993	0,999	0,998	0,754	0,928
Área PRC	0,971	0,988	0,919	0,39	0,878
Estadístico Kappa	0,98	0,989	0,834	0,606	0,914

Nota. En la tabla se presentan los valores de las métricas para cada modelo donde M1 corresponde al modelo del árbol de decisión; M2 es K-Nearest Neighbors (KNN); M3 es red bayesiana; M4 es optimización mínima secuencial (SMO); M5 es red neuronal artificial (ANN). Tomado de (Alouhali et al., 2018).

Los autores pudieron concluir que el mejor algoritmo es el árbol de decisiones, ya que refleja valores altos en todas las métricas, además de contar con una alta precisión y presentar una menor complejidad en términos de potencia computacional requerida. Una alternativa viable es el uso del algoritmo KNN, ya que posee una precisión ligeramente mayor, pero es un poco más complejo y el tiempo de tratamiento de datos es más alto, junto a una mayor exigencia de hardware. A excepción de la optimización mínima secuencial, todos los modelos propuestos pueden aplicarse en la detección de patadas dando el tiempo suficiente para tomar las medidas preventivas.

5.5 Optimización del sistema de gas lift utilizando algoritmos genéticos

En la industria del petróleo una de las variables más importantes a mantener es la tasa de producción, la cual depende de la presión del yacimiento y, con el paso del tiempo, esta tiende a disminuirse, lo que a su vez reduce la tasa de producción, por lo que se hace necesario el uso de un método de levantamiento artificial para mantener el caudal de aceite y con ello la rentabilidad. La simulación es una herramienta poderosa para visualizar, analizar y optimizar cualquier proceso

de recuperación secundaria o terciaria, pero es imprescindible invertir tiempo y recursos, además de contar con la suficiente cantidad de datos para llevar a cabo dicha simulación.

Para llevar a cabo la optimización de este procedimiento existen varios métodos (por ejemplo, programación lineal, de pendiente, algoritmo de programación dinámica, entre otros), sin embargo, la mayoría de los métodos poseen diversas limitaciones y no tienen en cuenta factores importantes como la contrapresión ejercida. Por su parte, los algoritmos genéticos (GA), que también son un método de optimización, poseen una alta capacidad para integrar eficazmente la información y no son afectados por las limitaciones que tienen los otros métodos.

(AlJuboori et al., 2020) estudian el uso de algoritmos genéticos para optimizar el sistema de levantamiento artificial de *gas lift* en un campo del Medio Oriente. Su técnica propuesta parte de un conjunto de soluciones de las cuáles se generan nuevos grupos o generaciones que serán sometidas a un proceso de selección natural (concepto de biología evolutiva) hasta obtener la solución óptima global. Este proceso de aplicación se define a continuación:

1) Generación de la población inicial: para este caso, se generan 4 tasas de inyección de gas diferentes que serán tratadas como cromosomas, es decir, pasarán por fases de selección, cruce y mutación.
2) Evaluación de la función de aptitud: este criterio permite cuantificar la eficiencia de la solución dentro del procedimiento. En el actual contexto, el caudal de aceite es la representación de dicho concepto.
3) Finalización del proceso: en el caso de que los resultados cumplan con los requerimientos para la optimización, la técnica GA habrá culminado.

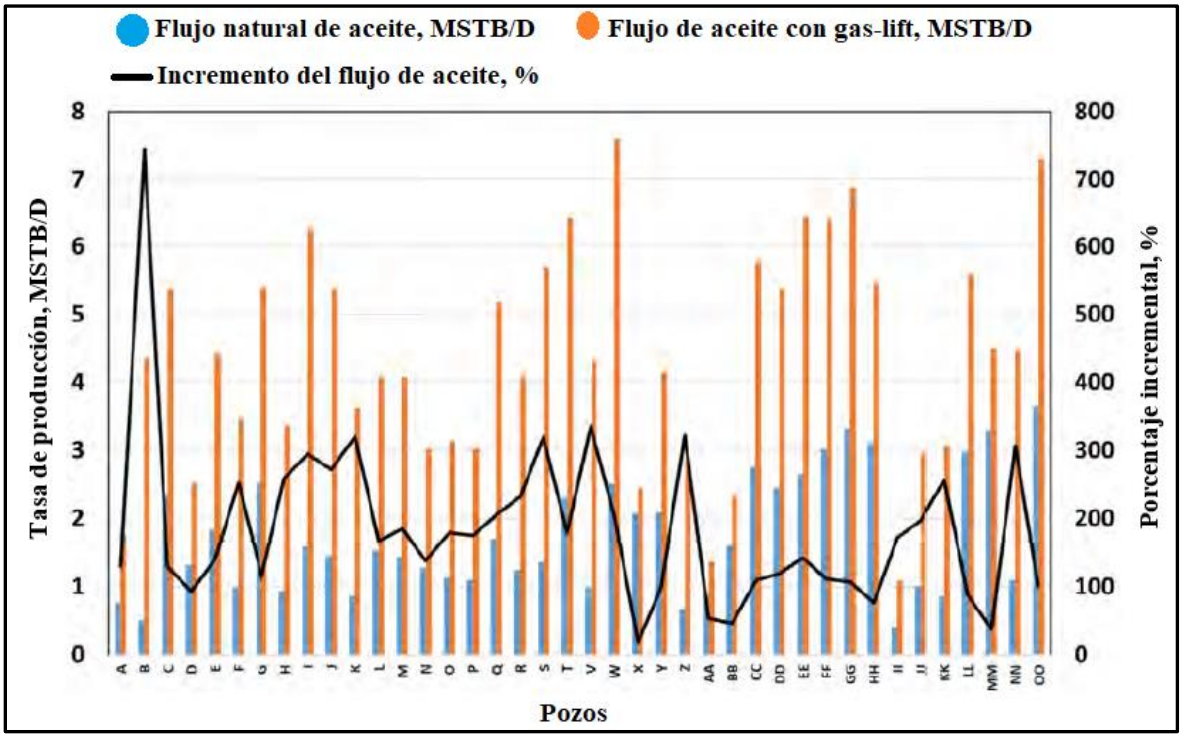
4) Si aún no se cumplen los criterios: es necesario seleccionar 2 cromosomas que serán cruzados (y en algunos casos mutados) para obtener una nueva generación con su respectiva desviación estándar.
5) Método iterativo: el paso anterior se repite cuantas veces se precise hasta llegar a un resultado óptimo global que se manifieste con un valor aceptable en la función de aptitud. Al cumplir dicha condición el proceso habrá finalizado.

Con base en lo anterior, se procede a realizar el modelado y la simulación numérica, para lo cual se consideró una red con sistema de inyección de *gas lift* para 43 pozos de un campo petrolero del Medio Oriente. Para esta sección se debe contar con información del yacimiento (reservas, presión, profundidad, si está o no asociado a un acuífero, etc.), tasa de producción promedio, propiedades de los fluidos, datos de completamiento, tuberías, índice de productividad, información PVT, entre otros; para obtener modelos de flujo de pozo, construcción de pozo, diseño de elevación de gas y construcción de campo. Además, se debe realizar un análisis de sensibilidad para el corte de agua y presión del yacimiento.

Los resultados se pueden visualizar en la Figura 47:

Figura 47

Resultados de la simulación para la tasa de producción de aceite para cada pozo



Nota. Adaptado de (AlJuboori et al., 2020)

Con el objetivo de determinar la viabilidad de todo el proyecto, se llevó a cabo un análisis financiero que, a su vez, complementa todo el proceso de mejoramiento rentable del campo.

Tras realizar el estudio, los autores concluyeron que la técnica de algoritmo genético permite modelar de manera eficiente un volumen considerable de datos para maximizar la producción de petróleo mediante la determinación de la tasa de gas a inyectar. A partir del análisis de resultados se afirma que, en pozos con corte de agua relativamente alto, el sistema de *gas lift* presenta mejores valores de recuperación de aceite. Finalmente, gracias al estudio económico, se logró establecer un beneficio neto respecto a la producción de petróleo antes y después de aplicar el levantamiento artificial, además de aumentar el ciclo de vida del campo.

5.6 Diagnóstico y predicción de problemas en sistemas de bombeo mecánico mediante el uso de Machine Learning

El bombeo mecánico es un método de levantamiento artificial ampliamente usado en la industria a nivel mundial. Este tipo de recuperación se basa en el uso de una bomba de varilla o de subsuelo reciprocante para generar la presión suficiente para que el fluido del yacimiento pueda desplazarse desde el fondo hasta superficie. Estos equipos son muy eficientes, sin embargo, pueden presentar fallas de varios tipos por distintas causas, por lo tanto, es necesario predecir todas estas problemáticas antes de que se manifiesten para aplicar un mantenimiento adecuado. Gracias a la información recopilada en superficie, es posible conocer el estado de la bomba a través del cálculo de las condiciones de fondo del pozo.

(Bangert, 2019), en su estudio, busca generar predicciones acertadas en ese sistema que, generalmente, utiliza un diagrama denominado tarjeta de dinamómetro. En un solo pozo, la cantidad de información no es grande, pero en el caso de varias perforaciones el volumen de datos generados aumenta, aspecto que complica un diagnóstico confiable y eficaz pero que puede ser alcanzado a través del Machine Learning. Para emplear esta técnica, primero se debe reducir la cantidad de datos mediante la ingeniería de características y, posteriormente, se procede a generar el modelo que debe ser retroalimentado con datos de entrenamiento y de prueba para asegurar resultados satisfactorios. A fin de clasificar y operar las entradas (caracterizar las tarjetas de dinamómetro) se pueden usar los siguientes métodos:

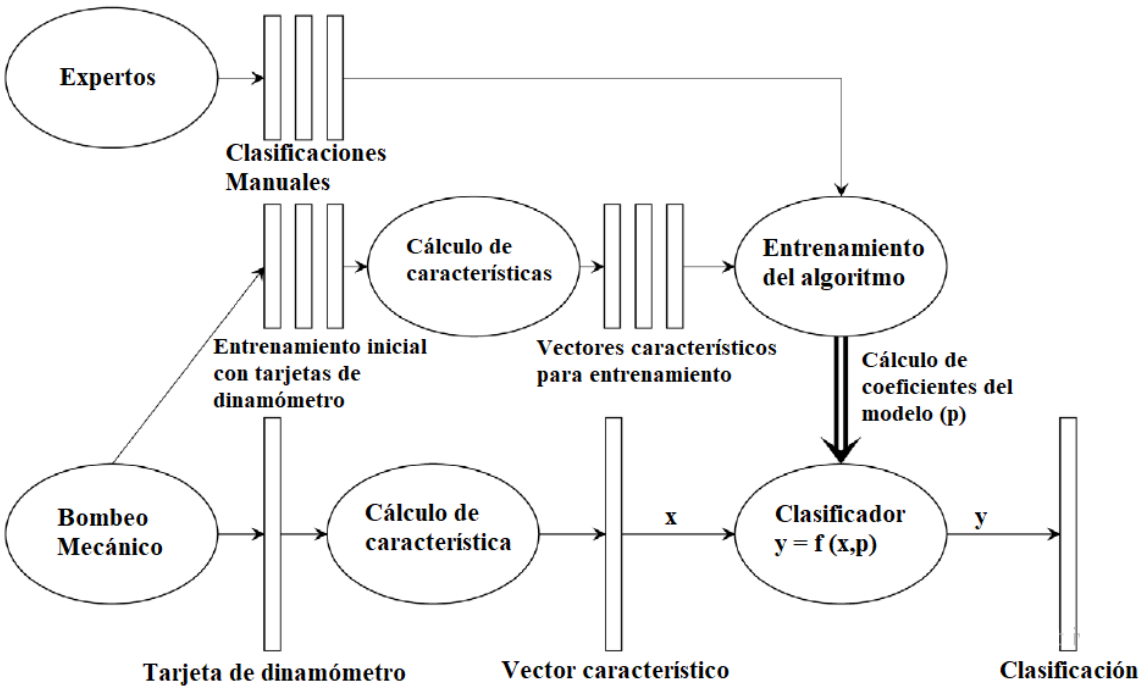
- Métodos basados en la biblioteca: se genera un conjunto de categorías conocidas y cada nueva entrada será clasificada de acuerdo a una función de distancia (métrica) donde se compara el valor de la nueva tarjeta con cada miembro de la biblioteca. Sus desventajas radican en que, para mediciones en tiempo real, los conjuntos deben ser pequeños y es

difícil elegir la función de distancia. Para este caso se emplearon 2166 tarjetas dando un error del 5%.

- Métodos basados en modelos: uno de estos es basado en la división en segmentos y la carga promedio de cada segmento (para 6101 tarjetas llegó a un error de 2,2%). Otro método es basado en el centroide, donde se calculan áreas para luego utilizar Machine Learning apoyado en una red neuronal (230 tarjetas de entrenamiento y 100 de prueba), logrando un error del 11%. La última opción es utilizar una serie de Fourier para descomponer la tarjeta (se manejaron 102) y se tuvo un error del 5%.
- Métodos basados en segmentos: se enfoca en determinar los puntos clave de las tarjetas (apertura y cierre). Para ello, existen 2 métodos, el código cadena (geometría pura) y otro fundamentado sobre aleatoriedad. Se utilizaron 6132 tarjetas clasificadas manualmente y se obtuvo un error de 24%. En otro caso se implementó el algoritmo de máquinas de vectores de soporte (SVM) y con un número bastante menor de entradas (88 de entrenamiento y 40 de prueba) se obtuvo un error del 2%.
- Otros métodos: para estos casos se formulan en términos de valores exactos para que puedan evaluarse en un sistema informático automatizado.

Figura 48

Elementos de un proceso de clasificación



Nota. Adaptado de (Bangert, 2019).

Después de una revisión exhaustiva, (Bangert, 2019) define la serie de Fourier de un momento combinada con las coordenadas del centroide como modelo elegido, con una tasa de precisión de 99,8 ± 0,1%. Con el objetivo de ser validado, este modelo debe pasar por cuatro fases específicas: recopilación de datos, generación de datos de entrenamiento, generación de funciones y aprendizaje automático. Finalmente se ejecuta la técnica con 35292 tarjetas clasificadas manualmente utilizando 85% de la información para entrenamiento y el 15% restante para prueba.

El autor afirma que el modelo es lo suficientemente preciso para aplicarlo en un campo real con la ventaja de generar una alerta automática ante cualquier anormalidad, ahorrándoles a los expertos el trabajo de monitorear y diagnosticar y permitiendo pasar directamente a la etapa de mantenimiento preventivo.

5.7 Uso de Machine Learning para mejorar la perforación direccional

La perforación direccional es un proceso que aborda una dificultad mayor respecto a la convencional (perforación vertical): mantener una tasa de penetración adecuada (ROP) junto a la conservación de una trayectoria óptima (ángulo). Cualquier desviación en el plan de perforación acarrea un aumento de costos inmediato. Con la intención de evitar estos desvíos, es necesario ejecutar acciones precisas y oportunas para llevar a cabo este procedimiento de manera eficiente, para lo cual, una alternativa a considerar es la implementación de la inteligencia artificial.

El aprendizaje automático o Machine Learning se aplica en este caso con el objetivo de mantener un sistema de perforación controlado y, a su vez, sostener una ROP máxima y minimizar las acciones correctivas de dirección. Para este estudio, el tipo de perforación aplicada es basada en motores de fondo de pozo, donde la sarta manejará 2 operaciones distintas (rotar y deslizar), además, los datos de entrada fueron obtenidos a partir de datos históricos y simulaciones.

Para este proyecto, (Pollock et al., 2018) implementaron un sistema de aprendizaje automático que entrena una red neuronal artificial, específicamente, un perceptrón multicapa. Lo que se busca es que el sistema logre emular la capacidad de un perforador direccional experto para tomar decisiones frente a sucesos nuevos o inesperados. Esta técnica automatizada requiere llevar a cabo las siguientes tareas: formulación de información (preparación y análisis de datos), construcción y evaluación de redes neuronales, simulación y retroalimentación y, de esta manera, ofrecer la posibilidad de asesorar a un perforador a través de la supervisión.

El primer paso consiste en dividir los datos recopilados en 3 grupos: entrenamiento, validación y prueba para alimentar y entrenar la red de manera adecuada, de tal forma que esta pueda partir de la información de entrada para llegar a los datos objetivo (acciones que tomó el

ingeniero de perforación frente a un problema determinado). Los resultados de la red se comparan con las decisiones que hubieran tomado los expertos para verificar la eficiencia de la misma.

El segundo paso consiste en interconectar el sistema con un programa de simulación y verificar su desempeño, es decir, cumplir con el objetivo propuesto minimizando las desviaciones, retrasos o problemas técnicos que se puedan presentar.

El tercer paso es realizar una demostración con guía humana directa, en otras palabras, el sistema se evaluará con un perforador direccional experto a partir de las veces que este se vea obligado a cambiar el curso de la perforación (realizar ajustes) y de su propio criterio para considerar si el programa es útil y confiable.

