UNIVERSIDAD
NACIONAL
DE COLOMBIA

MÉTODO PARA EL CÁLCULO DEL PORCENTAJE DE AGUA EN EMULSIONES AGUA-CRUDO USANDO APRENDIZAJE DE MÁQUINAS

Sebastian Echeverri Parra

Universidad Nacional de Colombia

Facultad de Minas

Maestría en Ingeniería – Ingeniería de Sistemas

2023

MÉTODO PARA EL CÁLCULO DEL PORCENTAJE DE AGUA EN EMULSIONES AGUA-CRUDO USANDO APRENDIZAJE DE MÁQUINAS

Sebastian Echeverri Parra

Trabajo Final presentado como requisito parcial para optar al título de:

Magíster en Ingeniería – Ingeniería de Sistemas

Directores:

John Willian Branch Bedoya, Ph.D.

Camilo Franco Ariza, Ph.D

Farid Cortés, Ph.D

Universidad Nacional de Colombia

Facultad de Minas

2023

Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada con el respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

_____

Sebastian Echeverri Parra

Fecha 31/01/2023

Agradecimientos

Al profesor John Willian Branch, por acoger el proyecto, y por su acompañamiento y guía a lo largo de todo el proceso. Gracias por su esmero en fomentar espacios de divulgación y transferencia de conocimiento entre toda la comunidad académica.

Al profesor Farid Cortés, Camilo Franco y el equipo de laboratorio por su apoyo en la recolección de muestras, comentarios e ideas que hicieron parte de brindarme todo el contexto posible sobre el mundo de las emulsiones.

A la Universidad Nacional de Colombia por varios años de formación, apoyo y aprendizaje constante, tanto académico como personal.

A mi esposa, mi familia y todas las personas que mostraron su apoyo durante todo el proceso de desarrollo y ejecución de este trabajo.

Al Todopoderoso papá Dios quien es mi maestro, mentor y compañero inseparable

Método para el cálculo del porcentaje de agua en emulsiones de agua-crudo usando aprendizaje de máquinas.

Resumen

Se realizó un estudio sobre las emulsiones de agua en aceite (W/O) en crudos pesados provenientes de las imágenes adquiridas del laboratorio de Fenómenos de Superficie Michael Polanyi, mediante el uso de herramientas de microscopia óptica. El interés por el estudio de estas imágenes se crea con el fin de determinar el porcentaje de agua contenido en una emulsión llamado por las siglas %BSW (Basic Sediment and Water) sin hacer uso de técnicas químicas que requieren inversión de recursos. Para abordar la solución se emplea un análisis de características, determinando las principales variables que pueden ser extraídas de las imágenes obtenidas del laboratorio, como lo es la cantidad de partículas, el tamaño, la masa, la señal, la excentricidad, el épsilon, la ubicación de las partículas, todo realizado por medio de un método de visión por computador que permite identificar estos atributos en las emulsiones con un algoritmo basado en el seguimiento de partículas. Una vez son extraídas las variables se emplean técnicas estadísticas para identificar tendencias, correlaciones y selección de variables para un posterior uso de técnicas de aprendizaje de máquina supervisado como lo son los Bosques Aleatorios, Regresión Logística Multinomial, Redes Neuronales y de la Potenciación del Gradiente, evaluando en cada modelo el rendimiento de acuerdo a la capacidad de predicción del contenido de agua %BSW con varias medidas de validación y también por medio de imágenes encontradas en la literatura se comprueba el rendimiento del método seleccionado, con respecto a los resultados indicados por el autor del artículo. Al final se presenta una interfaz de consulta con las soluciones previamente mencionadas.

Palabras clave: Emulsiones (W/O), Aprendizaje de máquinas Supervisado, Visión por computador, %BSW

Method for calculating the percentage of water in crude-water emulsions using machine learning.

Abstract

A study was carried out on the water-in-oil (W/O) emulsions in heavy crudes from the images acquired from the Michael Polanyi Surface Phenomena laboratory, through the use of optical microscopy tools. Interest in the study of these images is created in order to determine the percentage of water contained in an emulsion called %BSW (Basic Sediment and Water) without using chemical techniques that require investment of resources. To approach the solution, an analysis of characteristics is used, determining the main variables that can be extracted from the images obtained from the laboratory, such as the number of particles, the size, the mass, the signal, the eccentricity, the epsilon, the location of the particles, all carried out by means of a computer vision method that allows the identification of these attributes in the emulsions with an algorithm based on particle tracking. Once the variables are extracted, statistical techniques are used to identify trends, correlations and selection of variables for a later use of supervised machine learning techniques such as Random Forests, Multinomial Logistic Regression, Neural Networks and Gradient Boosting, evaluating the performance of each model according to the predictive capacity of the %BSW water content with several validation measures and also through images found in the literature, the performance of the selected method is verified, with respect to the results indicated by the article author. At the end, a consultation interface is presented with the previously mentioned solutions.

Keywords: Emulsions (W/O), Supervised Machine Learning, Computer Vision, %BSW

Tabla de contenido

Lista de figuras

Lista de tablas

Lista de ecuaciones

1. Introducción

1.1 Motivación

Las emulsiones son definidas como un sistema compuesto por dos líquidos inmiscibles, uno de los cuales se dispersa como pequeñas gotas (la fase dispersa o interna) a lo largo de la otra (la fase continua o externa) (Mahdi Jafari & Bhandaric, 2008). Son clasificadas como aceite en agua (O/W) o agua en aceite (W/O) y cuenta con propiedades fisicoquímicas como la concentración (dispersión de las partículas), viscosidad, tensión superficial, tamaño de la gota, volumen y temperatura.

Las emulsiones de agua en aceite, claramente, hacen parte de los principales problemas con los que la industria del petróleo tiene que batallar durante la recuperación, tratamiento, operaciones de producción y transporte del crudo (Wong, Lim, & Dol, 2015) (Langevin, Poteau, & Argillier, 2004). La presencia de emulsiones reduce la calidad del crudo, afectando generalmente el comportamiento del flujo del fluido y puede conducir a problemas significativos de aseguramiento del flujo (Wong, Lim, & Dol, 2015) (Omer & Pal, 2010), conducir a una mayor caída de presión (Pettersen & Sjöblom, 2012), lo que indica una mayor pérdida de energía, ya que se requiere una mayor energía de bombeo para mantener el flujo de masa deseado, siendo el agua un contaminante para la mayoría de los derivados del petróleo al reducir las propiedades lubricantes y favoreciendo en la degradación del aceite. Haciendo que los productores de petróleo estén obligados a emplear una variedad de técnicas de calentamiento y desemulsionantes químicos con el fin de aumentar la velocidad y la eficiencia de separación agua en crudo (reducir el contenido de agua). Estas técnicas

pueden ser costosas y consumir tiempo, siendo deseable desarrollar nuevos métodos menos costosos para la desestabilización de estas emulsiones (Blanco & Peguero, 2008).

En particular dentro del entendimiento de las emulsiones la caracterización y el modelado se han venido beneficiando de las técnicas de procesamiento de imágenes la cual es un área muy relevante de la informática con aplicaciones en el área química donde el análisis cuantitativo y la interpretación de imágenes digitalizadas son actualmente herramientas importantes para la toma de decisiones, donde el proceso básicamente consiste en una cámara digital o de video o en su defecto en el presente trabajo de un microscopio óptico, un software de computadora para el análisis de imágenes (Colucci, Morra, Zhang, Fissore, & Lamberti, 2020), que nos permita caracterizar cuantitativamente las propiedades complejas de tamaño, forma, color y distintos factores que contribuyan con el entendimiento de la emulsión.

La adquisición mediante microscopia óptica y el tratamiento de imágenes de emulsiones son pasos fundamentales para el cálculo de la distribución del tamaño y forma de las partículas contenidas en ello, el objetivo del presente trabajo es, sin lugar a dudas, desarrollar una metodología que permita determinar las características de las variables independientes que contribuyen con identificar en una emulsión W/O el porcentaje de agua (%BSW) contenido dentro de sus propiedades. Esto es importante dentro del análisis y la interpretación de los resultados, donde autores como Rod y Misek (Ribeiro, Guimarães, Madureira, & Cruz Pinto, 2004) han demostrado que se producen errores o sesgos cuando la técnica basada en un muestreo físico representativo cambiará drásticamente la composición general de la dispersión en la determinación de las características y la calidad del crudo.

1.2 Trabajos previos

Diferentes autores a lo largo de los años han propuesto diversas técnicas y métodos para la caracterización de emulsiones. Con base en (Henríquez, 2009), quien planteo un análisis completo tomando en cuenta una formulación, experimentación y análisis, describiendo mediante el uso de la microscopía óptica, las capacidades de video digital y software de análisis de imágenes, la caracterización de una emulsión W/O contabilizando varios miles de partículas y sus áreas proyectadas determinando la estabilidad de la emulsión por la cantidad de agua y aceite separados después de 30 días.

Otro tipo de análisis que se encuentra lo realizó (Kallevik, Brunsgaard Hansen, Sæther, Kvalheim, & Sjöblom, 2000), mediante análisis multivariado y perfiles espectroscópicos infrarrojos (NIR) en el rango de 1100- 2250 nm (nanómetros), comprueba que la composición de la fase de aceite (cantidad de agua en la muestra de emulsión) se puede predecir a partir de los espectros NIR y mediante una reducción de dimensionalidad (PCA) para evitar información redundante o correlacionada, selecciona los pesos de las variables para ser asignados a una regresión parcial de mínimos cuadrados(PSL). Esta misma técnica de PSL es aplicada en conjunto con un modelo de redes neuronales artificiales (ANN) por (Bampi, P. Scheer, & Castilhos, 2013) con el fin de predecir el tamaño promedio de una gota y el contenido de agua en un biodiésel.