Como resultado de la implementación de la red neuronal artificial, fue posible predecir las decisiones de perforadores direccionales expertos con un error de sólo 3%, gracias a lo cual los autores pudieron concluir que el enfoque de aprendizaje automático utilizado posee un gran potencial como sistema de recomendaciones para perforadores direccionales y, cuando sea refinado y perfeccionado, podría ser utilizado en una plataforma de perforación totalmente automatizada.

5.8 Predicción del IPR en pozos verticales de aceite mediante el uso de inteligencia artificial

(Basfar et al., 2018) realizaron un estudio enfocado en el comportamiento de afluencia de los fluidos al pozo (IPR), el cual representa una poderosa herramienta para el área de producción y se basa en la creación e interpretación de una curva que relaciona la presión con el caudal de los fluidos producidos. A partir de este medio, se puede diseñar el equipo de producción y los sistemas de levantamiento artificial, además de poder definir el mejor esquema para la extracción de hidrocarburos. Un parámetro muy importante que se debe establecer es el índice de productividad (J), que se entiende como la capacidad de un pozo para producir, es decir, el cambio que se presenta

en el flujo volumétrico cuando se varía la diferencia de presión entre el pozo y el yacimiento. Es posible determinar el IPR mediante correlaciones empíricas tales como Vogel, Fetkovich, Klins-Clark, Wiggins, Elias y colaboradores, Khasanov y colaboradores, entre otras, sin embargo, estas tienden a ser bastante imprecisas y están limitadas a ciertas condiciones, por lo que surge la necesidad de buscar alternativas que permitan predecir el IPR de manera confiable y rápida para optimizar este proceso.

Con lo anterior en cuenta, (Basfar et al., 2018) propusieron dos soluciones basadas en la inteligencia artificial: red neuronal artificial (RNA) y la técnica de lógica difusa (FL). La primera se centra en el uso de neuronas (puntos de tratamiento y transmisión de información) para hallar relaciones con base en los datos de entrada. Por su parte, la segunda alternativa se enfoca en la creación de un modelo que permite pronosticar parámetros a través de la información de entrada y salida.

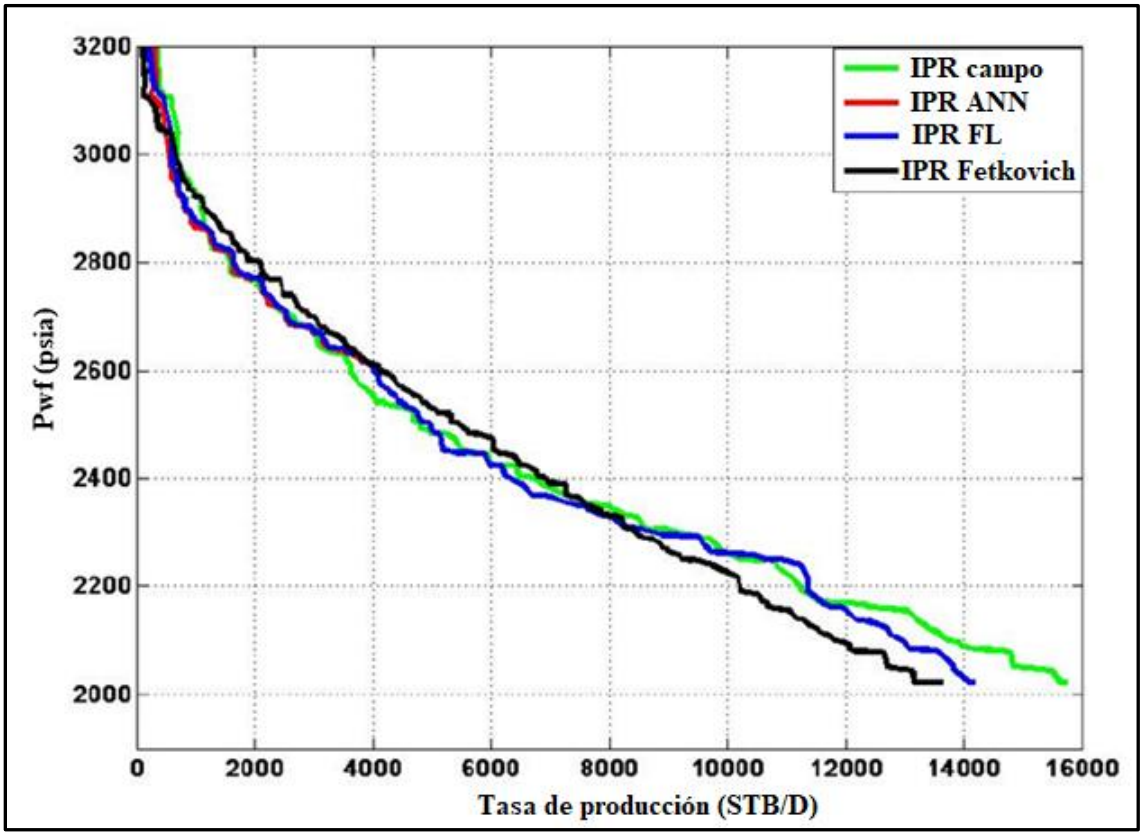
En este trabajo se utilizaron un total de 207 puntos coordenados para un yacimiento que tiene empuje por gas en solución y el pozo es vertical, con propiedades PVT variables y permeabilidad relativa definida.

Para la preparación y procesamiento de datos, se toma la tasa real de los pozos (que están produciendo por debajo de la presión de burbuja). Los parámetros a tener en cuenta son: presión del yacimiento, API del crudo, presión del fondo del pozo y los caudales de petróleo, gas y agua. Para ambas técnicas se agruparon los datos en 3 conjuntos: 70% entrenamiento, 15% validación y 15% para prueba. Con el objetivo de definir cuál es el mejor modelo predictivo, se llevó a cabo una comparación entre las 2 técnicas propuestas, los datos históricos de campo y un método convencional para obtener el IPR (Fetkovich).

Los resultados de la implementación de la red neuronal artificial y el modelo de lógica difusa son comparados con el IPR calculado mediante la correlación de Fetkovich (que fue la más precisa para los datos analizados) y con el IPR real del campo. Lo anterior se puede evidenciar en la Figura 49:

Figura 49

Comparación del IPR (real, Fetkovich, RNA y lógica difusa)



Nota. Adaptado de (Basfar et al., 2018).

El error porcentual promedio asociado a cada modelo de IPR fue de 13,5% para Fetkovich, 12,6% para la red neuronal y 1,22% para la lógica difusa. Se concluye que la técnica de lógica difusa es la mejor elección para resolver este problema, ya que presenta la menor desviación respecto a los datos reales y el menor error porcentual promedio.

5.9 Estimación de la porosidad de un yacimiento utilizando redes neuronales artificiales

La porosidad es la relación entre el volumen poroso (espacio de la roca que puede ser ocupado por fluidos) y el volumen total de la roca. Esta propiedad es fundamental para la caracterización de yacimientos y puede ser estimada por varios métodos: registros de pozo, corazonamiento, simulación, entre otros; sin embargo, pese a que estos suelen ser efectivos, requieren tiempo y costos adicionales. Una propuesta es utilizar como punto de partida los parámetros de perforación para determinar las propiedades de la formación, aunque esto representa un nuevo reto debido a la dificultad para generar o encontrar una conexión o dependencia entre los datos de perforación y las características del yacimiento.

A fin de enfrentar el desafío de procesar y tratar los datos para encontrar relación entre los parámetros de perforación y las propiedades de la formación, (Al-Abduijabbar et al., 2020) deciden recurrir a la inteligencia artificial y, en particular, a las redes neuronales artificiales (RNA), ya que estas pueden superar las limitaciones impuestas por la información disponible al imitar el sistema nervioso biológico para resolver problemas complejos de ingeniería y establecer las interacciones entre las variables dependientes e independientes dentro de un procedimiento.

La técnica de RNA aprovechó los datos suministrados de pozos horizontales, llamados A y B, separados 5 kilómetros uno de otro y ubicados en la misma sección de carbonato. El proceso que se lleva a cabo se describe de la siguiente manera:

1. Preparación de los datos: se calculó la porosidad a partir de la interpretación de registros obtenidos a través de sensores de fondo de pozo (LWD). Sólo se tienen en cuenta los datos de perforación medibles en superficie, es decir, que puedan ser tomados en tiempo real como la tasa de penetración (ROP), peso de la broca (WOB), tasa de bombeo (GPM), velocidad de rotación (RPM) y presión del tubo vertical (SPP).

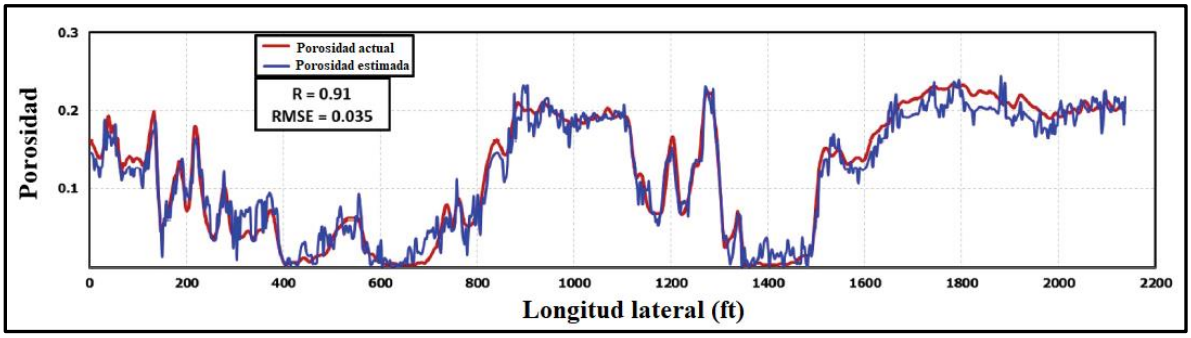
Posteriormente se hace un filtrado teniendo en cuenta sólo el tramo perforado de interés y, finalmente, se normalizan los datos, dando como resultado 2000 puntos disponibles del pozo A y 800 puntos disponibles del pozo B para un total de 2800.

2. Retroalimentación: los datos se distribuyeron de tal manera que 70% de ellos se destinarán para entrenamiento y 30% para prueba (sólo se usaron los del pozo A). El modelo ANN consta de una capa de entrada, dos ocultas y una de salida (cada nivel oculto cuenta con 30 neuronas).
3. Validación: para este caso se utilizó únicamente la información del pozo B donde se emplearon los parámetros de perforación medidos en superficie para estimar la porosidad y después se compararon los registros de la herramienta de fondo (LWD).

En la Figura 50 se presenta la comprobación del modelo en el tramo especificado:

Figura 50

Porosidad estimada y real a lo largo de la longitud lateral del pozo B (validación del modelo)



Nota. La gráfica ilustra la comparación entre la porosidad estimada por el modelo de RNA (línea azul) y la porosidad real (línea roja), presentando un coeficiente de correlación de 0,91 y un error cuadrático medio de 0,035. Adaptado de (Al-Abduijabbar et al., 2020).

Los autores concluyeron que el uso de esta técnica es muy prometedor debido a sus resultados. Además, según ellos, es posible mejorarla con el entrenamiento de pozos adicionales y la integración de un modelo desarrollado para obtener la estimación en tiempo real de la

porosidad mientras se perfora. El modelo actualmente sólo puede ser usado para predecir porosidad y para nuevos conjuntos de datos que compartan el mismo rango.

5.10 Predicción y prescripción de la operación en la unidad de endulzamiento de gas H_2S mediante Big Data Analytics

En un estudio realizado por (Cadei et al., 2019), los autores abordan el problema del sulfuro de hidrógeno (H_2S) que se presenta en el gas producido. Este gas es asociado con dificultades como generar impacto ambiental, poseer capacidad corrosiva que daña los equipos y ocasionar inconvenientes en el transporte, ya que el gas de venta no debe exceder cierto límite para el contenido de H_2S, según la norma vigente para transporte por gasoductos. El proceso de endulzamiento consiste en separar este contaminante tóxico del flujo de hidrocarburos a través de varios mecanismos como son: aminas, solventes, tamices moleculares, membranas, etc.

Los autores proponen una técnica que se basa en una poderosa infraestructura de Big Data donde se aplicará el Machine Learning (Deep Learning) y minería de datos, que a su vez se alimenta de datos suministrados en tiempo real, con el fin de predecir eventos casi al instante dentro del sistema de endulzamiento.

Para este caso de estudio, se toman los datos de un campo de petróleo y gas en el sur de Europa que actualmente cuenta con 5 líneas de producción que tratan el flujo multifásico (petróleo, gas y agua) proveniente de 27 pozos productores con 0,5 - 1,5% de contenido molar de H_2S y 5 - 30% de CO_2. Para llevar a cabo el proceso, se contará con sensores para un monitoreo antes y después del procesamiento del flujo.

A continuación, se presenta la secuencia que sigue la técnica:

1. Recopilación y agregación de datos: dentro de la planta de tratamiento se hizo un enfoque al procedimiento de endulzamiento del gas para reducir el conjunto de datos.

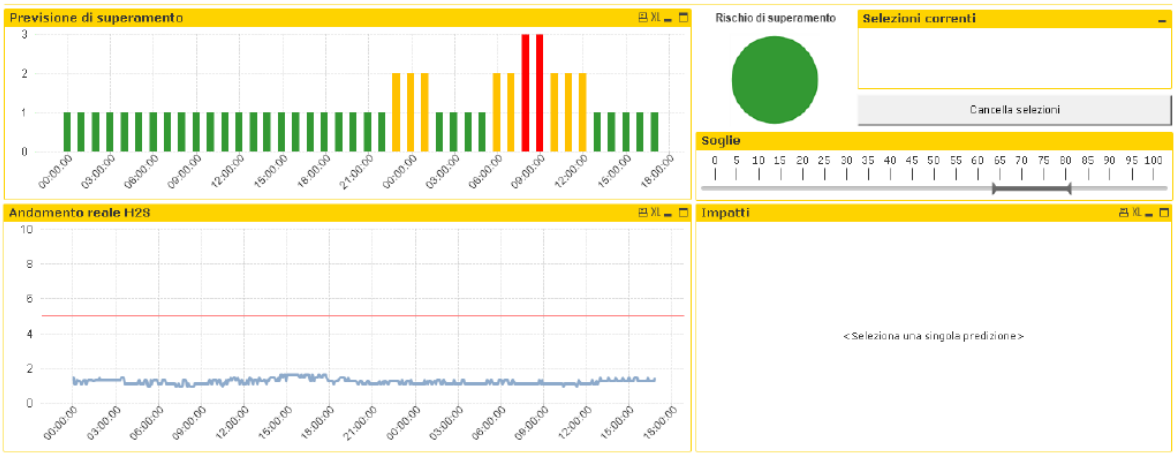
Se aplica interpolación inteligente y una correcta limpieza para garantizar un flujo limpio y constante de información a la infraestructura de Big Data. También, es posible enriquecer los datos con fuentes externas como análisis químicos e informes de mantenimiento.

2. Ingeniería y selección de características: el objetivo de esta fase es extraer información adicional mediante la transformación de datos. A su vez, la entrada pasa por diferentes fases que son: generación de características autorregresivas, generación de características sintéticas y selección de características.
3. Modelo de aprendizaje automático: para esta fase, las características seleccionadas en la anterior fase (tanto las sintéticas como las originales) sirven como entrenamiento del modelo. Posteriormente, se efectúa una validación cruzada teniendo en cuenta los requisitos impuestos por administradores y operadores de la planta.
4. Visualización: finalmente, se programa el modelo para que se ejecute con frecuencia horaria determinada. Los resultados son presentados en tablas, pero, además, cada punto se clasifica en tres clases respecto al umbral de concentración de ácido sulfúrico (bajo, medio y alto), cada una representada con un color (verde, amarillo y rojo respectivamente).

Los resultados del sistema implementado se pueden apreciar en el panel de control del programa que se presenta en la Figura 51:

Figura 51

Panel de control para los operadores del sitio que monitorean la predicción de H_2S



Nota. La ilustración corresponde a una muestra de la interfaz del sistema desarrollado. En la esquina superior derecha se muestra, en tiempo real, el tipo de riesgo cuando la concentración de H_2S traspasa el umbral definido; en la esquina superior izquierda, en el gráfico de barras, se muestra la tendencia durante el último día; en la esquina inferior izquierda, se muestran los datos del sensor H_2S para evaluar rápidamente el comportamiento de la unidad de endulzamiento de gas y validar la predicción del modelo; finalmente, en la esquina inferior derecha, en caso de una predicción positiva, se muestran los impactos de las características. Tomado de (Cadei et al., 2019).