Dentro de las propuestas de los diferentes autores se identifica una tendencia de selección de instrumentos NIR, modelos empíricos PLS, ANN para el análisis de contenido y tamaño de gota promedio de agua en emulsiones en conjunto con una validación de la utilidad del método mediante

la precisión de la predicción superior al 90% en todos los casos para cada método empleado (Araujo, Santos, M., Fortuny, & Montserrat, 2008) (Balabin, Lomakina, & Safieva, 2011).

1.3 Descripción del problema

En el proceso de adquisición y análisis de las emulsiones, se presenta actualmente una metodología de prueba de botella como estándar que es mecánico o repetitivo con la posibilidad de sesgos en el resultado, que requiere de tiempo y recursos (material y personal) (Olalekan S, Mahmoud, Al Shehri, & Sultan, 2021). Por lo antepuesto, en el presente trabajo final de maestría, se propone un método basado en aprendizaje de máquinas para realizar una predicción del cálculo del porcentaje de agua de acuerdo a las características identificadas en las micrografías de las emulsiones de agua en aceite (W/O). Este método ofrece una solución que permite a partir de una imagen capturada desde un microscopio óptico llegar a obtener el porcentaje de agua contenida (%BSW).

1.4 Objetivos

1.4.1 Objetivo general

Proponer un método para la predicción del contenido de agua en una emulsión de un crudo pesado empleando técnicas de aprendizaje de máquinas.

1.4.2 Objetivos específicos

- Identificar las variables y parámetros que influyen en determinar el contenido de agua en una emulsión de un crudo pesado.

- Proponer un método para la predicción del contenido de agua en una emulsión de un crudo pesado empleando técnicas de aprendizaje de máquinas supervisado.

- Evaluar el desempeño del método de predicción propuesto con respecto a los resultados obtenidos y a través de los resultados en la literatura.

1.5 Alcance

En este trabajo final de maestría, se llega al diseño e implementación de un método basado en técnicas de aprendizaje de máquinas, para la predicción del porcentaje de agua contenido en una emulsión de un crudo pesado mediante la información adquirida en el laboratorio (imágenes extraídas desde el microscopio óptico).

1.6 Estructura del documento

En el presente documento se ilustra el proceso de investigación realizado para llegar al método de predicción del contenido de agua sobre la emulsión, así como el contexto sobre el tema central de investigación y una serie de resultados y reflexiones resultado de todo el proceso.

En el capítulo 2 se recoge una explicación teórica acerca del problema sobre los diferentes métodos aplicados desde la visión por computador para la identificación de las variables principales

y los recientes aportes, por parte del aprendizaje de máquinas para dicho problema, argumentando cómo y en qué escenarios resulta especialmente útil e interesante la aplicación de estas técnicas, al momento de realizar una predicción sobre la imagen de la emulsión. Por otro lado, en el capítulo 3 se realiza una revisión de la literatura existente en la actualidad acerca de este tema, mencionando aquellos estudios e investigaciones publicadas que han sido más relevantes, de cara a la realización del proyecto y que más han inspirado el proceso de investigación y el desarrollo del método propuesto.

La arquitectura y detalles del diseño de dicho método, así como su desarrollo e implementación, su posterior evaluación y puesta a prueba, se indican de forma extensa en el capítulo 4. Consecuentemente, a continuación, en el capítulo 5 se hace referencia al tercer objetivo específico definido, mostrando un análisis del método propuesto, contrastándolo principalmente con los métodos e información encontrada en la literatura, mostrando los resultados obtenidos con la implementación del método. Y para finalizar el desarrollo de cada una de las actividades mencionadas, en el capítulo 6 se presenta una arquitectura e interfaz de consulta para el usuario de interés.

Por último, el capítulo 7 y 8, recoge una serie de conclusiones obtenidas a lo largo del desarrollo del proyecto, así como reflexiones e ideas de trabajo futuro al respecto de este tema, indicando finalmente una serie de propuestas bien definidas que serán realizadas con posterioridad.

2. Marco teórico de referencia

A continuación, se desarrollarán cada uno de los conceptos que han sido claves para entender el desarrollo del problema específico de investigación y la solución propuesta para este.

Emulsiones (O/W)

Una emulsión es una dispersión de dos líquidos inmiscibles que se estabilizan generalmente con moléculas tensioactivas (Aranberri, Binks, Clint, & Fletcher, 2006). Su tamaño de gota suele ser del orden de micras, por lo que presentan un aspecto lechoso que consiste en gotitas de agua en una fase continua de aceite (Figura 1), en la industria petrolera, las emulsiones de agua en aceite son más comunes (la mayoría de las emulsiones de campos petroleros producidas son de este tipo) (Lendínez Gris, 2015).

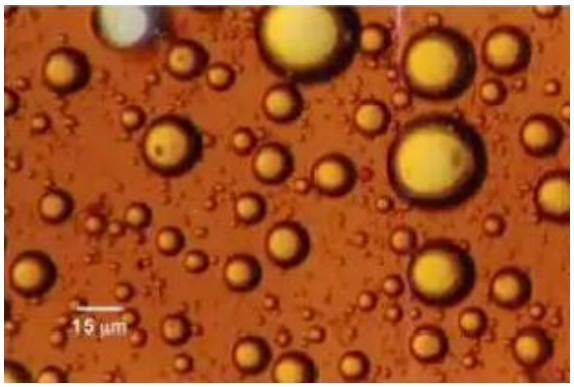

Figura 1: Fotomicrografía de una emulsión de agua en aceite (SPE INTERNATIONAL, 2015).

Tamaño de la gota

La distribución del tamaño de las gotas normalmente se representa por un histograma o una función de distribución de algún tipo.

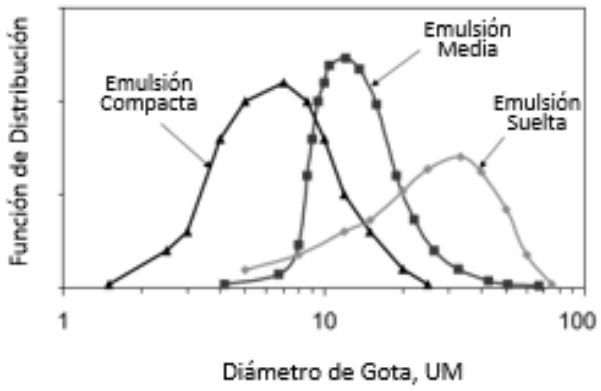

Figura 2: Distribución del tamaño de la gota medida en micrones (Kokal, 2005).

Dividida en tres tipos como lo son; la emulsión suelta que indica gotas grandes y de distribución amplia siendo fácil de descomponer, emulsión compacta con pequeñas gotas y una distribución cerrada siendo difícil de romper, emulsión media con gotas con una distribución cerradas con gotas tanto pequeñas como grandes demorando un mayor tiempo en romperse (Kokal, 2005).

Generalmente, las emulsiones que tienen gotas de tamaño más pequeño serán las más estables. Para la separación del agua, las gotas tienen que coalescer (unirse para formar una fase continua), y cuanto más pequeñas sean las gotas, más tiempo tardará en separarse.

La distribución del tamaño de las gotas afecta la viscosidad de la emulsión, siendo mayor cuando las gotas son más pequeñas y también es mayor cuando la distribución del tamaño de gota es compacta (Kokal, 2005).

Microscopia y análisis de imágenes

La observación directa o la fotográfica en microscopia óptica son el método más simple, y el único que puede considerarse como absoluto; permite al operador pronunciar un juicio, subjetivo, acerca del tamaño o de la forma de las gotas. Sin embargo, se vuelve extremadamente monótono y a menudo es difícil hacer medidas objetivas, como la determinación del tamaño de la gota. (Albert & Méndes, 2014).

Son operaciones de determinación de tamaño y conteo de gotas. En los aparatos modernos, la imagen es analizada por un detector fotoeléctrico de barrido, semejante a una filmadora de televisión, que transforma la información óptica en una señal de video. Dicha señal, está luego manipulada por un sistema computarizado, cuyo análisis está limitado solo por la sofisticación del programa y la capacidad de computación del aparato (Figura 3), conviene recordar que el análisis de imagen está siempre limitado por la precisión del microscopio que se usó para tomar la foto. (Albert & Méndes, 2014).

Figura 3: Equipo de microscopia óptica (Servicios Científicos Técnicos, Universidad de Oviedo, s.f.).

Prueba %BSW (Basic Sediment and Water)

Corresponde al contenido de agua libre y sedimentos que hay en una muestra de hidrocarburos (limo, arena) que trae el crudo, permitiendo evaluar la calidad del crudo mediante tubos de centrifugado en forma de cono de 100 ml (Utria Robinson, 2017).

Dado por (Ecuación 1):

Ecuación 1: Fórmula del %BSW (Utria Robinson, 2017).

$$\%\text{BSW} = \frac{\text{Vol. Agua y sedimientos}}{\text{Vol. total}} * 100$$

El determinar el valor del %BSW nos indica cuanta cantidad de agua y sedimentos se están produciendo por cada barril de aceite que se tienen, conocido también como el corte agua. Siendo

importante el proceso porque el identificar esta propiedad en los crudos, regula tanto su calidad como su precio (Forero, Ortíz, Nariño, Díaz, & Peña, 2008).

Inteligencia Artificial

La inteligencia artificial (IA) es la rama de las ciencias computacionales que se encarga del diseño y construcción de sistemas capaces de realizar tareas asociadas con la inteligencia humana, mediante la construcción de artefactos que pueden desarrollar conocimiento, aprendiendo de la experiencia, leyendo y procesando textos escritos en lenguajes naturales, razonar con el conocimiento adquirido (capaz de realizar tareas tales como explicar, planificar, diagnosticar, etc.) y actuar racionalmente.

Sus aplicaciones van desde el reconocimiento en imágenes o video de objetos y personas, hasta el habla y la traducción automática de textos, pasando por el diagnóstico y tratamiento de enfermedades y la toma de decisiones que han venido en aumento dado que cada vez hay una mayor disponibilidad de datos, tecnológicos y financieros, al igual que, el avance en las técnicas de aprendizaje computacional donde la transparencia y la confiabilidad de las técnicas son factores claves en la representatividad y en las decisiones fundamentadas (Ocampo, 2018) debido a que la proporción de errores significativamente menores, en las máquinas que realizan las mismas tareas que sus contrapartes humanas (Rouhiainen, 2018).

Visión por computador

La visión por computador es un campo en el que se estudian las distintas características y descriptores de las imágenes obtenidas mediante instrumentos de captura de imágenes (cámaras,

espectrómetros, instrumentos de resonancia magnética, instrumentos de rayos X). Haciendo posible tomar decisiones inteligentes con la información de los descriptores y características obtenidas mediante esta técnica. Entre los procesos más comunes de visión por computador con técnicas de inteligencia artificial se encuentra la detección, la clasificación y el rastreo de objetos, y la estimación de pose, ubicación y volumen de objetos (Toquica Cáceres, 2020).

Las imágenes manejadas por el computador son bidimensionales debido a que el proceso de muestreo de la escena del mundo real se realiza, normalmente, a través de una rejilla rectangular, donde cada uno de sus elementos se conoce como un píxel (picture element). Cada píxel puede almacenar valores de un tipo concreto, dependiendo del formato de representación de la imagen.

Estos procesos de reconocimiento de imágenes, memorización de la información, interpretación permiten (Maisa, 2019):

- Automatizar tareas repetitivas de inspección.
- Realizar controles de calidad ante posibles métodos tradicionales.
- Realizar inspecciones de objetos evitando contacto físico.
- Inspeccionar el 100% de la producción a mayor velocidad.
- Reducir tiempos en procesos automatizados.
- Verificar la diversidad de piezas u objetos con cambios frecuentes de producción.

Aprendizaje de Máquinas

Es uno de los enfoques principales de la inteligencia artificial (Rouhiainen, 2018), siendo un campo basado en la informática, las matemáticas y la estadística centrado en el diseño de modelos (a veces inspirados biológicamente) para el análisis automático de datos a gran escala. A diferencia de los algoritmos creados para obtener soluciones exactas, los modelos de aprendizaje de máquinas tienen un fuerte enfoque en soluciones aproximadas. Existen tres áreas principales dentro del aprendizaje automático:

- El aprendizaje supervisado, donde el foco está en la predicción precisa basada en conjuntos de datos etiquetados u organizados previamente para indicar como tendría que ser categorizada la nueva información, siendo necesaria la intervención humana para proporcionar retroalimentación (Rouhiainen, 2018).
- El aprendizaje no supervisado, donde el objetivo es encontrar grupos comunes en los datos sin contar con ninguna etiqueta u organizado previamente para indicar cómo tendría que ser categorizada la nueva información, sino que se debe encontrar la manera de ser clasificada por sí mismo sin la necesidad de requerir la intervención humana (Gershgorn, 2022).
- En el aprendizaje por refuerzo los algoritmos aprenden de la experiencia, donde se le da algún tipo de “premio y castigo” cada vez que aciertan o falla el objetivo (Gershgorn, 2022). (Bagnato, 2020).

Red Neuronal Artificial (ANN)

Es un procesamiento de información inspirado en la forma en que los sistemas nerviosos biológicos, como el cerebro, procesa información. El elemento clave de este paradigma es la estructura novedosa del sistema de procesamiento de información. Se compone de una gran cantidad de productos altamente interconectados con elementos de procesamiento (neuronas) que trabajan al unísono para resolver problemas específicos o con el fin de ayudarse mutuamente para obtener la mejor respuesta posible al estímulo de entrada. Un ANN está configurado para una aplicación específica, como reconocimiento de patrones o clasificación de datos, a través de un proceso de aprendizaje. Con cada presentación, la señal a la salida de cada neurona dependerá de tres funciones: una función de propagación, que multiplica cada una de las entradas por un peso específico, una función de activación que decide si aplicar o no la función de propagación, y una función de transferencia, que se encarga de acotar la señal que transmite cada neurona y de facilitar la comunicación entre ellas (Figura 4) (Leoca, 2017).

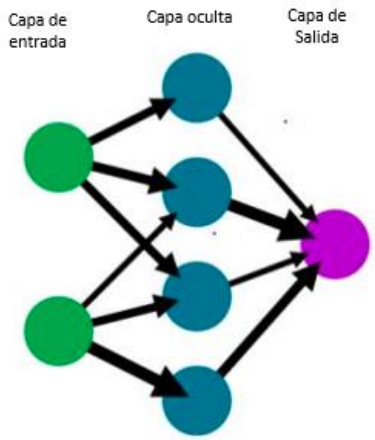

Figura 4: Esquema básico de una Red Neuronal (Leoca, 2017).

Red Neuronal Recurrente (RNN)

Es un tipo de red neuronal artificial especializada en procesar datos secuenciales o series temporales cuya arquitectura permite que la red obtenga memoria artificial contribuyendo a realizar predicciones de lo que sucederá en un futuro a partir de datos históricos (Cañadas, 2021).

Una neurona recurrente transmite la información hacia adelante, pero también tiene la característica de enviar la información hacia atrás (Figura 5). Por lo tanto, en cada paso, la neurona recurrente recibe datos de las neuronas anteriores, pero también recibe información de ella misma en el paso anterior, la RNN se compone de tres capas y la capa oculta (Hidden layer) está conectada tanto en dirección a la capa de entrada como a la de salida. (Digital Guide Ionos, 2020).

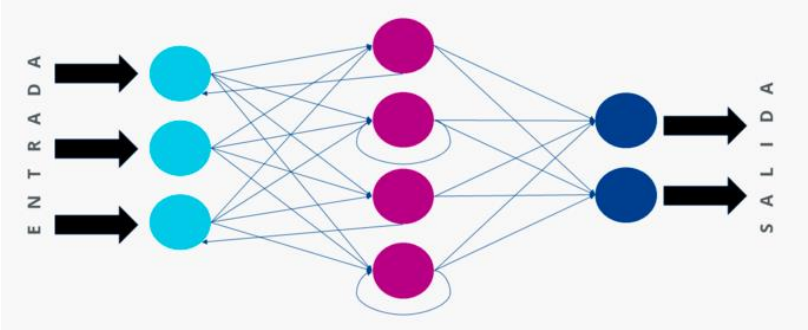

Figura 5: Red Neuronal Recurrente (Digital Guide Ionos, 2020).

Regresión Logística Multinomial

Consiste en generalizar el método de regresión logística para problemas multiclase, es decir, con más de dos posibles resultados discretos. Cabe anotar, que se trata de un modelo que se utiliza, para predecir las probabilidades de los diferentes resultados posibles de una distribución

categórica como variable dependiente, dado un conjunto de variables independientes (que pueden ser de valor real, valor binario, categórico, etc.).

La regresión logística multinomial se conoce por una variedad de otros nombres, incluyendo regresión multiclase LR, la regresión multinomial, función SoftMax regression, Logit multinomial, clasificador de máxima entropía (MaxEnt), etc. (Pando Fernández & San Martín Fernández, 2004).

Algunas consideraciones sobre los datos que se deben de tener para aplicar estos tipos de modelos son:

- Datos: La variable dependiente debe ser categórica. Las variables independientes pueden ser factores o covariables. En general, los factores deben ser variables categóricas y las covariables deben ser variables continuas.
- Supuestos: Se asume que la razón de las ventajas de cualquier par de categorías es independiente de las demás categorías de respuesta. De igual manera, dado un patrón en las covariables, se asume que las respuestas son variables multinomiales independientes. (Statistics, 2021).

El modelo de un regresor logístico multinomial nos permite obtener la probabilidad p_i de tener la clase i (Ecuación 2):

Ecuación 2: Fórmula Regresión Logística Multinomial (López, 2017) (Wikipedia, s.f.).

$$\mathrm{Pi} = \frac{e^{Z_i}}{\sum_{j=1}^{k} e^{Z_j}} \text{ con } Z_i = \alpha_{i,1}x_1 + \alpha_{i,2}x_2 + \cdots + \alpha_{i,n}x_n + \beta_i$$

Asumiendo k clases, un vector x de características, los coeficientes $\alpha_{i,1}, \ldots, \alpha_{i,n}$ del modelo estimado y el sesgo o error β (Pando Fernández & San Martín Fernández, 2004).

Árboles Aleatorios

Consiste en un algoritmo utilizado para resolver problemas de clasificación y regresión. Tendiendo a combinar cientos de árboles de decisión que luego son entrenados cada uno de los árboles en una muestra diferente de las observaciones.