Finalmente, los autores determinaron que el algoritmo implementado en el sistema desarrollado presenta una precisión del 78% en la predicción del traspaso del H_2S del límite normativo establecido, con entre 2 y 10 horas de antelación. Pese a que su precisión no es completamente de fiar, los autores consideran que el sistema muestra un alto impacto potencial en la gestión de activos, siendo especialmente útil en el mejoramiento del control de equipos, la optimización de la producción garantizando un alto nivel de calidad en el gas de venta y en la reducción de los riesgos potenciales asociados a la presencia del H_2S en el gas de venta.

6. Conclusiones

La conclusión más importante a la cual se llegó tras haber realizado la investigación es que la utilización de la ciencia de datos puede generar resultados con gran precisión, a menudo mayor que aquella que poseen los resultados obtenidos mediante la metodología convencional, además, los produce en un tiempo significativamente menor, por lo que debería considerarse la posibilidad de incluir la formación en la ciencia de datos dentro del pénsum de la carrera con lo que los futuros profesionales, al entrar al ámbito laboral, contarían con herramientas y habilidades que les permitirían proponer soluciones innovadoras y eficientes a los problemas que, en el ejercicio de su labor, puedan encontrar.

Existe un muy amplio y variado conjunto de técnicas de la ciencia de datos y, con frecuencia, es posible utilizar diferentes técnicas con el propósito de llevar a cabo un mismo procedimiento. Sin embargo, el desempeño de las técnicas utilizadas para resolver un procedimiento en particular puede variar significativamente al comparar unas con otras, por lo que es tarea del investigador probar diferentes alternativas y elegir la que mejor se adapte a sus necesidades.

Las redes neuronales artificiales son la técnica de ciencias de datos más utilizada, puesto que es la más versátil y la que suele ofrecer los mejores resultados en los procedimientos donde se aplica. No obstante, puede presentar una alta dificultad para ser implementada en función de la complejidad del problema que pretende resolver.

El área más beneficiada actualmente por la ciencia de datos, es decir, aquella donde más estudios se han realizado, es la de yacimientos. Ello se debe a que las principales aplicaciones corresponden a la caracterización de yacimientos, especialmente, la estimación de las propiedades de la roca y del fluido, y también la interpretación de registros de pozo.

La industria petrolera tiene un enorme potencial para la aplicación de la ciencia de datos considerando que es de las pocas industrias que generan enormes cantidades de datos todo el tiempo y en todas las etapas de la vida útil de sus proyectos. Todos los datos que se han ido almacenando a lo largo del tiempo podrían ser procesados con el fin de extraer información útil y potencialmente utilizable como soporte en la toma de decisiones o en la planeación y diseño de tareas y procedimientos.

7. Recomendaciones

Considerando que la presente investigación corresponde a un estudio exploratorio y, por lo tanto, general, se recomienda profundizar en el tema desarrollando un estudio detallado sobre la aplicación de técnicas de la ciencia de datos sobre algún procedimiento en particular e, idealmente, utilizando información de campo, para así poder medir el desempeño de las técnicas con información de primera mano.

Como alternativa a la recomendación anterior, podría realizarse un estudio sobre la aplicación de una técnica de la ciencia de datos en particular (por ejemplo, las redes neuronales artificiales) sobre diferentes procedimientos de la ingeniería de petróleos. Ello con el fin de determinar el potencial de uso de dicha técnica dentro de la industria petrolera.

Se recomienda analizar el aporte que la ciencia de datos ha hecho a la industria petrolera durante la situación de pandemia y los beneficios que ha tenido su aplicación.

Se sugiere realizar un estudio profundo y minucioso sobre las implicaciones que podría tener la formación en ciencia de datos dentro de la carrera de ingeniería de petróleos en la UIS.

Ello implica un análisis de los requisitos de hardware, software, capacitación docente, reestructuración del plan educativo de la carrera, etc.

El presente proyecto de investigación fue realizado considerando exclusivamente el sector upstream (exploración y producción) de la industria petrolera. En consecuencia, se recomienda investigar sobre el potencial de la aplicación de la ciencia de datos en los otros sectores de la industria.

Referencias Bibliográficas

Abbas, A. K., Assi, A. H., Abbas, H., Almubarak, H., & Saba, M. Al. (2019). Drill bit selection optimization based on rate of penetration: Application of artificial neural networks and genetic algorithms. *Society of Petroleum Engineers - Abu Dhabi International Petroleum Exhibition and Conference 2019, ADIP 2019*. https://doi.org/10.2118/197241-ms

ActiveWizards. (2019, Agosto). *Artificial Intelligence vs. Machine Learning vs. Deep Learning: What is the Difference?* KDnuggets. https://www.kdnuggets.com/2019/08/artificial-intelligence-vs-machine-learning-vs-deep-learning-difference.html

Aggarwal, C. C. (2018). *Neural Networks and Deep Learning: A Textbook*. Springer. doi:10.1007/978-3-319-94463-0

Ahmed, A. A., Elkatatny, S., Abdulraheem, A., & Mahmoud, M. (2017). Application of artificial intelligence techniques in estimating oil recovery factor for water derive sandy reservoirs. *Society of Petroleum Engineers - SPE Kuwait Oil and Gas Show and Conference 2017*. https://doi.org/10.2118/187621-ms

Al-Abduijabbar, A., Al-Azani, K., & Elkatatny, S. (2020). Estimation of reservoir porosity from drilling parameters using artificial neural networks. *Petrophysics*, 61(3), 318–329. https://doi.org/10.30632/PJV61N3-2020a5

Al-AbdulJabbar, A., Elkatatny, S., Mahmoud, M., & Abdulraheem, A. (2018). Predicting Rate of Penetration Using Artificial Intelligence Techniques. *SPE*. https://doi.org/10.2118/192343-MS

Al-Amoudi, L. A., Geri, B. S. B., Patil, S., & Baarimah, S. O. (2019). Development of artificial intelligence models for prediction of crude oil viscosity. *SPE Middle East Oil and Gas Show and Conference, MEOS, Proceedings*. https://doi.org/10.2118/194741-ms

Alarifi, S., Al Nuaim, S., & Abdulraheem, A. (2015). Productivity Index prediction for oil horizontal wells using different artificial intelligence techniques. *SPE Middle East Oil and Gas Show and Conference*. https://doi.org/10.2118/172729-ms

Alavala, C. (2007). *Fuzzy Logic and Neural Networks: Basic concepts and applications*. New Age International.

Al-Hameedi, A. T. T., Alkinani, H. H., Dunn-Norman, S., Flori, R. E., Alsaba, M. T., Amer, A. S., & Al-Bazzaz, W. H. (2019). An assessment of the impact of rheological properties on rate of penetration using data mining techniques. *International Petroleum Technology Conference 2019, IPTC 2019*. https://doi.org/10.2523/19446-ms

AlJuboori, M., Hossain, M., Al-Fatlawi, O., Kabir, A., & Radhi, A. (2020). Numerical simulation of gas lift optimization using genetic algorithm for a middle east oil field: Feasibility study. *International Petroleum Technology Conference 2020, IPTC 2020*. https://doi.org/10.2523/iptc-20254-ms

Alloush, R. M., Elkatatny, S. M., Mahmoud, M. A., Moussa, T. M., Ali, A. Z., & Abdulraheem, A. (2017). Estimation of geomechanical failure parameters from well logs using artificial intelligence techniques. *Society of Petroleum Engineers - SPE Kuwait Oil and Gas Show and Conference 2017*. https://doi.org/10.2118/187625-ms

Al-Mudhafar, W. J. (2017). Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms. *Journal of Petroleum Exploration and Production Technology*. https://doi.org/10.1007/s13202-017-0360-0

Alouhali, R., Aljubran, M., Gharbi, S., & Al-Yami, A. (2018). Drilling through data: Automated kick detection using data mining. *Society of Petroleum Engineers - SPE International Heavy Oil Conference and Exhibition 2018, HOCE 2018*. https://doi.org/10.2118/193687-MS

Alshaikh, A., Magana-Mora, A., Gharbi, S. Al, & Al-Yami, A. (2019). Machine Learning for Detecting Stuck Pipe Incidents: Data Analytics and Models Evaluation. *International Petroleum Technology Conference*. https://doi.org/10.2523/IPTC-19394-MS

Aming, A. (2021). Artificial Intelligence AI / Machine Learning ML Drives Increased Capital Efficiency and Minimizes Geological Risk in E&P Operations. *SPE Trinidad and Tobago Section Energy Resources Conference*. https://doi.org/10.2118/200978-MS

Anifowose, F., Ayadiuno, C., & Reshedan, F. (2019). Feature Selection Based Hybrid Machine Learning Approach to Formation Cementation Factor Prediction. *SPE Kuwait Oil & Gas Show and Conference*. https://doi.org/10.2118/198074-MS

Bahga, A., & Madisetti, V. (2019). *Big Data Analytics: A Hands-On Approach*. Arshdeep Bahga & Vijay Madisetti.

Baldini, D., Piazza, L., & Barbanotti, L. (2020). Artificial Intelligence and Machine Learning Techniques Provide Operations Geologists with an Automated and Reliable Lithology-

Fluid Pattern Recognition Assistant: A Case History in a Clastic Reservoir in West Africa. *International Petroleum Technology Conference*. https://doi.org/10.2523/IPTC-19701-MS

Bangert, P. (2019). Diagnosing and Predicting Problems with Rod Pumps using Machine Learning. *SPE Middle East Oil and Gas Show and Conference*. 1–13. https://doi.org/10.2118/194993-ms

Barrio, M. (2018). *INTERNET DE LAS COSAS*. Reus.

Basfar, S., Baarimah, S. O., Elkatany, S., AL-Ameri, W., Zidan, K., & ALdogail, A. (2018). Using artificial intelligence to predict IPR for vertical oil well in solution gas derive reservoirs: A new approach. *Society of Petroleum Engineers - SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition 2018, SATS 2018*, 1–12. https://doi.org/10.2118/192203-ms

Belozerov, B., Bukhanov, N., Egorov, D., Zakirov, A., Osmonalieva, O., Golitsyna, M., Reshytko, A., Semenikhin, A., Shindin, E., & Lipets, V. (2018). Automatic Well Log Analysis Across Priobskoe Field Using Machine Learning Methods. *SPE Russian Petroleum Technology Conference*. https://doi.org/10.2118/191604-18RPTC-MS

Bhavsar, P., Safro, I., Bouaynaya, N., Polikar, R., & Dera, D. (2017). Machine Learning in Transportation Data Analytics. En M. Chowdhury, M. Apon, & K. Dey (Edits.), *Data Analytics for Intelligent Transportation Systems* (págs. 283-307). Elsevier Inc. https://doi.org/10.1016/B978-0-12-809715-1.00012-2

BP. (2021). *Statistical Review of World Energy 2021*. bp website. https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html

Bravo, R. J. C., Nieves, E. J., & Arhuata, L. A. (2020). Predictive data mining techniques for economic evaluation of unconventional resources: The tight gas of Argentina. *Offshore Technology Conference Brasil 2019, OTCB 2019*. https://doi.org/10.4043/29954-ms

Brownlee, J. (2017*). Master Machine Learning Algorithms: Discover how they work and implement them from Scratch*. Publicación independiente.

Cadei, L., Camarda, G., Montini, M., Rossi, G., Fier, P., Bianco, A., Lancia, L., Loffreno, D., Corneo, A., Milana, D., Carrettoni, M., & Silvestri, G. (2019). Prediction and Prescription of Operation Upset in H2S Gas Sweetening Unit: Implementation of an Innovative Big Data Analytics Procedure. *Offshore Mediterranean Conference and Exhibition.*

Cady, F. (2017). *The Data Science Handbook*. Wiley. https://doi.org/10.1002/9781119092919

Carvajal, G., Maucec, M., & Cullick, S. (2018). *INTELLIGENT DIGITAL OIL AND GAS FIELDS.* Gulf Professional Publishing. https://doi.org/10.1016/C2015-0-02216-X

Chandel, A., & Sood, M. (2014). Searching and Optimization Techniques in Artificial Intelligence: A Comparative Study and Complexity Analysis. *IJARCET*, 866-871.

Chen, C., Han, X., Zhang, W., Zhang, Y., & Zhou, F. (2021). A New Artificial Intelligence Method to Predict Water Flooding Performance in Layered Reservoir. *International Petroleum Technology Conference*. https://doi.org/10.2523/IPTC-21317-MS

Cirani, S., Ferrari, G., Picone, M., & Veltri, L. (2018). *INTERNET OF THINGS Architectures, Protocols and Standards*. John Wiley & Sons, Inc. doi:10.1002/9781119359715

Conway, D. (2010, 30 de septiembre). *The Data Science Venn Diagram.* Dataists. http://www.dataists.com/2010/09/the-data-science-venn-diagram/

Cooper, S. (2018). *Data Science from Scratch - The #1 data science guide for everything a data scientist needs to know: Python, linear algebra, statistics, coding, applications, neural networks, and decision trees.* (Publicación independiente).

Cornel, S., & Vazquez, G. (2020). Use of Big Data and Machine Learning to Optimise Operational Performance and Drill Bit Design. *SPE Asia Pacific Oil & Gas Conference and Exhibition.* https://doi.org/10.2118/202243-MS

Cumming, J., Riggins, V., Hodson, P., & Walker, B. (2020). Advanced Well Planning Using Natural Language Processing NLP and Data Science Models: Maximizing the Value of Data to Mitigate Costs and Risks in New Wells. *Abu Dhabi International Petroleum Exhibition & Conference*. https://doi.org/10.2118/203280-MS

Dang, C., Nghiem, L., Fedutenko, E., Gorucu, E., Yang, C., & Mirzabozorg, A. (2018). Application of artificial intelligence for mechanistic modeling and probabilistic forecasting of hybrid low salinity chemical flooding. *Proceedings - SPE Annual Technical Conference and Exhibition.* https://doi.org/10.2118/191474-ms

Duarte, S. B., De Jesus, C. M., Da Silva, V. F., Arouca Sobreira, M. C., Cristofaro, R. A. L., de Lima, L., De Mello E Silva, F. G., Marques De Sa, C. H., De Oliveira Berto, F. M., Backheuser, Y., Loureiro, S. de A., De Almeida Waldmann, A. T., & Fioriti, L. D. S.

(2018). Artificial Intelligence Use to Predict Severe Fluid Losses in Pre-Salt Carbonates. *SPWLA 59th Annual Logging Symposium.*

Elgendy, M. (2020). *Deep Learning for Vision Systems*. Manning Publications.

Elzenary, M., Elkatatny, S., Abdelgawad, K. Z., Abdulraheem, A., Mahmoud, M., & Al-Shehri, D. (2018). New technology to evaluate equivalent circulating density while drilling using artificial intelligence. *Society of Petroleum Engineers - SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition 2018*. https://doi.org/10.2118/192282-ms

Escuela de Ingeniería de Petróleos - UIS. (2015). *Proyecto Educativo de Reforma Académica del Programa Ingeniería de Petróleos*.

Franceschetti, D. R. (Ed.). (2018). *Principles of Robotics & Artificial Intelligence*. Grey House Publishing.

Frankish, K., & Ramsey, W. M. (2014*). The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press. https://doi.org/10.1017/CBO9781139046855

Géron, D. (2019). *Machine Learning for Beginners: Step-by-Step Guide to Learning and Mastering Machine Learning for Absolute Beginners with Real Examples.* (Publicación independiente).

Gholami, R., Moradzadeh, A., Maleki, S., Amiri, S., & Hanachi, J. (2014). Applications of artificial intelligence methods in prediction of permeability in hydrocarbon reservoirs. *Journal of Petroleum Science and Engineering*, 122, 643–656. https://doi.org/10.1016/j.petrol.2014.09.007

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

Gupta, A., & Soumya, U. (2020). Well log interpretation using deep learning neural networks. *International Petroleum Technology Conference 2020, IPTC 2020*. https://doi.org/10.2523/iptc-19678-abstract

Gupta, I., Samandarli, O., Burks, A., Jayaram, V., McMaster, D., Niederhut, D., & Cross, T. (2021). Autoregressive and Machine Learning Driven Production Forecasting - Midland Basin Case Study. *SPE/AAPG/SEG Unconventional Resources Technology Conference.* https://doi.org/10.15530/urtec-2021-5184

Hackeling, G. (2014). *Mastering Machine Learning with scikit-learn.* Packt Publishing.

Han, J., Kamber, M., & Pei, J. (2012*). Data Mining: Concepts and Techniques* (Tercera ed.). Elsevier.

Han, J., Sun, Y., & Zhang, S. (2019). A Data Driven Approach of ROP Prediction and Drilling Performance Estimation. *International Petroleum Technology Conference*. https://doi.org/10.2523/IPTC-19430-MS

He, Q., Zhong, Z., Alabboodi, M., & Wang, G. (2019). Artificial intelligence assisted hydraulic fracturing design in shale gas reservoir. *SPE Eastern Regional Meeting*. https://doi.org/10.2118/196608-ms

Heaton, J. (2015). *Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks*. Heaton Research, Inc.