Las predicciones finales de los árboles aleatorios o bosque aleatorio se realizan promediando las predicciones de cada árbol individual que tiende a sobreajustarse (overfitting) a los datos de entrenamiento, pero el bosque aleatorio puede mitigar ese problema al promediar los resultados de predicción de diferentes árboles. Esto le da al algoritmo de bosques aleatorios una mayor precisión predictiva más que un solo árbol de decisión. (Cardellino, 2021).

El algoritmo funciona completando los siguientes pasos (Figura 6):

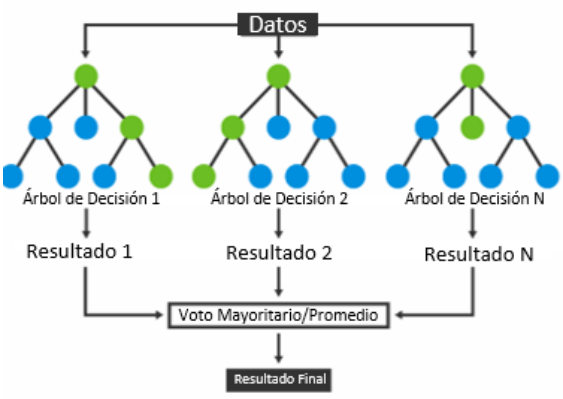

Figura 6: Bosques Aleatorios (TIBCO, s.f.).

- Paso 1: El algoritmo selecciona muestras en forma aleatoria de la base de datos proporcionada.
- Paso 2: El algoritmo creará un árbol de decisión para cada muestra seleccionada. Luego obtendrá un resultado de predicción de cada árbol creado.
- Paso 3: A continuación, se realizará la votación para cada resultado previsto. Para un problema de clasificación, usará la moda, y para un problema de regresión, usará la media.
- Paso 4: Y finalmente, el algoritmo seleccionará el resultado de predicción más votado como predicción final. (Cardellino, 2021).

Potenciación del Gradiente o Aumento de Gradiente

Potenciación del gradiente (Gradient Boosting), es una técnica de aprendizaje automático utilizado para el análisis de la regresión y para problemas de clasificación estadística, el cual produce un modelo predictivo en forma de un conjunto de modelos de predicción débiles (son ligeramente mejor que el azar), típicamente árboles de decisión. Construye el modelo de forma

escalonada como lo hacen otros métodos de boosting que es combinar los resultados de varios clasificadores débiles para obtener un clasificador robusto, y los generaliza, permitiendo la optimización arbitraria de una función de pérdida diferenciable. Los algoritmos de potenciación optimizan una función de coste sobre el espacio de funciones, eligiendo iterativamente una función (hipótesis débil) que apunta en la dirección del gradiente negativo. Esta visión de gradiente funcional de potenciación ha llevado al desarrollo de algoritmos de potenciación en muchas áreas del aprendizaje automático y estadísticas más allá de la regresión y la clasificación (Fafalios, Charonyktakis, & Tsamardinos, 2020).

El funcionamiento del Aumento de Gradiente está compuesto por tres elementos (Figura 7):

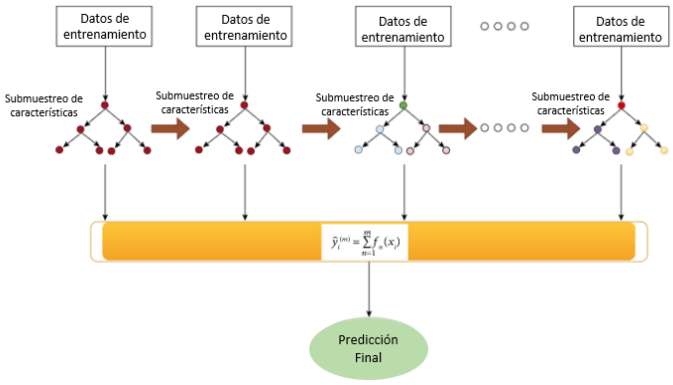

Figura 7: Potenciación del Gradiente (Muhammad Shahani & Muhammad, 2021).

1. Una función de pérdida a optimizar: Depende del tipo de problema que se está resolviendo, y debe ser diferenciable (Brownlee, 2016).

2. Un aprendiz débil para hacer predicciones: Los árboles de decisión se utilizan como el alumno débil en el aumento de gradiente, Específicamente, se utilizan árboles de regresión que arrojan valores reales para divisiones y cuya salida se puede sumar, lo que permite agregar salidas de modelos posteriores y "corregir" los residuos en las predicciones. Son elegidos los mejores puntos de división en función de las puntuaciones de pureza como Gini o para minimizar la perdida. Los árboles de decisión elegidos son muy cortos al ser de una única división, siendo restringidos de formas específicas ya sea por el número máximo de capas, nodos y hojas (Brownlee, 2016).

3. Un modelo aditivo para agregar estudiantes débiles para minimizar la función de pérdida: Los árboles se agregan de uno en uno y los árboles ya existentes en el modelo no se modifican. Es utilizado un procedimiento de descenso de gradiente para minimizar la pérdida al agregar árboles. Debido a que el descenso gradiente lo que hace es minimizar el conjunto de submodelos de aprendizaje débil (árboles de decisión), después de ser calculada la perdida, se debe de agregar un árbol al modelo que reduzca la perdida (para que continúe el gradiente), por parametrización del árbol son modificados los parámetros y se mueve en la dirección correcta (reduciendo la perdida residual), Se van agregando una cantidad fija de árboles o el entrenamiento se detiene una vez que la pérdida alcanza un nivel aceptable o ya no hay una mejora sobre un conjunto de datos de validación externa (Brownlee, 2016).

Algoritmo de seguimiento de partículas

Consiste en una implementación completa del algoritmo de Crocker-Grier para localizar características redondas o similares a manchas en las imágenes sobre un fondo negro (Crocker & Grier, 1995). Los principales pasos son:

Detección de partículas: Se realiza una detección de partículas en cada cuadro de la secuencia de imágenes o vídeo. Esto implica identificar las áreas de interés que corresponden a las partículas y diferenciarlas del fondo. Se pueden utilizar métodos de segmentación o detección de bordes para este propósito.

Atribución de identidad: Una vez que se han detectado las partículas en el cuadro actual, se asigna una identidad única a cada partícula. Esto permite realizar un seguimiento individual de las partículas a lo largo del tiempo. Para esto, se pueden utilizar diferentes criterios, entre alguno de ellos está la distancia más cercana, para vincular las partículas detectadas en el cuadro actual con las partículas detectadas en el cuadro anterior.

Predicción de ubicación: Después de asignar las identidades de las partículas en el cuadro actual, se realiza una predicción de la ubicación de las partículas en el siguiente cuadro. Esto se hace utilizando modelos de movimiento o trayectorias previas de las partículas. La predicción se basa en la velocidad y la dirección estimada de las partículas.

Actualización y corrección: Una vez que se ha realizado la predicción de ubicación, se compara con las detecciones reales en el siguiente cuadro. Se ajustan las predicciones iniciales según las nuevas detecciones y se corrigen los errores de seguimiento. Esto se logra utilizando métodos de filtrado, como el filtro de Kalman, que permite fusionar la información de las predicciones y las nuevas detecciones para obtener una estimación más precisa de la ubicación de las partículas.

Análisis de trayectorias: Finalmente, se analizan las trayectorias de las partículas a lo largo del tiempo para extraer información cuantitativa, como la velocidad, la aceleración, el comportamiento de difusión, etc. Esto proporciona una comprensión más profunda del movimiento de las partículas y puede revelar patrones o características de interés en el sistema estudiado.

El algoritmo de Crocker-Grier ha demostrado ser efectivo en realizar un seguimiento preciso de partículas individuales a lo largo del tiempo, permitiendo avances importantes en la comprensión de diversos fenómenos físicos y biológicos (Neves Miranda, 2019).

Métricas de rendimiento del modelo

Son las estadísticas resumidas de desempeño del modelo como la precisión, sensibilidad, la especificidad, la precisión equilibrada de las categorías individuales y el área bajo la curva (ROC), las características de cada una de las métricas son:

- La precisión es la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) entre el número total de casos examinados (Ecuación 3).

Ecuación 3: Fórmula del porcentaje precisión (Bagui & Mink, 2022).

$$\%\ \text{Precisión} = \frac{\text{Numero de verdaderos positivos} + \text{Numero de verdaderos negativos}}{\text{Numero total de predicciones}} * 100\%$$

- La sensibilidad es la proporción de muestras positivas verdaderas que el modelo clasifica correctamente (Ecuación 4).

Ecuación 4: Fórmula del porcentaje sensibilidad (Bagui & Mink, 2022).

$$\% \text{ Sensibilidad} = \frac{\text{Numero de verdaderos positivos}}{\text{Numero de verdaderos postivos} + \text{Numero de falsos negativos}} * 100\%$$

- La especificidad, también conocida como tasa de verdaderos negativos, es la proporción de muestras de verdaderos negativos que se clasifican como negativas (Ecuación 5).

Ecuación 5: Fórmula del porcentaje especificidad (Bagui & Mink, 2022).

$$\% \text{ Especificidad} = \frac{\text{Numero de verdaderos negativos}}{\text{Numero de verdaderos postivos} + \text{Numero de falsos negativos}}$$

- La métrica de precisión equilibrada proporciona una medida de la precisión de cada clase individual teniendo en cuenta los verdaderos positivos, los falsos positivos y los falsos negativos. Esta es una buena medida para evaluar conjuntos de datos desequilibrados (Ecuación 6). (García, Mollineda, & Sánchez, 2009).