Hou, X., Yang, J., Yin, Q., Chen, L., Cao, B., Xu, J., Meng, L., Zhang, Y., Xu, D., & Zhao, X. (2019). Automatic Gas Influxes Detection in Offshore Drilling Based on Machine Learning

Technology. *SPE Gas & Oil Technology Showcase and Conference.* https://doi.org/10.2118/198534-MS

Hou, X., Yang, J., Yin, Q., Liu, H., Chen, H., Zheng, J., Wang, J., Cao, B., Zhao, X., Hao, M., & Liu, X. (2020). Lost Circulation Prediction in South China Sea using Machine Learning and Big Data Technology. *Offshore Technology Conference*. https://doi.org/10.4043/30653-MS

Kantardzic, M. (2020). *DATA MINING: Concepts, Models,Methods, and Algorithms* (Tercera ed.). John Wiley & Sons, Inc.

Karaboga, D., & Kaya, E. (2018). *Adaptive network based fuzzy inference system (ANFIS) training approaches: a comprehensive survey*. Springer.

Kassambara, A. (2017*). Machine learning essentials*. STHDA.

Kelleher, J., & Tierney, B. (2018). *Data Science*. The MIT Press.

Khan, H., & Louis, C. (2021). An Artificial Intelligence Neural Networks Driven Approach to Forecast Production in Unconventional Reservoirs – Comparative Analysis with Decline Curve. *International Petroleum Technology Conference*. https://doi.org/10.2523/IPTC-21350-MS

Khan, M. R., Tariq, Z., & Abdulraheem, A. (2018). Machine Learning Derived Correlation to Determine Water Saturation in Complex Lithologies. *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition*. https://doi.org/10.2118/192307-MS

Klie, A., Klie, H., Vuong, D., Chaban, F., & Chaban, N. (2020). Automated Lease Operating Statements for Cost Optimization and Reserve Evaluation Using Artificial Intelligence. *SPE*. https://doi.org/10.2118/201710-MS

Konoshonkin, D., Shishaev, G., Matveev, I., Volkova, A., Rukavishnikov, V., Demyanov, V., & Belozerov, B. (2020). Machine Learning Clustering of Reservoir Heterogeneity with Petrophysicaland Production Data. *Society of Petroleum Engineers - SPE*. https://doi.org/10.2118/200614-ms

Kranz, M. (2017). *Internet of Things*. LID.

Lake, L. W. (Ed.). (2007). *Petroleum Engineering Handbook* (Vols. I - VII). Society of Petroleum Engineers.

Li, Y., & Han, Y. (2017). Decline curve analysis for production forecasting based on machine learning. *Society of Petroleum Engineers - SPE Symposium: Production Enhancement and Cost Optimization 2017*. https://doi.org/10.2118/189205-ms

Liang, Y., & Zhao, P. (2019). A Machine Learning Analysis Based on Big Data for Eagle Ford Shale Formation. *SPE Annual Technical Conference and Exhibition*. https://doi.org/10.2118/196158-MS

Lin, A., Alali, M., Almasmoom, S., & Samuel, R. (2018). Wellbore instability prediction using adaptive analytics and empirical mode decomposition. *Society of Petroleum Engineers - IADC/SPE Drilling Conference and Exhibition*, 1–10. https://doi.org/10.2118/189598-ms

Liu, Y., Chen, C., Zhao, H., Wang, Y., & Han, X. (2021). A Robust Method to Predict Fluid Properties Based on Big Data and Machine Learning Algorithms. *International Petroleum Technology Conference.* https://doi.org/10.2523/IPTC-21356-MS

Loretti, R. A., Da Costa, V. F. P., Memoria, D. G. D. O., Barbosa, A. N., Oliveira, H. L. S., Wegner, I. R., & Zank, C. A. C. (2019). Data Science and Business Intelligence Techniques for Learning from Environmental Accident Analysis for Offshore Oil Fields. *Offshore Technology Conference Brasil.* https://doi.org/10.4043/29725-MS

Lucas, P. J., & van der Gaag, L. (1991). *Principles of Expert Systems*. Addison-Wesley.

Lyons, W., Plisga, G., & Lorenz, M. (Edits.). (2015). *Standard Handbook of Petroleum and Natural Gas Engineering* (Tercera ed.). Gulf Professional Publishing.

Masini, S. R., Goswami, S., Kumar, A., & Chennakrishnan, B. (2019). Decline Curve Analysis Using Artificial Intelligence. *Abu Dhabi International Petroleum Exhibition & Conference*. https://doi.org/10.2118/197932-MS

Masoudi, R., Mohaghegh, S. D., Yingling, D., Ansari, A., Amat, H., Mohamad, N., Sabzabadi, A., & Mandel, D. (2020). Subsurface Analytics Case Study; Reservoir Simulation and Modeling of Highly Complex Offshore Field in Malaysia, Using Artificial Intelligent and Machine Learning. *SPE Annual Technical Conference and Exhibition*. https://doi.org/10.2118/201693-MS

Merayo, D., Rodríguez-Prieto, A., & Camacho, A. M. (2019). Comparative analysis of artificial intelligence techniques for material selection applied to manufacturing in Industry 4.0. *Procedia Manufacturing*, 41, 42–49. https://doi.org/10.1016/j.promfg.2019.07.027

Miller, R. S., Rhodes, S., Khosla, D., & Nino, F. (2019). Application of Artificial Intelligence for Depositional Facies Recognition - Permian Basin. *SPE/AAPG/SEG Unconventional Resources Technology Conference.* https://doi.org/10.15530/urtec-2019-193

Mohamed, I. M., Mohamed, S., Mazher, I., & Chester, P. (2019). Formation Lithology Classification: Insights into Machine Learning Methods. *SPE Annual Technical Conference and Exhibition.* https://doi.org/10.2118/196096-MS

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning.* The MIT Press.

Mokhatab, S., Valappil, J. V., Mak, J. Y., & Wood, D. A. (2014). Chapter 5: Natural Gas Liquefaction Cycle Enhancements and Optimization. En S. Mokhatab, J. V. Valappil, J. Y. Mak, & D. A. Wood. *Handbook of Liquefied Natural Gas* (págs. 229-257). Elsevier Inc.

Mueller, J. P., & Massaron, L. (2018). *Artificial Intelligence for dummies*. Wiley.

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists.* O'Reilly Media, Inc.

Negara, A., Ali, S., AlDhamen, A., Kesserwan, H., Nair, A., & Aleid, Z. (2018). Utilizing Artificial Intelligence Techniques for Predicting Rock Failure Parameters. *SPE*. https://doi.org/10.2118/192340-MS

Noshi, C. I. (2019). Application of Data Science and Machine Learning Algorithms for ROP Optimization in West Texas: Turning Data into Knowledge. *Offshore Technology Conference*. https://doi.org/10.4043/29288-MS

Noureldien, D. M., & El-Banbi, A. H. (2015). Using artificial intelligence in estimating oil recovery factor. *Society of Petroleum Engineers - SPE North Africa Technical Conference and Exhibition 2015*. https://doi.org/10.2118/175867-ms

Ojukwu, C., Smith, K., Kadkhodayan, N., Leung, M., & Baldwin, K. (2020). Reservoir Characterization, Machine Learning and Big Data – An Offshore California Case Study. *SPE Nigeria Annual International Conference and Exhibition.* https://doi.org/10.2118/203642-MS

Okpo, E. E., Dosunmu, A., & Odagme, B. S. (2016). Artificial neural network model for predicting wellbore instability. *Society of Petroleum Engineers - SPE Nigeria Annual International Conference and Exhibition*. https://doi.org/10.2118/184371-ms

Olah, C. (2015, 27 de agosto). *Colah's Blog*. Retrieved from Understanding LSTM Networks: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Ounsakul, T., Sirirattanachatchawan, T., Pattarachupong, W., Yokrat, Y., & Ekkawong, P. (2019). Artificial lift selection using machine learning. *International Petroleum Technology Conference 2019, IPTC 2019*. https://doi.org/10.2523/19423-ms

Ozdemir, S. (2016). *Principles of Data Science*. Packt Publishing.

Pankaj, P., Geetan, S., MacDonald, R., Shukla, P., Sharma, A., Menasria, S., Xue, H., & Judd, T. (2018). Application of Data Science and Machine Learning for Well Completion Optimization. *Offshore Technology Conference*. https://doi.org/10.4043/28632-MS

Parapuram, G. K., Mokhtari, M., & Hmida, J. Ben. (2017). Prediction and Analysis of Geomechanical Properties of the Upper Bakken Shale Utilizing Artificial Intelligence and

Data Mining. *SPE/AAPG/SEG Unconventional Resources Technology Conference.* https://doi.org/10.15530/URTEC-2017-2692746

Pascual, D., Pla, F., & Sánchez, S. (2007). Algoritmos de agrupamiento. En R. Quirós, F. Pla, J. M. Badía, & M. Chover (Edits.), *Métodos informáticos avanzados* (págs. 163-175). Publicacions de la Universitat Jaume I.

Pinzón, L. C., & Martínez, S. N. (2019). *Un modelo de medición de la calidad de los servicios que ofrece la biblioteca central de la UIS-Bucaramanga basado en lógica difusa.* (Tesis de pregrado). Universidad Industrial de Santander, Bucaramanga.

Pollo Cattaneo, M. F., Pytel, P., Vegega, C., Ramón, H. D., Deroche, A., Straccia, L., . . . Acosta, M. P. (2016). Implementación de sistemas inteligentes para la asistencia a alumnos y docentes de la carrera de Ingeniería en Sistemas de Información. *XVIII Workshop de Investigadores en Ciencias de la Computación*, (págs. 662-666). Obtenido de http://sedici.unlp.edu.ar/handle/10915/53034

Pollock, J., Stoecker-Sylvia, Z., Veedu, V., Panchal, N., & Elshahawi, H. (2018). Machine Learning for Improved Directional Drilling. *Offshore Technology Conference.* https://doi.org/10.4043/28633-MS

Porras, L., Hawkes, C., & Islam, A. (2020). Evaluation and Optimization of Completion Design using Machine Learning in an Unconventional Light Oil Play. *SPE/AAPG/SEG Unconventional Resources Technology Conference.* https://doi.org/10.15530/urtec-2020-2938

Rahmanifard, H., Alimohamadi, H., & Gates, I. (2020). Well Performance Prediction in Montney Formation Using Machine Learning Approaches. *SPE/AAPG/SEG Unconventional Resources Technology Conference*. https://doi.org/10.15530/urtec-2020-2465

Ray, S. (9 de Septiembre de 2017*). Commonly used Machine Learning Algorithms*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Rezaei, A., Siddiqui, F., Dindoruk, B., & Soliman, M. Y. (2020). Utilizing a Global Sensitivity Analysis and Data Science to Identify Dominant Parameters Affecting the Production of Wells and Development of a Reduced Order Model for the Eagle Ford Shale. *SPE/AAPG/SEG Unconventional Resources Technology Conference.* https://doi.org/10.15530/urtec-2020-2751

Sambo, C. H., Hermana, M., Babasari, A., Janjuhah, H. T., & Ghosh, D. P. (2018). Application of artificial intelligence methods for predicting water saturation from new seismic attributes. *Offshore Technology Conference Asia 2018, OTCA 2018*. https://doi.org/10.4043/28221-ms

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal,* 210-229.

Sarkar, D., Bali, R., & Sharma. (2018). *Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems.* Apress. https://doi.org/10.1007/978-1-4842-3207-1

Shadravan, A., Tarrahi, M., & Amani, M. (2015). Intelligent cement design: Utilizing machine learning algorithms to assure effective long-term well integrity. *Carbon Management Technology Conference 2015: Sustainable and Economical CCUS Options, CMTC 2015.* https://doi.org/10.7122/440236-ms

Shadravan, A., Tarrahi, M., & Amani, M. (2017). Intelligent Tool To Design Drilling, Spacer, Cement Slurry, and Fracturing Fluids by Use of Machine-Learning Algorithms. *SPE Drilling & Completion,* 32(02), 131–140. https://doi.org/10.2118/175238-PA

Shelley, R., Melcher, H., & Oduba, O. (2021). Machine Learning and Artificial Intelligence Provides Wolfcamp Completion Design Insight. *SPE/AAPG/SEG Unconventional Resources Technology Conference.* https://doi.org/10.15530/urtec-2021-5572

Sheppard, C. (2016). *Genetic Algorithms with Python*. Independiente.

Shi, X., Zhou, Y., Zhao, Q., Jiang, H., Zhao, L., Liu, Y., & Yang, G. (2019). A New Method to Detect Influx and Loss During Drilling Based on Machine Learning. *International Petroleum Technology Conference.* https://doi.org/10.2523/IPTC-19489-MS

Sierra, D. M., Rojas, A. A., & Araque, V. S. (2020). Low salinity water injection optimization in the Namorado Field using compositional simulation and artificial intelligence. *SPE Latin American and Caribbean Petroleum Engineering Conference Proceedings*. https://doi.org/10.2118/198995-ms

Sivanandam, S., & Deepa, S. (2008). *Introduction to Genetic Algorithms*. Springer. doi: 10.1007/978-3-540-73190-0

Solomatine, D. P., & Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of Hydroinformatics*, 3-22. https://doi.org/10.2166/hydro.2008.015

Solomatine, D., See, L. M., & Abrahart, R. J. (2008). Data-Driven Modelling: Concepts, Approaches and Experiences. En R. J. Abrahart, L. M. See, & D. P. Solomatine (Edits.), *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications* (págs. 17-30). Springer. https://doi.org/10.1007/978-3-540-79881-1

Spyropoulos, C. D. (2000). AI planning and scheduling in the medical hospital environment. *Artificial Intelligence in Medicine*, 101-111. https://doi.org/10.1016/S0933-3657(00)00059-2

Sun, H., & Belhaj, H. (2019). Incorporated Artificial Intelligence and Digital Imaging System for Unconventional Reservoirs Characterization. *SPE Annual Technical Conference and Exhibition.* https://doi.org/10.2118/195834-MS

Tan, L., & Wang, N. (2010). Future internet: The Internet of Things. *2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)*, (pp. 376-380). doi:10.1109/ICACTE.2010.5579543

Tarrahi, M., & Shadravan, A. (2016). Advanced Big Data Analytics Improves HSE Management. *SPE Bergen One Day Seminar*. https://doi.org/10.2118/180032-MS

Tavares, M., Filho, N. F., & Carrasquilla, A. (2019). Facies, data mining and artificial intelligence approaches in the characterization of a carbonate reservoir in Campos Basin — Southeastern Brazil. *SEG International Exposition and Annual Meeting.* https://doi.org/10.1190/segam2019-3199693.1

Tharwat, A. (2016). Principal component analysis – a tutorial. *International Journal of Applied Pattern Recognition (IJAPR)*, 3(3), 197-240. doi: 10.1504/IJAPR.2016.079733

Theobald, O. (2017). *Machine Learning for Absolute Beginners*. (Publicado independientemente).

Ticona, J. L. (2017). *Sistema inteligente para el diagnóstico y tratamiento de las cataratas*. (Tesis de pregrado). Universidad Mayor de San Andrés, La Paz, Bolivia.

Tierney, B. (13 de junio de 2012). *Data Science Is Multidisciplinary.* Oralytics. https://oralytics.com/2012/06/13/data-science-is-multidisciplinary/

Tortrakul, N., Pochan, C., Southland, S., Mala, P., Pichaichanlert, T., & Tangsawanich, Y. (2021). Drilling Performance Improvement Through use of Artificial Intelligence in Bit and Bottom Hole Assembly Selection in Gulf of Thailand. *IADC/SPE Asia Pacific Drilling Technology Conference.* https://doi.org/10.2118/201079-MS

Tulleken, H. (8 de octubre de 2009). *15 Steps to Implement a Neural Net.* code-spot. http://www.code-spot.co.za/2009/10/08/15-steps-to-implemented-a-neural-net/

University of Nevada, Reno. (s.f.). *What are intelligent systems?*. University of Nevada Website. https://www.unr.edu/cse/undergraduates/prospective-students/what-are-intelligent-systems

Zhan, Y., Lu, S., Xiang, T., & Wei, T. (2020). Application of convolutional neural network in random structural damage identification. *Structures,* 570-576. doi:10.1016/j.istruc.2020.11.056

Zhang, A. (2017). *Data Analytics: Practical Guide to Leveraging the Power of Algorithms, Data Science, Data Mining, Statistics, Big Data, and Predictive Analysis to Improve Business, Work, and Life.* (Publicación independiente).

Zhao, P., Dong, R., & Liang, Y. (2020). Regional to Local Machine-Learning Analysis for Unconventional FormationReserve Estimation: Eagle Ford Case Study. *SPE*. https://doi.org/10.2118/201351-MS

Apéndices

Apéndice A. Resumen de Inteligencia Artificial

Apéndice B. Resumen de Machine Learning

Apéndice C. Resumen de Redes Neuronales Artificiales

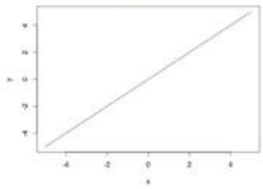
REDES NEURONALES ARTIFICIALES (RNA)

Una RNA es un modelo computacional y una arquitectura que simula las neuronas biológicas y la forma en que funcionan en nuestro cerebro. Se componen de una capa de entrada, una capa de resultados o de salida y una o varias capas ocultas.

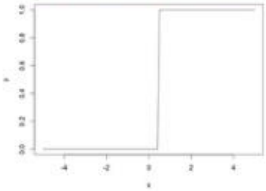
FUNCIÓN DE ACTIVACIÓN

Son uno de los componentes más importantes de una red neuronal artificial ya que son las encargadas de decidir si se activa o no una neurona. Las más comunes son:

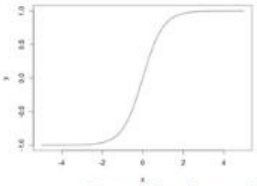
Función de activación lineal o identidad



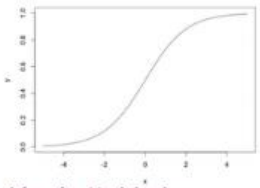
Función de activación de escalón o umbral



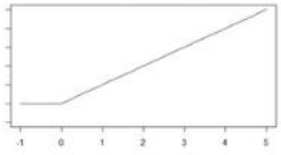
Función de activación tangente hiperbólica



Función de activación sigmoide o logística



Función de activación de Unidades Lineales Rectificada (ReLU)

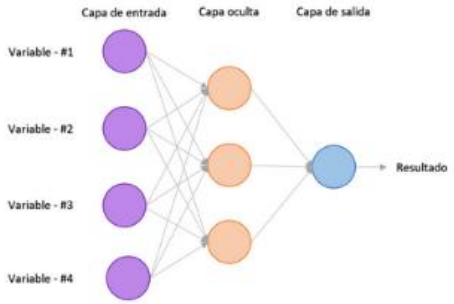


AREAS DE APLICACIÓN

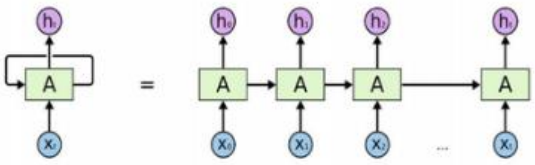
- Procesamiento de señales
- Reconocimiento de patrones
- Medicina (emisión de diagnósticos)
- Síntesis de habla (generación de voz)
- Reconocimiento de voz
- Negocios Procesamiento de imágenes
- Etc.

TIPOS DE RNA

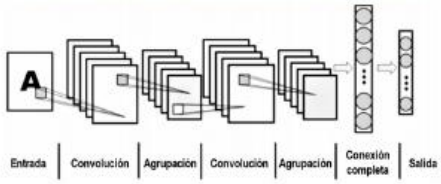
Red neuronal prealimentada (feed-forward)



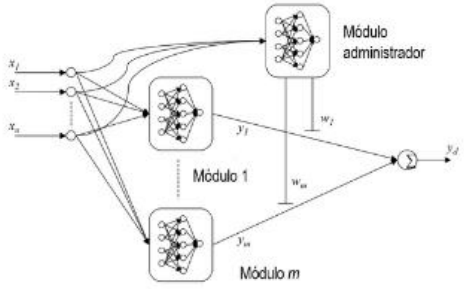
Redes Neuronales Recurrentes



Redes Neuronales Convolucionales



Redes Neuronales Modulares



Apéndice D. Resumen de Deep Learning

DEEP LEARNING

Es un subcampo de las redes neuronales artificiales que, a su vez, pertenecen al Machine Learning. Se fundamenta en el uso de niveles jerárquicos o capas de representaciones, las cuales facilitan las ejecución de las tareas propuestas. Básicamente, este enfoque resuelve problemas al introducir representaciones que se expresan en términos de representaciones más simples, por ejemplo, se puede representar el concepto de imagen de una persona combinando conceptos más simples, como esquinas y contornos, que a su vez se definen en términos de bordes. Las redes neuronales recurrentes y convolucionales pertenecen al Deep Learning.

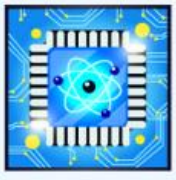
CARACTERÍSTICAS

- Sus algoritmos son de naturaleza iterativa y su complejidad aumenta dependiendo del número de capas y volumen de datos.
- Posee un enfoque dinámico, es decir, tiene una capacidad de mejorar y adaptarse continuamente a cambios en el patrón de información implícito.
- Representa una oportunidad de mejorar la precisión y el desempeño en aplicaciones.
- Posibilidad de ser más eficiente y simplificado en operaciones analíticas existentes.

APLICACIONES

- Análisis de imágenes médicas.
- Bioinformática.
- Salud pública.
- Localización de caras e identificación de emociones faciales.
- Detección, predicción y prevención de amenazas sofisticadas en tiempo real en el campo de la ciberseguridad.
- Identificación de clientes potenciales para empresas.

- Identificación de los niveles de confianza de los clientes para empresas.
- Procesamiento y reconocimiento de texto.
- Reconocimiento de voz.
- Clasificación de vídeos.
- Etc.



Apéndice E. Resumen de Data Mining

Apéndice F. Resumen de Algoritmos Genéticos

ALGORITMOS GENÉTICOS

Un algoritmo genético es un proceso iterativo que parte de un conjunto de soluciones obtenidos de manera aleatoria. A partir de ese conjunto cada solución se trata como un individuo y dicho individuo pasa por procesos de codificación, selección, combinación y mutación para generar nuevas soluciones y cada solución tendrá un valor de eficacia (a partir de una función de de aptitud también conocida como función "fitness") con respecto a un problema predeterminado.

Todo este proceso se asemeja o se inspira en la idea de la evolución en la rama de la genética, de ahí su nombre.

CARACTERÍSTICAS

- Método de búsqueda dirigida basada en probabilidad.
- Las soluciones están limitadas a un cierto rango porque su espacio de búsqueda es discreto. Se puede usar en espacios de búsqueda continuos cuando el rango de solución sea muy pequeño.
- La función de aptitud que marca el problema de optimización a resolver (o función adaptativa según su terminología), siempre es maximizada, y tiene que poder ser definida de forma que se nos indique si es buena o no cierta solución, premiando en el primer caso y penalizando en el segundo.
- Cada solución va a ser codificada, normalmente en forma binaria.

APLICACIONES

- Enrutamiento de tráfico y envío.
- Robótica.
- Optimización de procesos.
- Predicción precisa de fallas en redes eléctricas.
- Imágenes por resonancia magnética.
- Finanzas.
- Aeronáutica.
- Diseño de redes de agua potable.
- Plegamiento de proteínas.
- Optimización de portafolios.
- Optimización de algoritmos de Machine Learning
- Etc.

Apéndice G. Resumen de Big Data Analytics

Big Data Analytics

Este término hace referencia a la aplicación de técnicas, métodos y algoritmos que permitan procesar y analizar volúmenes inmensos de datos (lo que se denomina big data) a fin de obtener de ellos patrones, tendencias, correlaciones o cualquier otra información valiosa que ayude a la toma de decisiones o a entender mejor los eventos pasados.

Dentro de los análisis que se hacen a los datos, existen 4 tipos:

ANÁLISIS DESCRIPTIVO

A partir del estudio y análisis de los datos pasados, se enfoca en explicar los eventos que sucedieron.

ANÁLISIS DIAGNÓSTICO

Su objetivo es determinar la causa, es decir, por qué sucedió determinado acontecimiento.

ANÁLISIS PREDICTIVO

Busca pronosticar o predecir la ocurrencia de un evento o resultado o, dicho de forma más simple, responde a la incógnita de qué va a suceder.

ANÁLISIS PRESCRIPTIVO

Este tipo de análisis se centra en definir o encontrar la ruta o conjunto de procesos adecuado para lograr un resultado determinado. A diferencia del análisis predictivo la pregunta que responde es cómo lograr que suceda.

Aplicaciones

- Administración de la información de los clientes en una empresa.
- Entendimiento y optimización de los procesos dentro de los negocios.
- Cuantificación y optimización del rendimiento personal (ejercicio físico).
- Salud pública.
- Rendimiento de máquinas y dispositivos.
- Energía.
- Seguridad en operaciones bancarias o de compra y venta.
- Mejoramiento de la seguridad ciudadana y el cumplimiento de la ley.
- Optimización del flujo de tráfico en las ciudades.
- Etc.

Apéndice H. Ecuaciones utilizadas en los procedimientos de ingeniería

Ecuaciones para calcular porosidad a partir de registros de pozo

Ecuación para el registro density

$$\phi_{den} = \frac{\rho_{ma} - \rho_b}{\rho_{ma} - \rho_f} \quad (1)$$

Donde:

ϕ_{den}: Porosidad calculada del registro density (fracción)

ρ_{ma}: Densidad de la matriz (g/cm^3)

ρ_b: Densidad del registro (g/cm^3)

ρ_f: Densidad del fluido (Generalmente, el fluido es agua y la densidad es 1 g/cm^3)

Densidad de matriz

Matriz	Densidad (g/cm^3)
Arena	2.65
Arena calcárea	2.68
Caliza	2.71
Dolomita	2.87
Anhidrita	2.92
Arcilla	2.2 – 2.85

Densidad de fluidos

Fluido	Densidad (g/cm^3)
Agua fresca	1
Agua salada (>200000 ppm)	1.1
Petróleo	0.6 – 0.9
Gas	0.01 – 0.35

Ecuación para el registro neutron

$$\phi_{eN} = \phi_{Log} - (V_{sh} * \phi_{sh}) \quad (2)$$

Donde:

ϕ_{eN}: Porosidad efectiva en fracción a partir del registro neutron.

ϕ_{Log}: Porosidad medida del registro

V_{sh}: Volumen de arcilla

ϕ_{sh}: Porosidad neutrón registrada en las arcillas

Ecuación para el registro neutron-density

$$\phi_{N-D} = \sqrt{\frac{{\phi_N}^2 - {\phi_D}^2}{2}} \quad (3)$$

Donde:

ϕ_{N-D}: Porosidad del neutrón densidad

ϕ_N: Porosidad del neutrónico (Unidades de caliza)

ϕ_D: Porosidad del perfil de densidad (Unidades de caliza)

Ecuaciones para calcular la tasa de penetración (ROP)

Modelo de Maurer

$$ROP = \frac{K}{S^2} * \left(\frac{W}{d_b} - \frac{W_o}{d_b}\right)^2 * N \quad (4)$$

Donde:

ROP: Tasa de penetración

K: Constante de proporcionalidad

S: Resistencia a la compresión de la roca

W: Peso de la broca

Wo: Umbral de peso de la broca

d_b: Diámetro de la broca

N: Velocidad de rotación

Modelo de Bingham

$$ROP = K * \left(\frac{W}{d_b}\right)^{a_5} * N \quad (5)$$

Donde:

ROP: Tasa de penetración

K: Constante de proporcionalidad (incluye el efecto de la resistencia de la roca)

W: Peso de la broca

d_b: Diámetro de la broca

a_5: Exponente de peso sobre la broca

N: Velocidad de rotación

Modelo de Bourgoyne

$$\frac{dD}{dt} = \exp\left(a_1 + \sum_{j=2}^{8} a_j\, x_j\right) \quad (6)$$

Donde:

D: Profundidad del pozo

t: Tiempo

a_1: Resistencia de la formación

a_2: Compactación de la formación

a_3: Presión de poro

a_4: Presión diferencial

a_5: Exponente de peso sobre la broca

a_6: Velocidad de rotación

a_7: Desgaste de los dientes de la broca

a_8: Impacto de la hidráulica de la broca

Modelo de AbdulJabbar et al.

$$ROP = \frac{16.96 * W^{a} * N * T * SSP * Q}{{d_b}^2 * \rho * PV * UCS^{b}} \quad (7)$$

Donde:

ROP: Tasa de penetración

W: Peso de la broca

N: Velocidad de rotación

T: Torque

SSP: Presión en la tubería vertical (Standpipe Pressure)

Q: Tasa de flujo

d_b: Diámetro de la broca

ρ: Densidad del lodo

PV: Viscosidad plástica

UCS: Resistencia a la compresión uniaxial

a y b: Coeficientes que se obtienen mediante regresión no lineal

Ecuaciones para calcular el comportamiento de afluencia de un pozo (IPR)

Correlación de Vogel

$$\frac{Q_o}{(Q_o)_{max}} = 1 - 0.2\left(\frac{P_{wf}}{\bar{P}_r}\right) - 0.8\left(\frac{P_{wf}}{\bar{P}_r}\right)^2 \quad (8)$$

Donde:

Q_o: Tasa de flujo de aceite (STB/d)

$(Q_o)_{max}$: Máxima tasa de flujo de aceite (STB/d)

P_{wf}: Presión de fondo fluyente (psig)

$\bar{P}_r$: Presión promedio del yacimiento (psig)

Correlación de Fetkovich

$$Q_o = c(\bar{P}_r - P_{wf})^n \quad (9)$$

Donde:

Q_o: Tasa de flujo de aceite (STB/d)

$\bar{P}_r$: Presión promedio del yacimiento (psig)

P_{wf}: Presión de fondo fluyente (psig)

c: Coeficiente de flujo

n: Coeficiente de contrapresión

Para determinar el valor de los coeficientes n y c se debe realizar una prueba multitasa.

Correlación de Klins-Clark

$$\frac{Q_o}{(Q_o)_{max}} = 1 - 0.295\left(\frac{P_{wf}}{\bar{P}_r}\right) - 0.705\left(\frac{P_{wf}}{\bar{P}_r}\right)^d \quad (10)$$

Donde:

$$d = \left[0.28 + 0.72\left(\frac{\bar{P}_r}{P_b}\right)\right] * (1.24 + 0.001P_b)$$

Q_o: Tasa de flujo de aceite (STB/d)

$(Q_o)_{max}$: Máxima tasa de flujo de aceite (STB/d)

P_{wf}: Presión de fondo fluyente (psig)

$\bar{P}_r$: Presión promedio del yacimiento (psig)

P_b: Presión en el punto de burbuja (psig)

Correlación de Wiggins

$$\frac{Q_o}{(Q_o)_{max}} = 1 - 0.519167\left(\frac{P_{wf}}{\bar{P}_r}\right) - 0.481092\left(\frac{P_{wf}}{\bar{P}_r}\right)^2 \quad (11)$$

Donde:

Q_o: Tasa de flujo de aceite (STB/d)

$(Q_o)_{max}$: Máxima tasa de flujo de aceite (STB/d)

P_{wf}: Presión de fondo fluyente (psig)

$\bar{P}_r$: Presión promedio del yacimiento (psig)

Correlación de Khasanov et al.

$$\frac{Q_o}{(Q_o)_{max}} = \left[1 - \left(\frac{P_{wf}}{\bar{P}_r}\right)^m\right]^n \quad (12)$$

Donde:

Q_o: Tasa de flujo de aceite (STB/d)

$(Q_o)_{max}$: Máxima tasa de flujo de aceite (STB/d)

P_{wf}: Presión de fondo fluyente (psig)

$\bar{P}_r$: Presión promedio del yacimiento (psig)

m: Se asume con un valor de 1.6 para fines prácticos

n: Es un exponente de Fetkovich

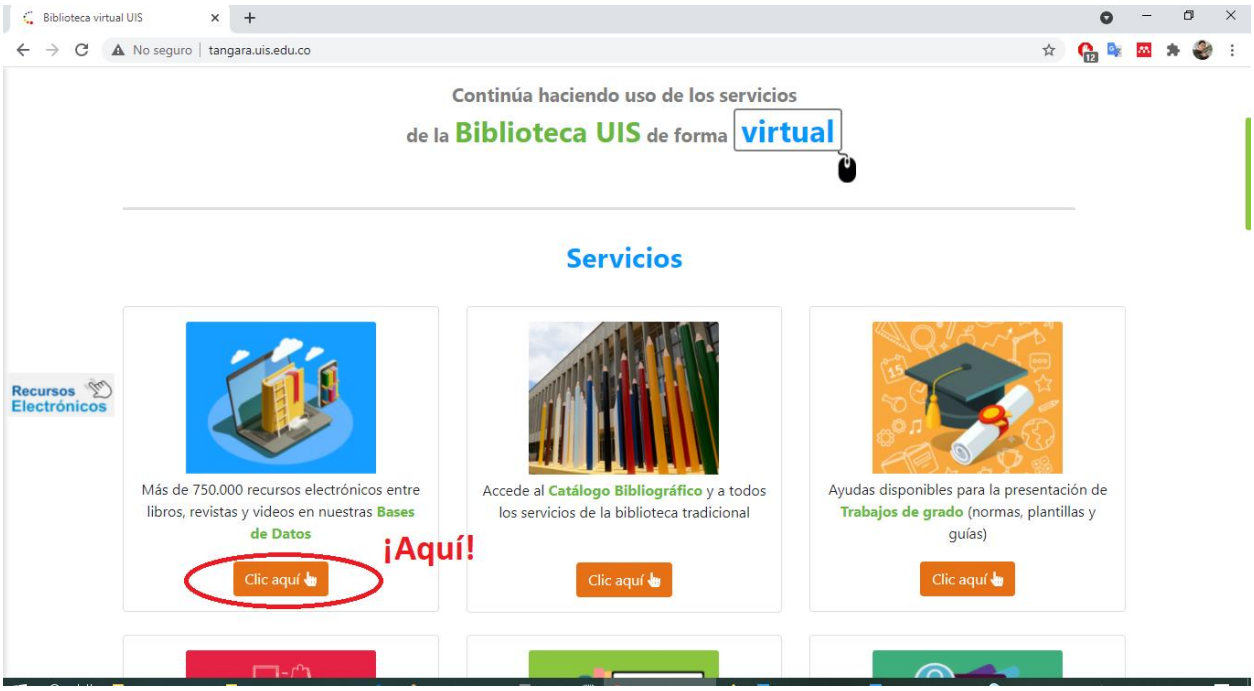
Apéndice I. Elaboración de un análisis bibliométrico con SCOPUS y VOSviewer.