Ecuación 6: Fórmula del porcentaje precisión equilibrada (García, Mollineda, & Sánchez, 2009).

$$\text{Precisión Equilibrada} = \frac{\text{sensibilidad} + \text{Especificidad}}{2}$$

- El área bajo la curva es una técnica para visualizar, organizar y seleccionar clasificadores en función de su rendimiento. El gráfico representa las compensaciones relativas entre los positivos verdaderos y falsos. Son gráficos bidimensionales sobre los cuales se trazan los clasificadores, siendo el eje “y” la tasa de verdaderos positivos y el eje “x” la tasa de falsos positivos (Bagui & Mink, 2022), una mejor manera de detectar el rendimiento o efectividad del

clasificador puede ser con solo mirar el área bajo la curva como lo presenta la Figura 8 (Di Sipio, 2021).

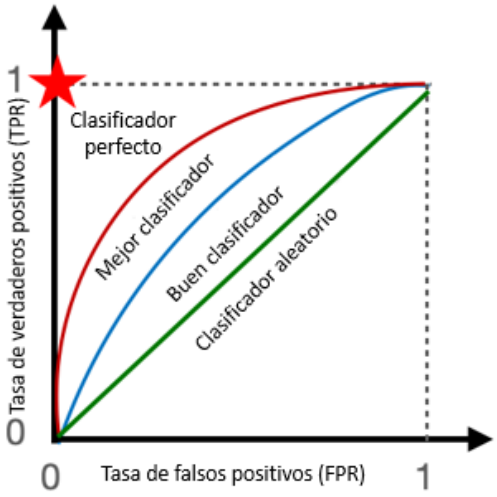

Figura 8: Área bajo la curva (Di Sipio, 2021).

3. Revisión sistemática de la literatura

3.1 Fuentes de información

Realizando una revisión sistemática de la literatura para comprender los diferentes usos en la detección y caracterización de emulsiones, se identifican varios contenidos de los cuales son extraídos información que indica el estado actual de la inteligencia artificial en las emulsiones:

Tabla 1: Palabras claves de búsqueda.

Grupo 1	Emulsiones agua en aceite, aceite en agua, contenido %BSW
Grupo 2	Inteligencia artificial, aprendizaje automático, aprendizaje profundo, visión por computador
Grupo 3	Dispersión de luz dinámica (DLS), distribución del tamaño de gotas, evaluación microscópica

Los artículos que cumplieron con este criterio de inclusión se caracterizaron por ser fundamentales para la construcción del estado del arte y la solución, correspondiendo al grupo 1 que incluía los términos asociados con el contexto de emulsiones, el grupo 2 con términos generales

de los tipos de métodos abarcados para el estudio de las emulsiones y en el grupo 3 las diferentes técnicas aplicadas para entender el comportamiento y características de las emulsiones.

Dentro de la búsqueda en Internet utilizando algunos de los motores de búsqueda más representativos, en cuanto artículos científicos e información académica que permitiera resolver cada uno de los objetivos mencionados. Se emplearon las bibliotecas digitales de la tabla 2 dados los diferentes contenidos en el ámbito de interés del presente trabajo.

Tabla 2: Recursos digitales de material bibliográfico.

Librería	**Tipo**	**URL**
Scopus	Biblioteca Digital	www.scopus.com
Sciencedirect	Biblioteca Digital	www.sciencedirect.com

La distribución de documentos o artículos vinculados al tema de interés indican un creciente aumento en el estudio de emulsiones de agua en aceite y con casos de solución aplicando técnicas de aprendizaje de máquinas.

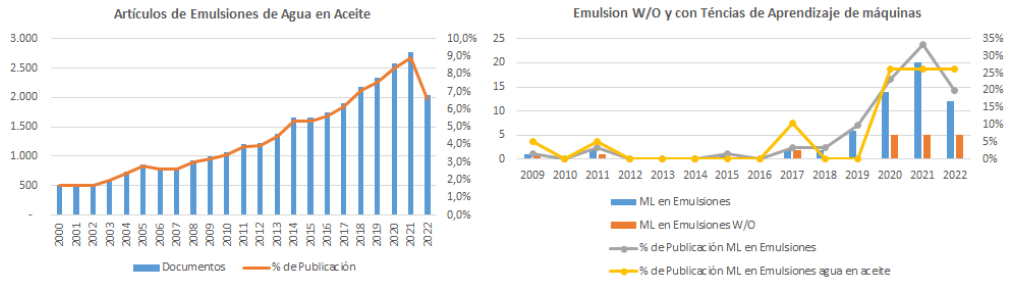

Figura 9: Evolución en investigaciones para emulsiones y W/O con aprendizaje de máquina (ML).

Indicando como durante los últimos cuatro años se han realizado más del 30% de las investigaciones desde los 20 años que se han estudiado los artículos con respecto a las emulsiones de agua en aceite, para el caso aplicado del análisis de emulsiones mediante técnicas de aprendizaje de máquinas (ML), el 87% de las investigaciones han sido realizadas durante ese mismo periodo (Figura 9).

Estos artículos se han encontrado en diferentes campos de la investigación, teniendo la mayor partición en el desarrollo del tema áreas como la química, la ingeniería y la ciencia de los materiales (Figura 10).

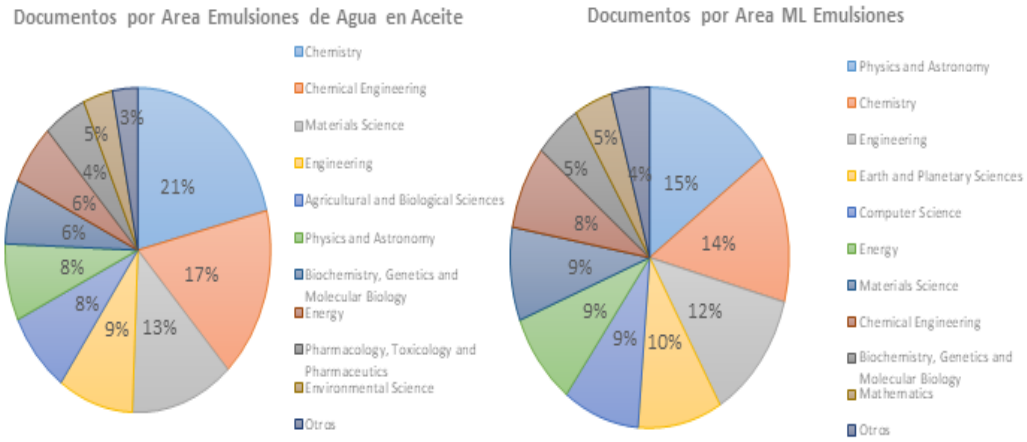

Figura 10: Tipos de áreas en investigaciones para emulsiones y W/O con aprendizaje de máquina (ML).

3.2 Trabajos relacionados

El proceso de revisión de la literatura ha supuesto una parte esencial de todo el desarrollo del proyecto, por el lado de los conocimientos teóricos y técnicos adquiridos y por la inspiración surgida con base en dicha revisión, que ha resultado muy útil para desarrollar ideas propias que permitan cubrir los objetivos planteados.

De forma esquemática, y con fines de organizar a alto nivel en periodos la literatura publicada en torno al tema de la caracterización y predicción del contenido de agua en emulsiones (%BSW), se han identificado varias formas de resolución para este tipo de investigación.

El primero de ellos está concentrado en los procesos de demulsificación química y mecánica llevando a cabo centrifugaciones para examinar la eliminación de agua para identificar la eficiencia y estabilidad de las emulsiones de agua en aceite, mediante un análisis DSC

(calorimetría diferencial de barrido) se determina el contenido de agua y la distribución del tamaño de las gotas en la emulsión, donde una alta viscosidad y la poca fluidez conducen a ser indicadores de un bajo contenido de agua y pequeñas gotas de agua en la emulsión dado al contacto limitado, al igual que cuando se tienen diferentes escenarios de temperatura a mayor nivel puede disminuir el agua contenida de pequeñas gotas y la estabilidad de las emulsiones formadas (Yudo Wardhono, Permana Pinem, Wahyudi, & Agustina, 2019).

Otro de los métodos que se aplica corresponde es la norma ASTM D96-88 (Pájaro & Álvarez, 2014), el cual es utilizado en la industria para tener un control sobre el contenido de agua presente en el crudo, siendo necesario medir su cantidad inicial en la muestra de estudio. Este método es estándar y consiste en determinar la cantidad de agua y sedimentos en el crudo por centrifugado. Se mezcla el petróleo (liviano o pesado) y se le es aplicado tolueno en cantidades volumétricas iguales y se llenan tubos de centrifugación de 100 mL. Se realiza un proceso de centrifugado para la muestra por 10 minutos y posteriormente se determina el volumen de agua y sedimentos en la parte inferior de los tubos, siendo llamado como la prueba de la botella (Pájaro & Álvarez, 2014).

Otro de los métodos es el térmico, que mejora el rompimiento o separación, reduce la viscosidad del aceite e incrementa la tasa de sedimentación del agua.