I.1 Requisitos previos

- Tener acceso al portal virtual de la biblioteca de la Universidad Industrial de Santander (https://bibliotecavirtual.uis.edu.co/login). En caso de no tenerlo, comunicarse con los encargados de la biblioteca.
- Tener descargado e instalado el software VOSviewer. Este software es libre y se puede obtener en la página oficial del desarrollador (https://www.vosviewer.com/)

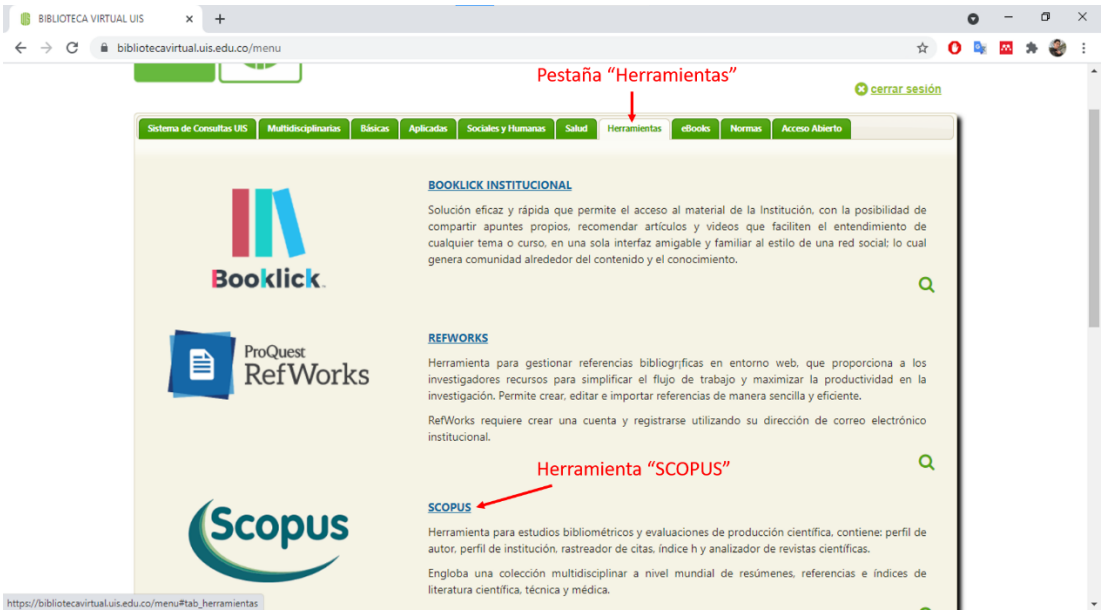
I.2 Procedimiento

Paso 1. Ingresar a la página web de la biblioteca y, en la sección "Servicios", acceder a los recursos electrónicos.



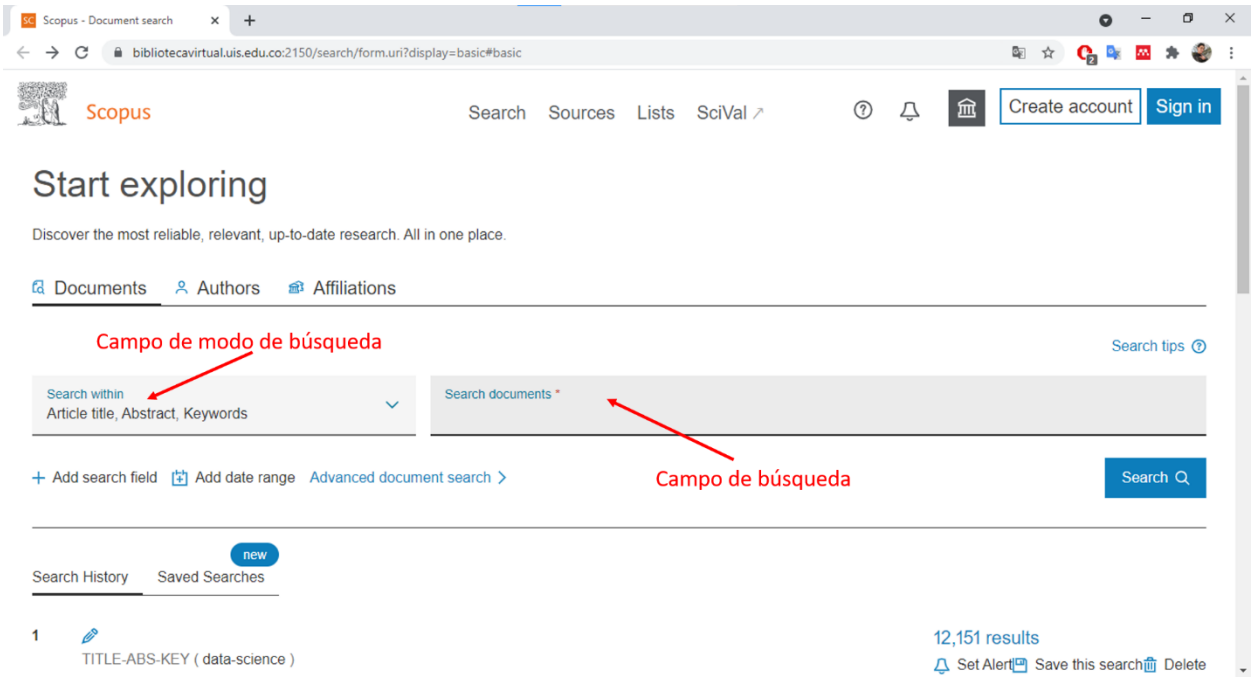
Paso 2. En la ventana que se abre, colocar las credenciales de acceso: usuario y contraseña y pulsar en el botón "Ingresar".

Paso 3. Primero, hacer clic sobre la pestaña "Herramientas" y, luego, seleccionar la herramienta "SCOPUS".



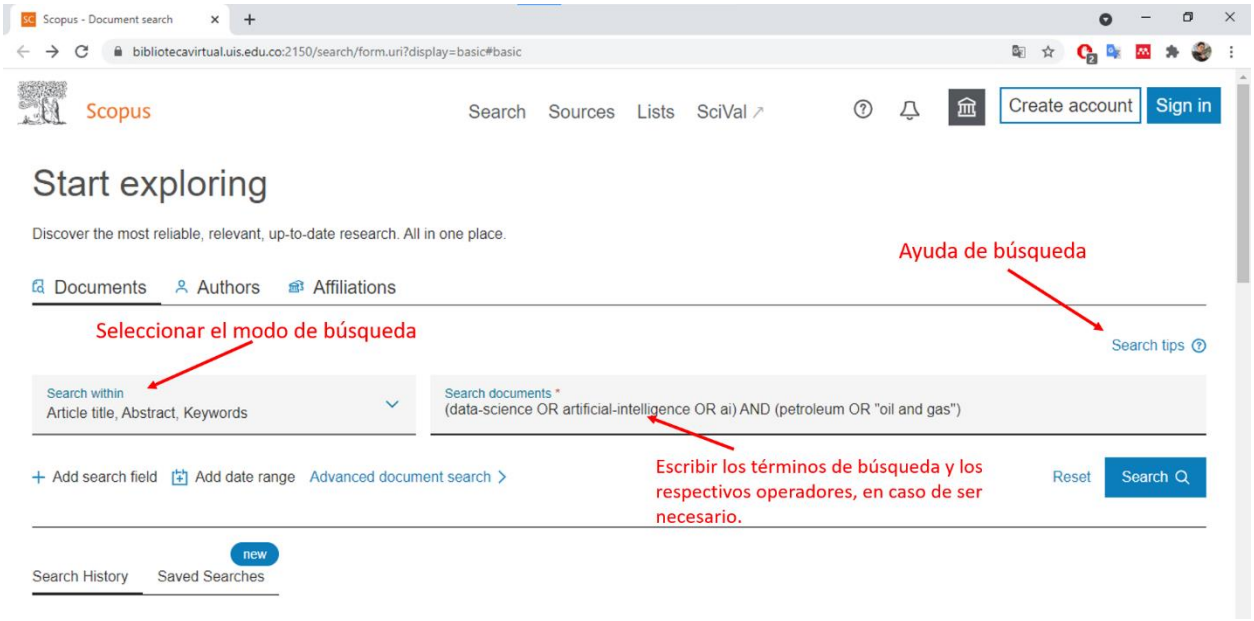
Paso 4. Se abrirá una nueva ventana con la herramienta SCOPUS. En esta ventana hay dos campos fundamentales:

- El campo de modo de búsqueda o “Search within” (“Buscar dentro”): indica dónde se buscarán los términos que el usuario proporcione. Se recomienda la opción “Article title, Abstract, Keywords”, ya que esta opción buscará los términos ingresados por el usuario en el título, el resumen y las palabras clave de los artículos que están dentro de la base de datos de SCOPUS.
- El campo de búsqueda: en este campo, el usuario ingresa los términos relacionados con la temática que desea investigar y sobre la cual hará el análisis bibliométrico. Por ejemplo, un usuario que desea realizar un análisis bibliométrico sobre la COVID-19, pondrá en el campo de búsqueda términos como “covid”, “covid-19”, “sars-cov-2”, etc.

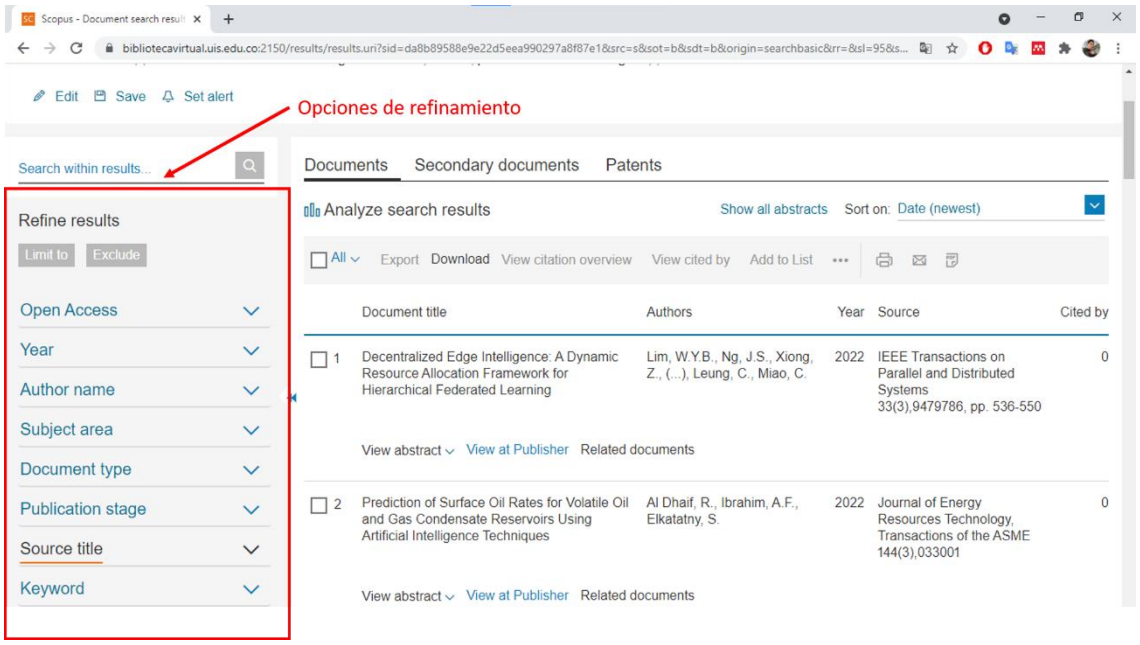


Paso 5. Seleccionar el modo de búsqueda en el campo de modo de búsqueda y, posteriormente, escribir los términos de la temática que se desea investigar en el campo de

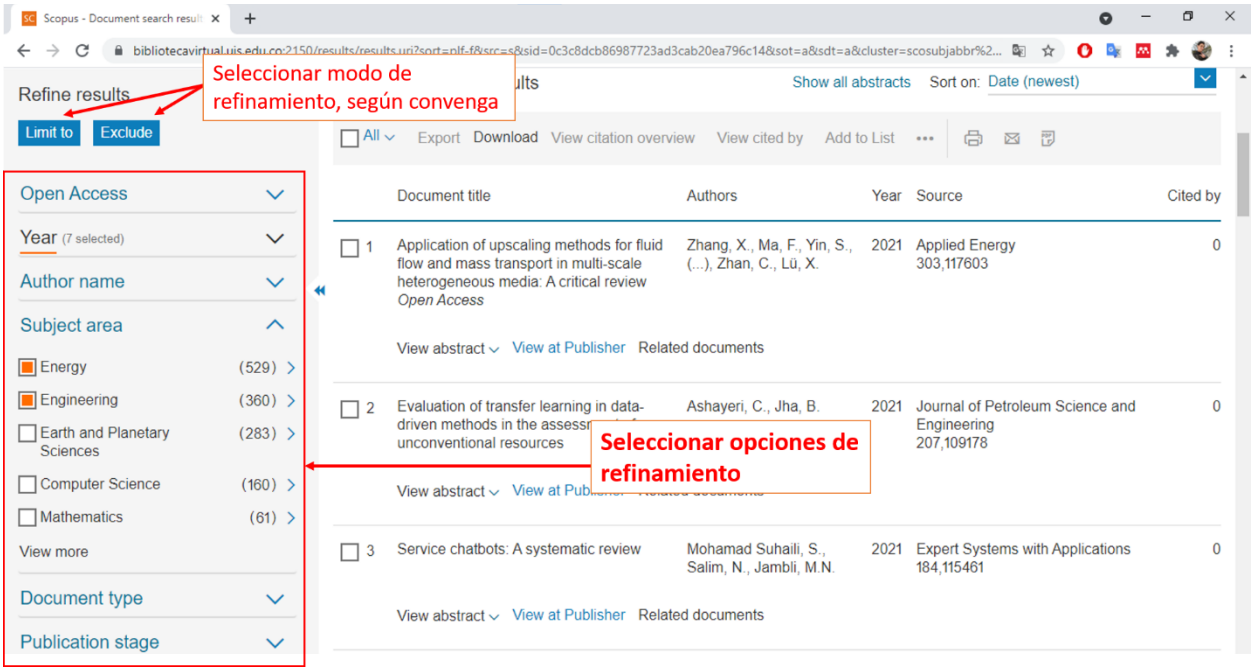
búsqueda. Los términos se pueden concatenar mediante el uso de operadores AND, OR, AND NOT y el uso de paréntesis (para más información, pulsar sobre la opción de "ayuda de búsqueda"). Una vez escritos los términos de búsqueda, hacer clic sobre el botón "Search".



Paso 6. Tras realizar la búsqueda, se procede a refinar los resultados, lo cual se hace con el fin de obtener resultados más precisos o que se ajusten mejor al interés del investigador. Para ello, en la parte izquierda de la pantalla que apareció tras hacer la búsqueda, hay un conjunto de opciones de refinamiento dentro de las que se incluye: año de publicación, nombre del autor, área de interés, tipo de documento, nombre de la revista o medio, palabras clave, instituciones, etc.



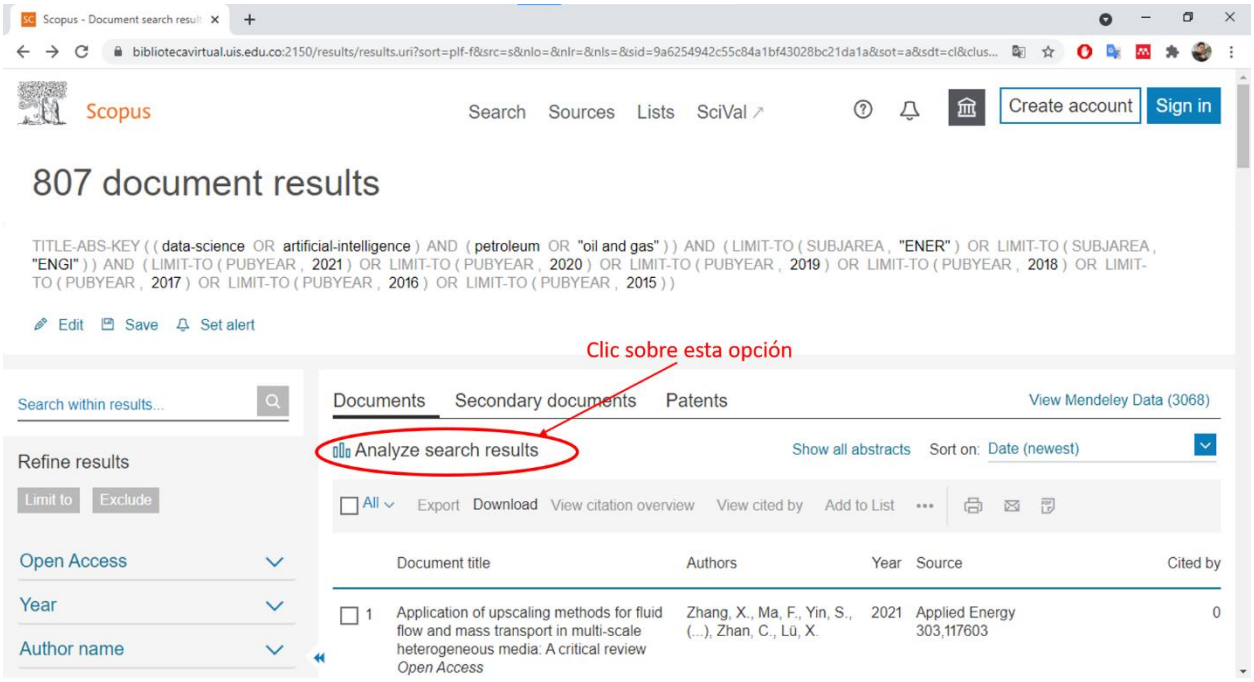
Paso 7. Seleccionar las opciones de refinamiento que mejor se adapten a las necesidades del investigador y, luego, hacer clic sobre una de las dos opciones que se habilitan: “Limit to” reduce los resultados a sólo aquellos que cumplen con las opciones de refinamiento; “Exclude” descarta los resultados que cumplen con las opciones de refinamiento y muestra los restantes.



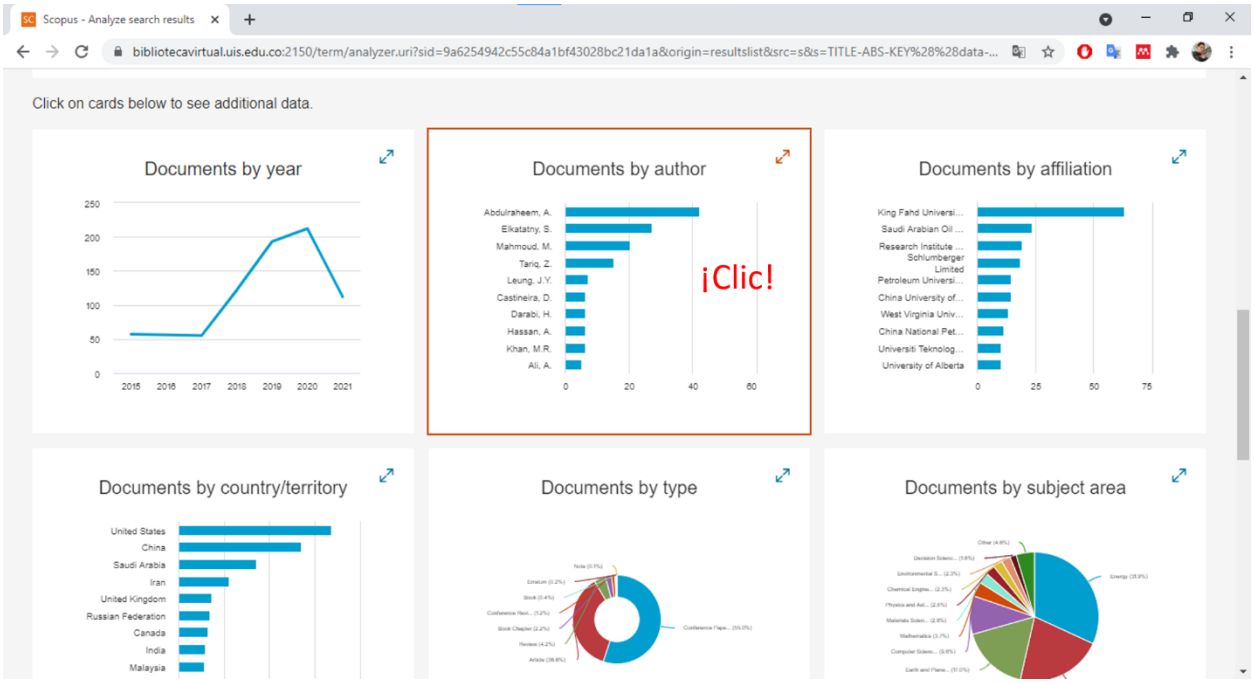
Paso 8. En este paso, es posible realizar un análisis de los resultados de la búsqueda aprovechando las opciones que ofrece la herramienta SCOPUS. La herramienta ofrece gráficas para los siguientes casos:

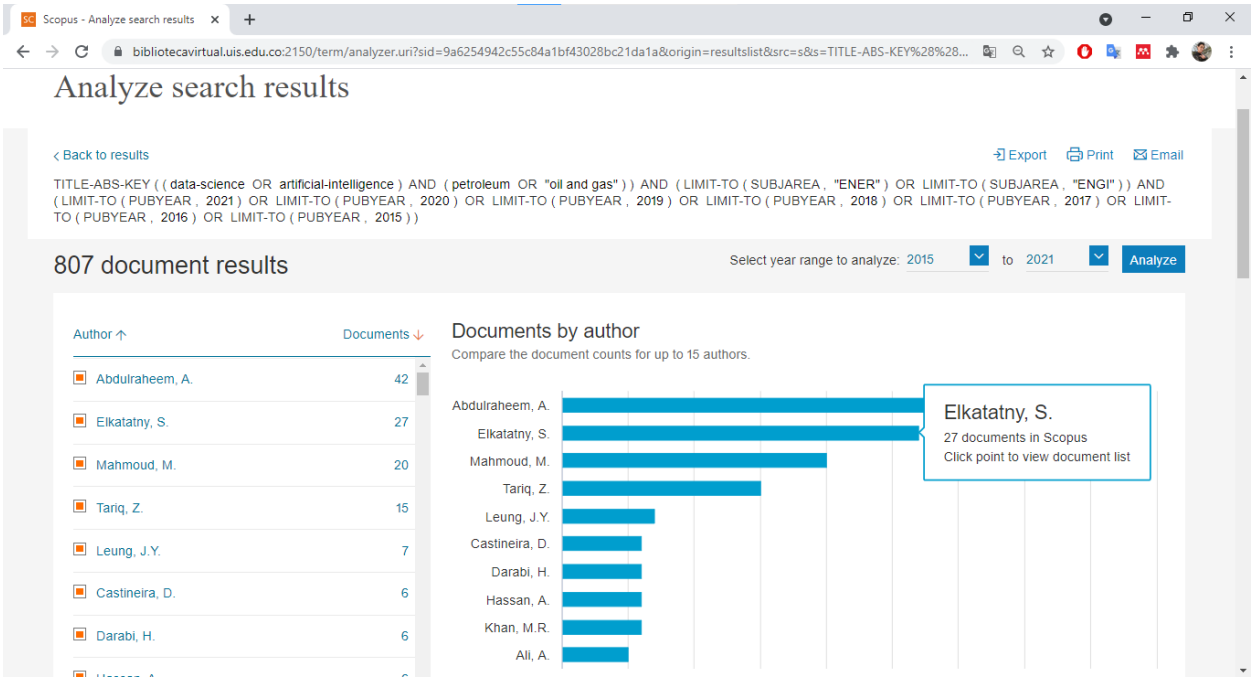
- Documentos por año
- Documentos por año por fuente (revistas y medios)
- Documentos por autor
- Documentos por afiliación (instituciones)
- Documentos por territorio o país
- Documentos por tipo
- Documentos por área temática
- Documentos por patrocinador o financiador

Para realizar el análisis, hacer clic sobre la opción “Analyze search results”



Al seleccionar la opción indicada previamente abre una pantalla como la que se muestra a continuación. Se puede hacer clic sobre la gráfica deseada para maximizarla y editarla.



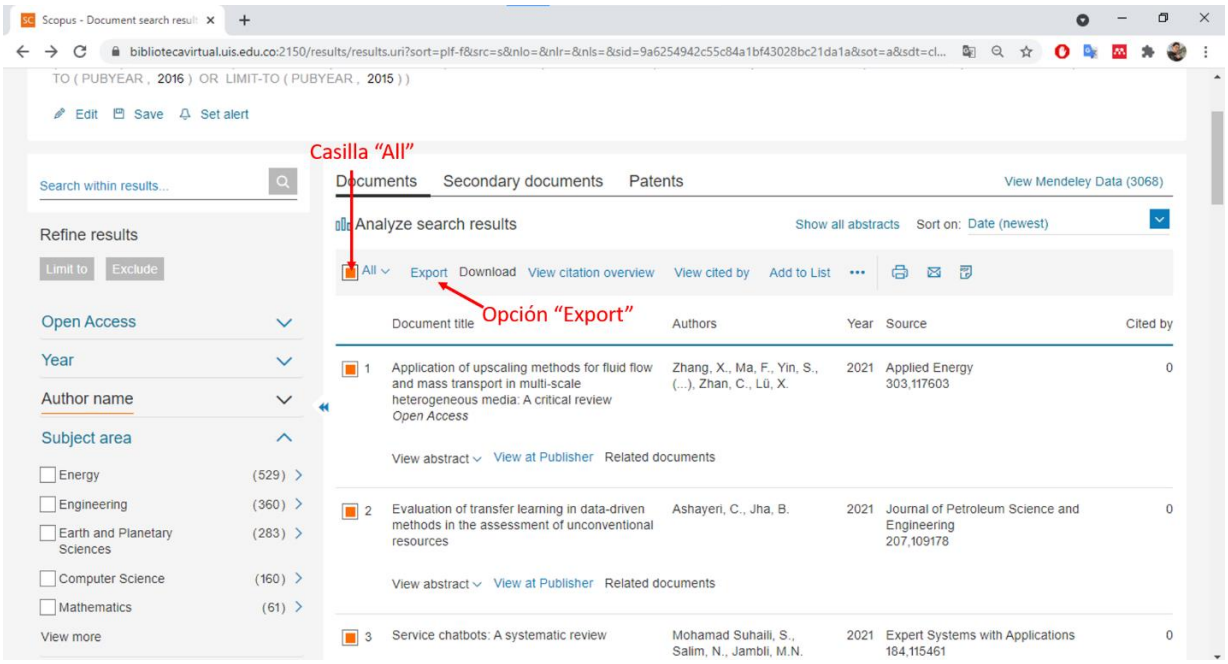


En algunas ocasiones, no es suficiente con las opciones de gráficas ofrecidas por SCOPUS y se requiere un análisis realizado sobre las palabras clave de los artículos obtenidos tras la búsqueda y refinamiento. En tales casos, es particularmente útil la herramienta VOSviewer, ya que permite elaborar redes bibliométricas basadas en palabras claves, entre otras opciones. La utilización de tal herramienta se explica en los próximos pasos.

Paso 9. Exportar los resultados de búsqueda de SCOPUS.

a. Seleccionar los resultados a exportar. Para seleccionarlos todos, se marca la casilla "All".

b. Hacer clic sobre la opción "Export".



c. Seleccionar el método de exportación. Para utilizar VOSviewer se requiere que se exporte en formato RIS (RIS format) o en formato .CSV. Para este ejemplo, se utilizará el formato RIS.

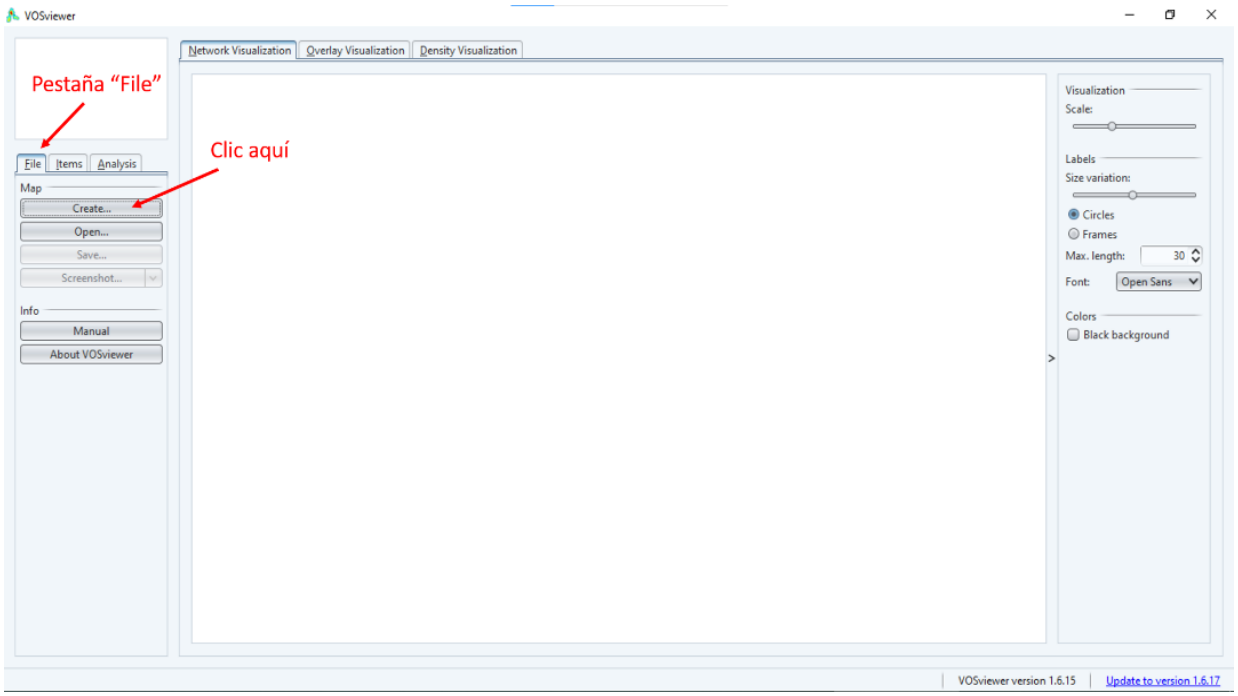
d. Seleccionar la información a exportar. Para la generación de una red bibliométrica basada en palabras clave, se recomienda marcar todas las opciones de “Abstract & keywords” y las opciones “Document title” y “Author(s)”, tal como se ilustra en la siguiente figura.

e. Para finalizar, dar clic en el botón “Export”.

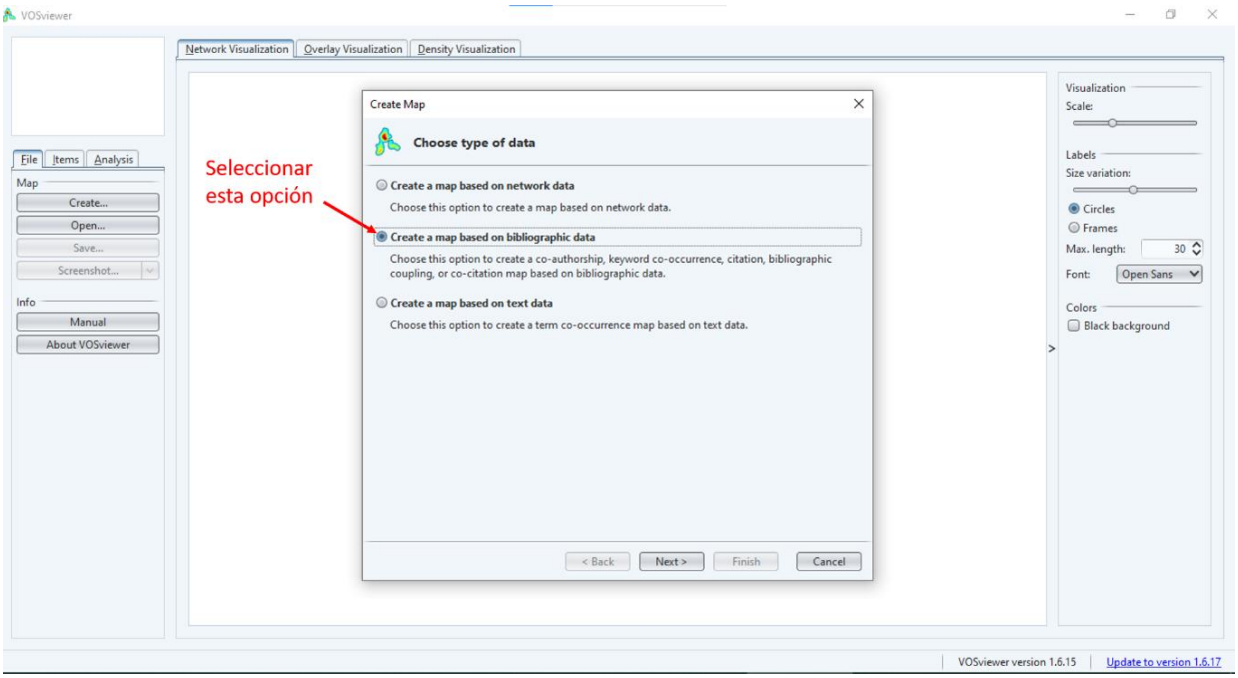
f. Al final de la exportación, se descargará un archivo llamado “scopus.ris”.