El aumento de temperaturas también da como resultado la desestabilización de las películas rígidas debido a la reducción de la viscosidad interfacial. El calor acelera el proceso de rompimiento de la emulsión. Incrementar la temperatura tiene algún efecto negativo por los costos químicos y de operación (Kokal, 2005)

Dentro de las formas mediante las cuales son caracterizadas las emulsiones de W/O es mediante la microscopia óptica es posible obtener una inspección visual de las emulsiones a nivel microscópico, capturando micrografías en las que se evidencia el tamaño de la gota y conteo (Pájaro & Álvarez, 2014). Siendo implementados algoritmos de segmentación de imágenes basada en histogramas mediante un software científico de procesamiento de imágenes llamado Fiji, versión 1.51 h (Unnikrishnan, Donovan, Macpherson, & Tormey, 2020). Fiji es una versión extendida de ImageJ (Schindelin, 2012). ImageJ, anteriormente conocido como NIH (National Institutes of Health) siendo un paquete de software científico de procesamiento de imágenes de código abierto, que tiene un historial comprobado de análisis de imágenes en diferentes áreas.

La técnica de espectroscopia de infrarrojo cercano ha sido utilizada ampliamente en la industria de los hidrocarburos con el fin de evaluar y caracterizar los productos del petróleo y para su procesamiento y refinación, teniendo como propósito analítico la estimación del contenido de agua y del tamaño de gota promedio en la emulsión, a partir de las variaciones de los enlaces químicos de las moléculas en la región de infrarrojo cercano (Pájaro & Álvarez, 2014) (Ametek Spectro Scientific, s.f.).

El análisis de las características principales del sistema emulsionado se encuentra como los diferentes usuarios abordan la distribución del tamaño de la gota, el contenido de la fase dispersa, la viscosidad, naturaleza química de las fases (Noboa, Márquez, & López, 2017), la temperatura, la presión y el tiempo de envejecimiento de la emulsión, tipo de crudo, considerando importante el efecto de algunas de estas variables en el proceso de demulsificación (Tolosa, 2016).

Dentro de los estudios actuales investigados se encuentra la aplicación del análisis de imágenes microscópicas combinado con algoritmos de aprendizaje automático y aprendizaje profundo en el proceso de la clasificación de muestras de emulsión. En varios de los casos la clasificación se basa en la variación de las características de las gotas observadas durante el proceso de emulsificación siendo considerada la variable tiempo. El análisis de visión por computadora, integrado con el aprendizaje automático, ha ganado gran atención recientemente en una variedad de industrias para lograr una evaluación rápida, precisa y objetiva de la calidad del objeto de estudio.

La técnica aplicada está basada en histograma (HBT) funciona calculando el histograma de los valores de gris en la imagen para observar los picos y valles. A esto le sigue el filtrado del ROI (Región de interés) utilizando el umbral de intensidad correspondiente (Unnikrishnan, Donovan, Macpherson, & Tormey, 2020). Los pasos describen el umbral y el filtrado del ROI con un ruido mínimo son:

1. Es calculado el histograma a partir de los valores de intensidad de píxel de cada imagen.
2. Luego, cada imagen se umbralizó utilizando el valor de intensidad media para resaltar el ROI.
3. Las imágenes con umbral se convirtieron a binarias (gotas en negro/fondo en blanco).
4. Se aplicó la segmentación de cuencas hidrográficas para separar las gotas superpuestas y obtener la imagen de salida.

5. Fueron analizadas las gotas detectadas con un tamaño de rango $\geq 1\ \mu m^2$ y un rango de circularidad de 0,00 a 1,00.

6. Para cada gota se obtuvieron trece características, que incluían tamaño, forma, centroide y orientación (Unnikrishnan, Donovan, Macpherson, & Tormey, 2020).

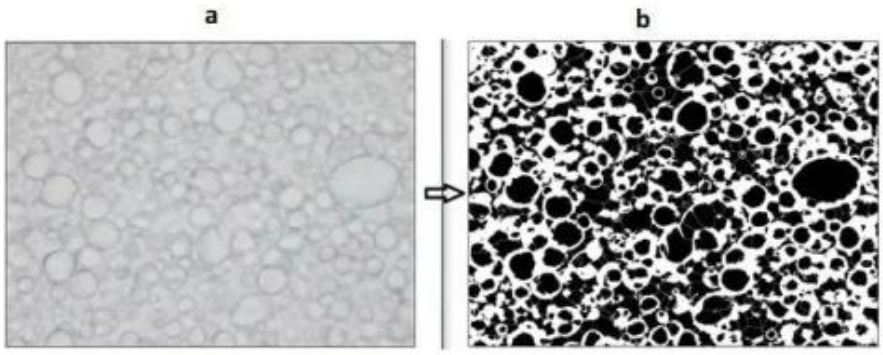

Figura 11: Detección de gotas. (a) Micrografía obtenida de una muestra de emulsión procesada durante 5 min. (b) Imagen de salida con las gotas detectadas usando el HBT. (Unnikrishnan, Donovan, Macpherson, & Tormey, 2020).

Los resultados en la (Figura 11 b) de una micrografía, es con la cual se inician los diferentes análisis para cada una de las muestras seleccionadas, extrayendo características como el tamaño, la forma, la orientación, las coordenadas de cada gota y el recuento, lo que incluyen el área de las gotas, el perímetro, el diámetro máximo de Feret (Feret), el diámetro mínimo de Feret (minFeret) y el recuento de gotas. Feret corresponde a la distancia máxima y minFeret corresponde a la distancia mínima entre los límites de la región de interés.

Para este caso, con un consolidado de 150 micrografías y generando cuatro categorías denominadas TAMU(Objetivo, Aceptable, Marginal e Inaceptable) basadas en la calidad de la

emulsión, correspondiendo a la variable respuesta, se emplearon mediante cinco características importantes de las gotas, como el área media de las gotas, el perímetro, tanto el diámetro mínimo y máximo de Feret el recuento de gotas y usó de los valores de intensidad de píxeles de la imagen como predictores se desarrollan modelos de clasificación supervisados utilizando algoritmos de aprendizaje automático y aprendizaje profundo.

Teniendo presente el autor para predecir la variable de clasificación los siguientes modelos:

- Regresión logística multinomial (MLR).
- Bosque aleatorio (RF).
- Red neuronal vainilla (VNN).
- Red neuronal convolucional (CNN).

Teniendo las estadísticas resumidas de desempeño del modelo como lo es la sensibilidad, la especificidad y la precisión equilibrada de las categorías individuales.

El modelo CNN se usó mediante un enfoque basado en píxeles, debido a la cantidad de datos limitados para ser utilizados en este método a comparación de los otros modelos de clasificación. Así que una de las soluciones empleada a cada una de las 150 micrografías obtenidas del lote inicial se cortó en cuatro cuartos y esto dio como resultado un total de 600 micrografías, teniendo aproximadamente el 80 % de las micrografías de cada una de las categorías de la variable objetivo seleccionada como conjunto de entrenamiento y el 20 % restante se retuvo para la

prueba. Los valores de intensidad de píxel de estas micrografías formaron las características de entrada del modelo CNN. Los pasos involucrados en el enfoque CNN son:

1. Las micrografías se transformaron a una resolución cuadrada de n × n píxeles, donde n = 500.
2. Luego se convirtieron a escala de grises.
3. Los valores de píxel se extrajeron de cada micrografía y se almacenaron en una matriz 2D de tamaño n × n.
4. El paso 3 se repitió para las 456 micrografías y, finalmente, se desarrolló una matriz de datos 3D de tamaño 456 × n × n.
5. Los datos de respuesta (variable independiente) se codificaron como 0 y 1 de acuerdo a la categoría.
6. Se entrena el modelo utilizando las 456 micrografías.
7. Es evaluado el error de entrenamiento y la precisión.

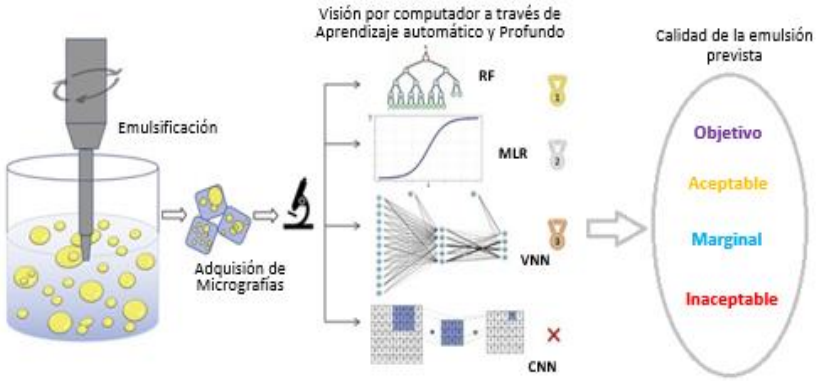

Figura 12: Representación esquemática del proceso de clasificación TAMU de micrografías (Unnikrishnan, Donovan, Macpherson, & Tormey, 2020).

Correspondiendo a una arquitectura (Figura 12) dada por una fase de adquisición de imágenes mediante microscopio óptico, técnicas de visión por computador y la aplicación de aprendizaje automático y profundo, clasificando la variable respuesta en 4 categorías asignadas por expertos según la caracterización previa.

De esta manera, concluye la revisión de la literatura referente al tema de caracterización y predicción de emulsiones, haciendo un especial énfasis en el papel del aprendizaje de máquinas supervisado. Se ha recorrido de forma selectiva toda la serie de estudios e investigaciones publicadas sobre el tema, empezando por los estándares más extendidos actualmente, pasando por soluciones más ingeniosas y complejas, recalcando algunos conceptos teóricos importantes y señalando las virtudes de cada una de las técnicas y formas de enfrentar este problema.

4. Método para la caracterización y predicción de emulsiones de agua en crudo empleando técnicas de aprendizaje de máquina

Los siguientes ítems a ser abordados constan de la descripción de la metodología de la solución, dando inicio desde la toma de muestras en laboratorio (extracción mínima de la muestra del crudo a ser analizada), la adquisición de las imágenes, la selección de las características y el entendimiento del comportamiento de las emulsiones mediante la modelación estadística y de aprendizaje de máquinas, validando los modelos implementados mediante las imágenes identificadas en la literatura.

Para el desarrollo del presente trabajo de acuerdo a la consecución de pasos identificados en los trabajos previos se propone la siguiente arquitectura (Figura 13):

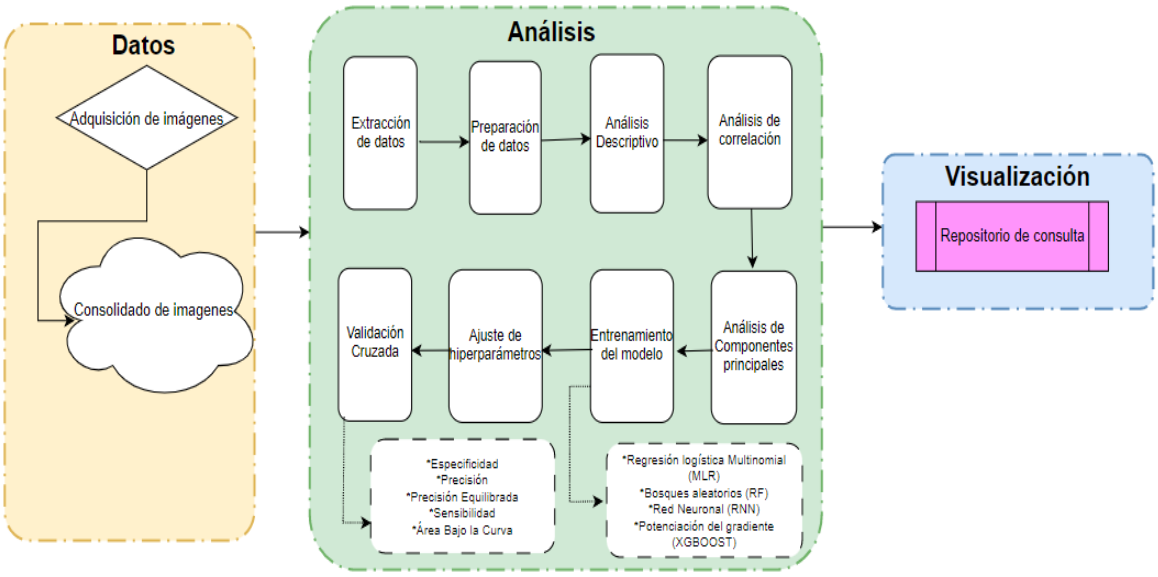

Figura 13: Arquitectura de la solución (elaboración propia).

Iniciando con una fase de adquisición de los datos mediante microscopia óptica, consolidando todas las imágenes en un solo repositorio en la nube debido a que todo el proceso es realizado en el aplicativo de “Google Colaboratory” (Google, 2017) el cual es un software de código libre en un servidor propio de Google, espacio en el que se desarrolla toda la fase de análisis que es la extracción de los datos (extracción de características de las imágenes), transformaciones, análisis descriptivo para entender tendencias y patrones, análisis de correlación para identificar relaciones crecientes o decrecientes, componentes principales para tener una mayor explicación y contribución con el menor número de variables, fase de modelación con varios métodos de aprendizaje de máquinas, encontrando los parámetros más óptimos mediante el ajuste de hiperparámetros, realizando una validación para cada uno de los modelos mediante las métricas de desempeño con el fin de encontrar el modelo más acertado y ser validado con las métricas de desempeño las imágenes de la literatura, para finalmente presentar una interfaz gráfica (Figura 13).

4.1 Materiales

Se utiliza el crudo pesado proveniente de diferentes muestras que fueron adquiridas en el laboratorio por parte del personal experto del grupo de investigación Michael Polanyi (MP) (Michael Polanyi, s.f.), de acuerdo a los lineamientos de protocolo ya definidos, se hace uso del microscopio óptico bajo los estándares de un aumento de 40X para cada muestra y con una

configuración de iluminación estándar a 7 V, con un proceso de Norma ASTM D96-88 (Figura 14).

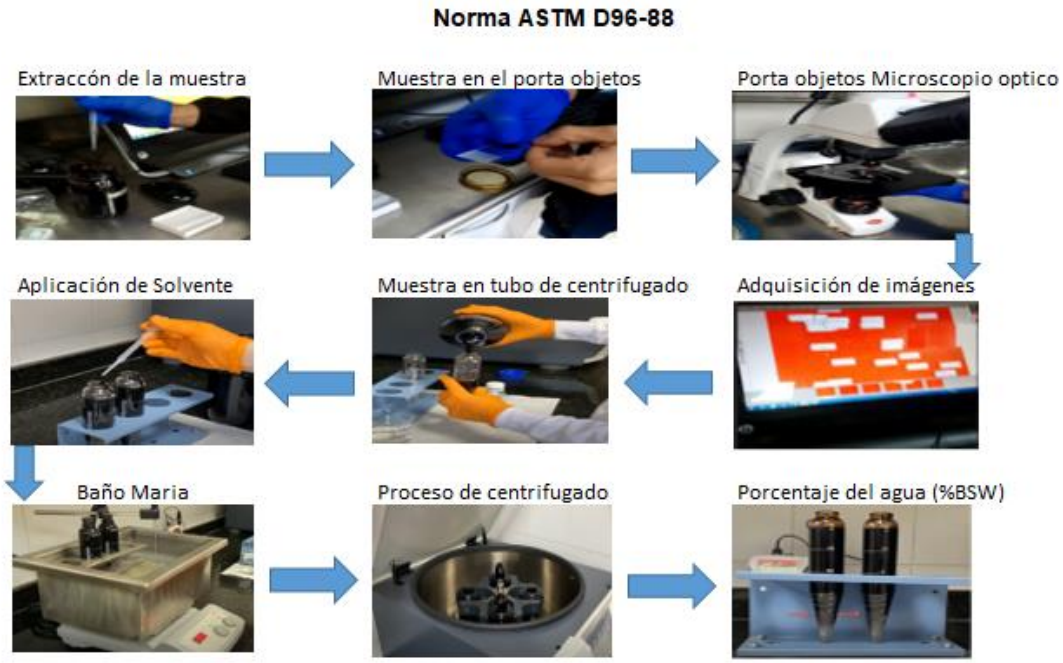

Figura 14: Proceso de adquisición y análisis según Norma ASTM D96-88 (Michael Polanyi, s.f.).

Se inicia con un método estandarizado de preparación de portaobjetos de muestra para lograr consistencia junto con todo el proceso de demulsificación, temperatura, centrifugado y medición del %BSW. Donde a cada uno de los crudos se le realizaron dos registros fotográficos, correspondiendo al momento tanto de ser recibidos y un registro final luego de un proceso de deshidratación (Aplicando el solvente Tolueno) a 40º C/15 horas y un proceso de centrifugado a 3000 RPM/3 horas, adquiriendo tanto un %BSW inicial y un %BSW final, obteniendo un total de 12 tipo de crudos para un total de 24 registros fotográficos (Figura 15).

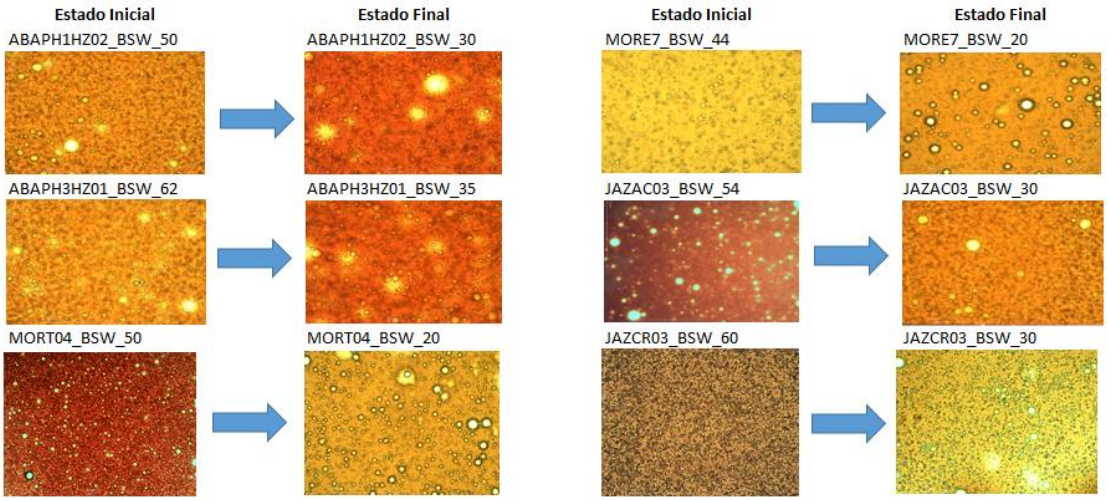

Figura 15: Adquisición de imágenes extraídas en una fase inicial y final con el porcentaje del %BSW (Michael Polanyi, s.f.).

4.2 Identificación de variables y parámetros

Haciendo uso de toda la información suministrada, se da inicio realizando los respectivos análisis de imagen buscando capturar las diferentes características encontradas en cada uno de los registros de acuerdo a las diferentes variables que son empleadas en el estudio de emulsiones, en ellas podemos identificar su distribución y comportamiento sobre las muestras seleccionadas.