Paso 10. Abrir el software VOSviewer.

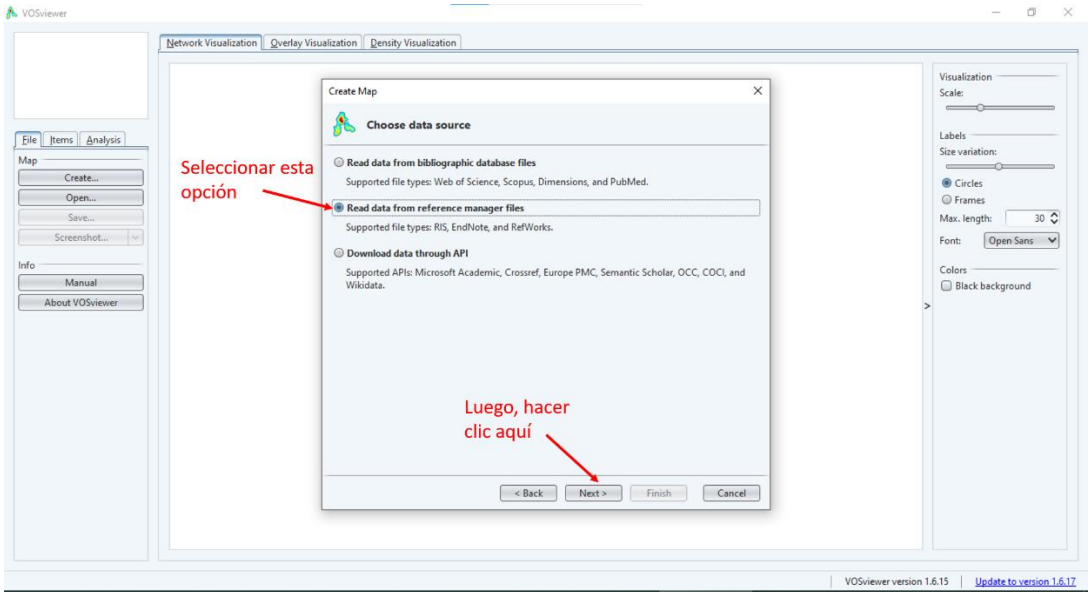
Paso 11. En la parte izquierda de la ventana, asegurarse que se encuentra en la pestaña “File”. Luego, hacer clic en el botón “Create…”.



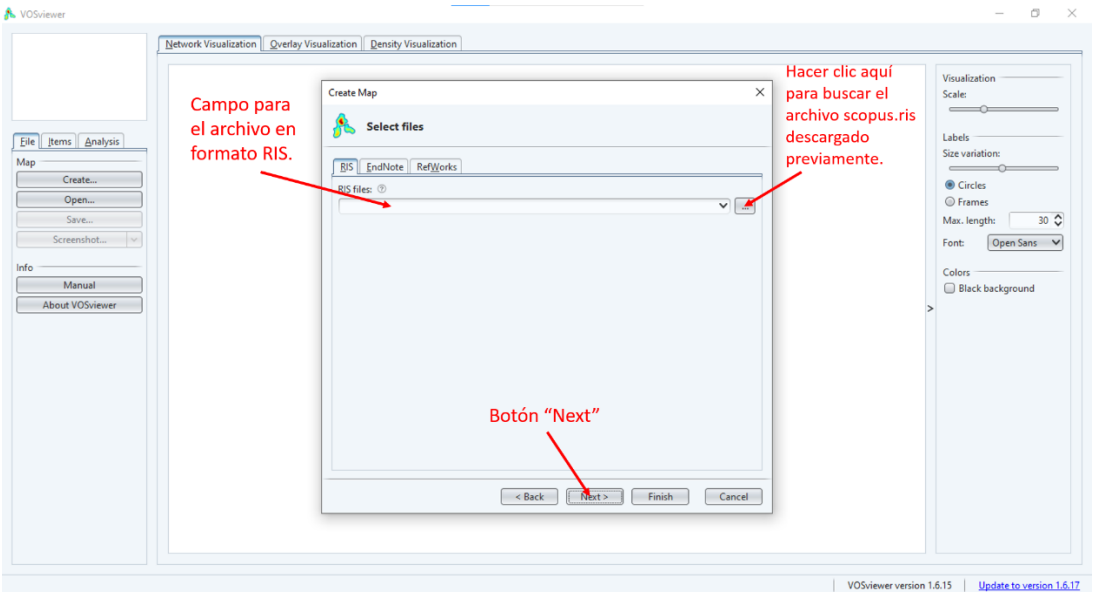
Paso 12. En la ventana que se abre, seleccionar la opción "Create a map based on bibliographic data" y pulsar en el botón "Next".



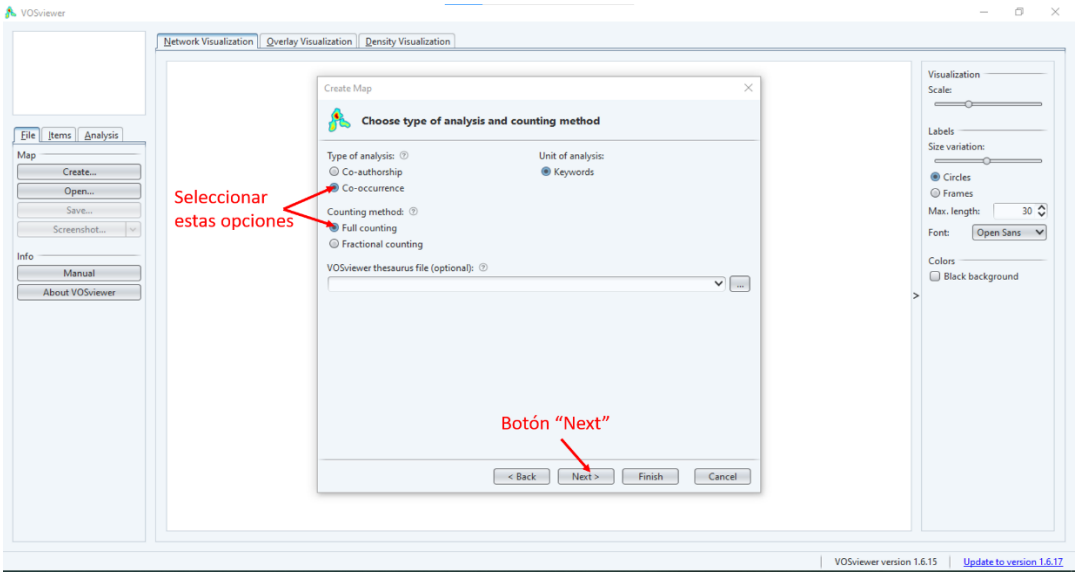
Paso 13. Seleccionar la opción “Read data from reference manager files” y, posteriormente, hacer clic en el botón “Next”.



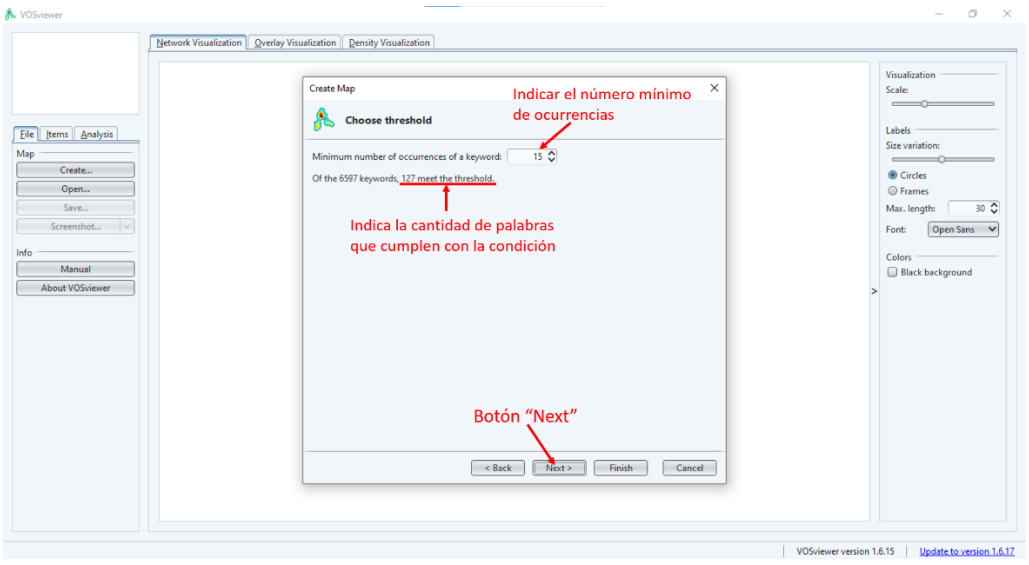
Paso 14. Hacer clic en el lugar indicado en la imagen para buscar el archivo scopus.ris que se descargó en el paso 9. Normalmente, tal archivo se guarda en la carpeta “Descargas” del computador. Finalmente, hacer clic en el botón “Next”.



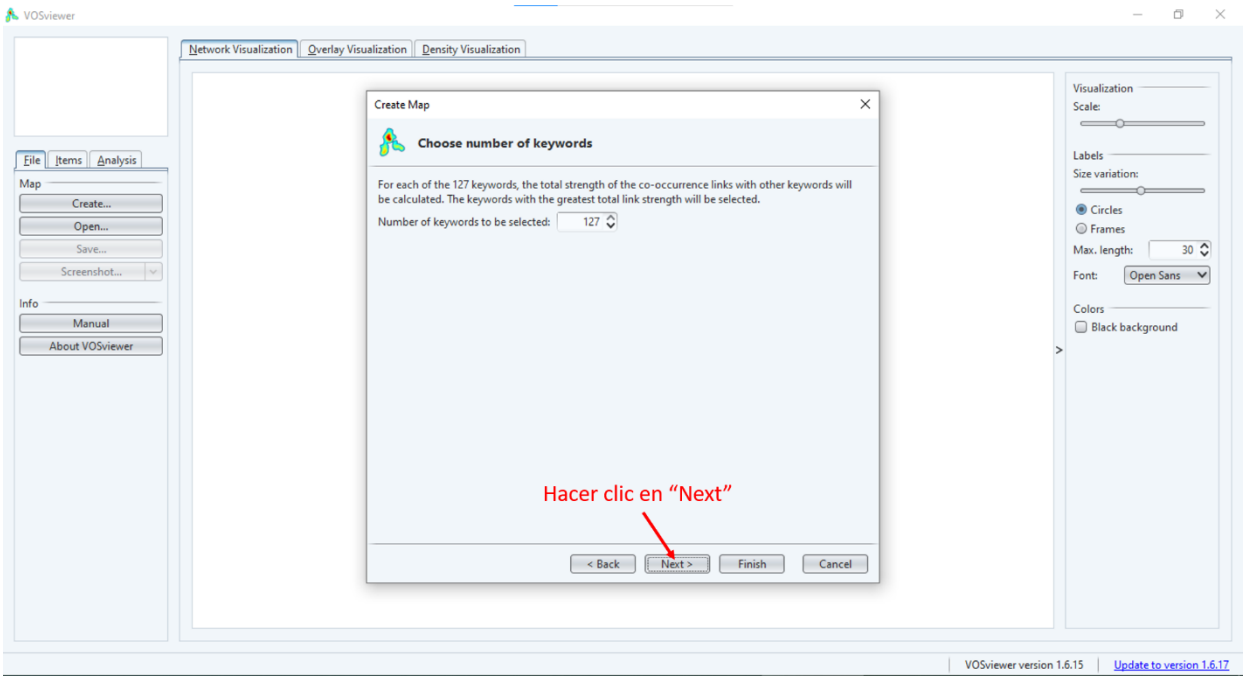
Paso 15. En tipo de análisis, seleccionar “Co-occurrence” y en el método de conteo, seleccionar “Full counting”. Luego, hacer clic en el botón “Next”.



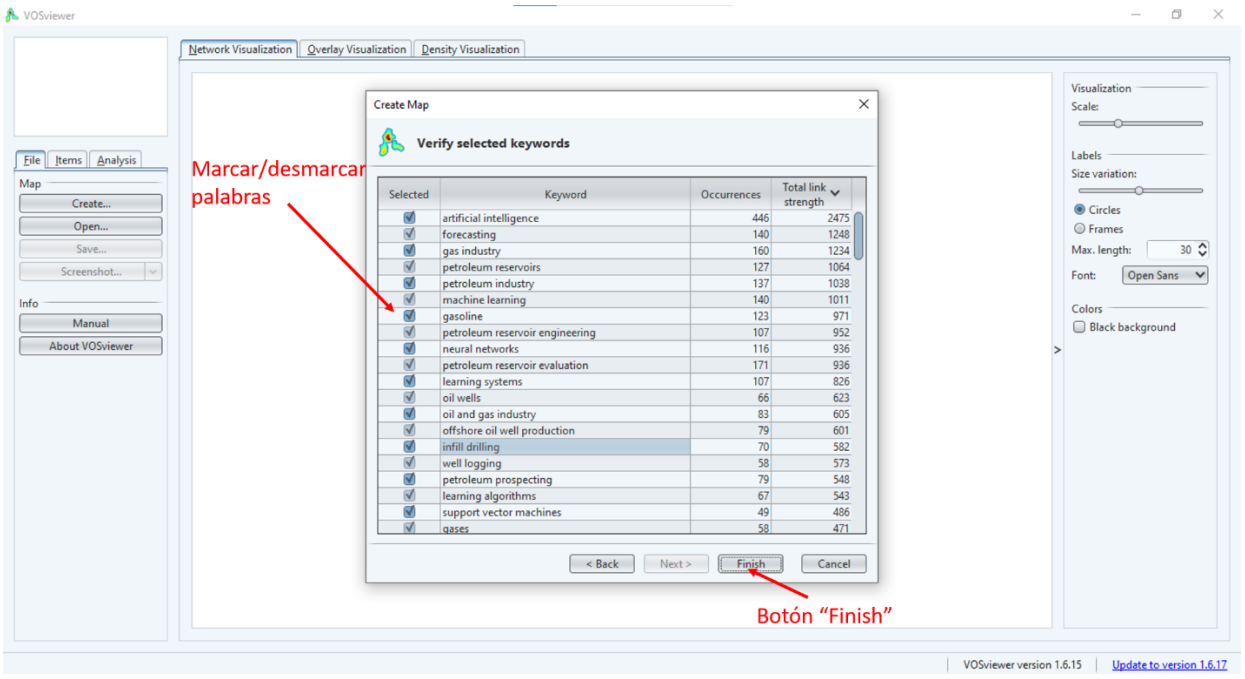
Paso 16. Seleccionar el mínimo número de ocurrencias de las palabras clave, es decir, la cantidad mínima de veces que aparecen las palabras clave en el título, resumen y sección de palabras clave de todos los documentos que fueron tenidos en cuenta tras la búsqueda y refinamiento en los pasos 5, 6 y 7. Esta condición dependerá de las necesidades del investigador. Finalmente, hacer clic en el botón “Next”.



Paso 17. En la ventana que aparece, hacer clic en el botón “Next”.



Paso 18. En la columna “Selected”, desmarcar aquellas palabras que no sean relevantes para el estudio y que, por lo tanto, no se tendrán en cuenta en la red bibliométrica. Finalmente, hacer clic en el botón “Finish”.



Paso 19. ¡La red ha sido generada! Con las opciones en el costado derecho de la ventana del software se puede personalizar la red: tamaño de fuente, colores, líneas, etc. En las opciones del costado izquierdo de la ventana del software es posible guardar la red actual para visualizarla luego, hacer una captura de pantalla o, bien, crear una nueva red.