4.2.1 Algoritmo de seguimiento de partículas (Trackpy)

La metodología planteada por parte del método de seguimiento de partículas (Figura 16) está dividido por tres pasos hasta un paso final que es la salida (Output) la información de detallada de las características de las imágenes, siendo implementado mediante la librería del software Python bajo la siguiente lógica:

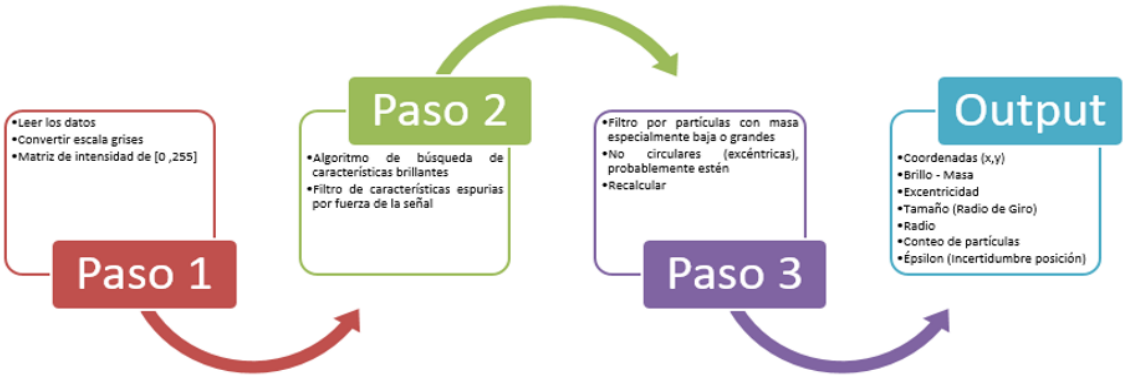

Figura 16: Método de Trackpy, mediante el algoritmo de seguimiento de partículas (Trackpy Contributors, Trackpy Documentation, 2015).

Iniciando con la lectura de las imágenes adquiridas, siendo escaladas a una escala de grises y matricial, continuando con un método de búsqueda de características brillantes y omitiendo aquellas características que generan ruido y a su vez masas que son demasiado grandes o demasiado pequeñas y no cuentan con algún tipo de forma circular definida. Por medio del método se obtienen atributos como lo es la posición, la masa (brillo total), tamaño (radio de giro del brillo), excentricidad, enumera la señal, una medida relacionada con el contraste, épsilon, la incertidumbre estimada en la posición de la característica, la cantidad de partículas (Allan, Caswell, Keim, Van der Wel, & Verweij, 2021) retornando por parte de la librería de Python, datos tabulares con 8 atributos y 4.768 registros extraídos de los 24 registros de micrografías, estas variables y cada uno de los registros son propias de la caracterización de cada una de las partículas detectadas para cada muestra de las emulsiones adquiridas en el laboratorio vinculando información que ya es propia de

la imagen como lo es nombre de la muestra de la emulsión, el porcentaje del agua calculado (%BSW) (Tabla 3)

Tabla 3: Conjunto de datos tabulados para cada una de las partículas de las muestras de las emulsiones mediante la librería Trackpy

nombre_emulsion	Coordenada y	Coordenada x	Masa (brillo Total)	Tamaño	Excentricidad	Señal	Épsilon	Cantidad de particulas	%BSW
ABAPH3HZ01_BSW_62,png	6,372898121	16,93867458	350,7372581	2,128651514	0,285777172	32,61058582	0,27951628	129	30
ABAPH3HZ01_BSW_62,png	161,4261364	103,9744318	488,4649451	2,097075886	0,330594921	33,99827032	0,209994615	34	19
ABAPH1HZ02_BSW_50,png	159,4641537	153,0036955	469,3842831	2,385593618	0,331016032	24,2844788	0,17143306	130	30
ABAPH3HZ06_BSW_48,png	159,7427766	14,5017618	492,2810774	2,186188951	0,394408472	32,2636647	0,277259136	97	29
MORAG05_BSW_84,png	159,2137509	352,0061266	509,6271337	2,285449878	0,42401543	30,87598019	0,145176783	85	50
MORAH204_BSW_50,png	158,7870631	186,6568903	863,486682	1,979403792	0,254455684	61,75196038	0,151884299	24	16
MORAO01_BSW_26,png	158,069477	249,1272443	888,8119242	1,510824849	0,214201091	88,46488707	0,164235068	85	16
MORT04_BSW_50,png	157,4638868	299,1653016	465,9150719	2,05018932	0,305245413	33,99827032	0,204844025	118	21
				X_i					Y_i

Siendo consideradas como las variables independientes (X_i) las encargadas de predecir la variable dependiente (Y_i) que corresponde al porcentaje de agua contenido dentro de la emulsión (%BSW). A continuación, se describen cada una de las variables haciendo uso de estadística descriptiva y del método de seguimiento de partículas para comparar la evolución entre el %BSW inicial (al momento de ser recibida la muestra en el laboratorio) y el %BSW final (luego de un proceso de desmulsificación).

Cantidad de Partículas

Corresponde al número de objetos detectados por el método (Trackpy Contributors, Trackpy Documentation, 2015). Identificando como en una fase con un %BSW final se identifica una menor cantidad de partículas (Tabla 4), en el único caso en donde se encuentra un comportamiento de un aumento de cantidad de partículas es la muestra ABAPH1HZ02.

Tabla 4: Cantidad de partículas detectadas por cada una de las muestras de las emulsiones y la diferencia entre el %BSW inicial y final.

Nombre_emulsión	% BSW Inicial	% BSW Final	%Diferencia	% BSW Inicial	% BSW Final	%Diferencia
	Cantidad_inicial	cantidad_final	dif_cantidad			
ABAPH1HZ02	240	246	2,5%	50%	30%	-20,0%
ABAPH3HZ01	323	181	-44,0%	62%	35%	-27,0%
ABAPH3HZ06	207	128	-38,2%	48%	30%	-18,0%
ABAPH7HZ03	309	272	-12,0%	52%	30%	-22,0%
ABAPH8HZ09	154	120	-22,1%	40%	20%	-20,0%
JAZAC03	243	209	-14,0%	54%	30%	-24,0%
JAZCR03	224	198	-11,6%	60%	30%	-30,0%
MORAG05	258	162	-37,2%	84%	40%	-44,0%
MORAH204	164	146	-11,0%	50%	16%	-34,0%
MORAO01	153	137	-10,5%	26%	15%	-11,0%
MORE7	175	153	-12,6%	44%	20%	-24,0%
MORT04	210	156	-25,7%	50%	20%	-30,0%

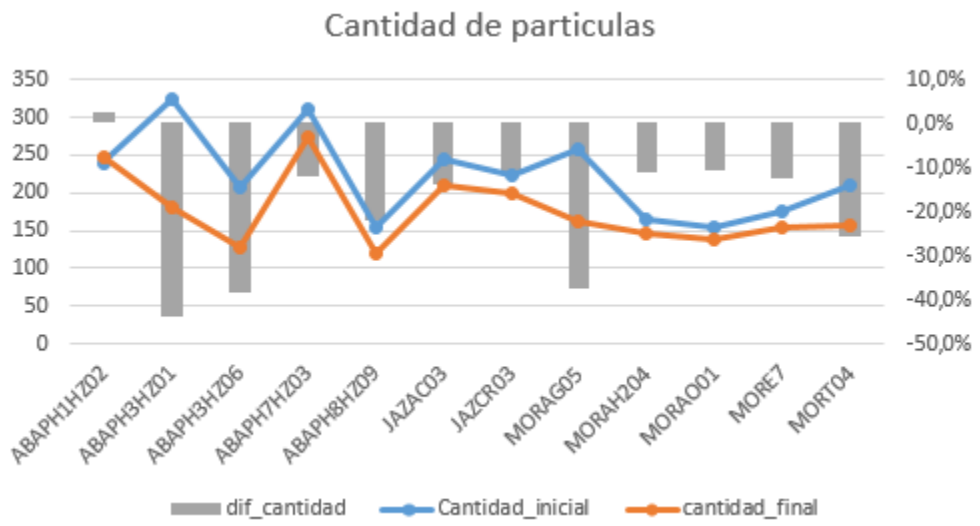

Figura 17: Distribución de la cantidad de partícula para cada muestra con el %BSW inicial y final.

Encontrando un patrón que indica que a mayor porcentaje de diferencia del %BSW inicial vs. final, mayor es la diferencia entre la cantidad de gotas detectadas en una fase inicial con respecto

la fase final luego de un proceso de demulsificación (Figura 17), indicando que a mayor porcentaje de agua extraída menos partículas de gotas van a ser detectadas en una fase final del %BSW.

La masa (brillo total):

Es la intensidad integrada dada por la suma del valor de intensidad de todos los píxeles bajo la máscara (brillo total de la mancha) (Trackpy Contributors, Trackpy Documentation, 2015). Encontrando que en una fase promedio inicial se identifica un mayor brillo con respecto a la fase final, en el único caso en donde se encuentra un comportamiento de aumento en el brillo promedio final es en las muestras MORE7 y ABAPH7HZ03 (Tabla 5).

Tabla 5: Evolución del brillo en una fase inicial y final promedio de cada una de las muestras.

	% BSW Inicial			% BSW Final			%Diferencia		
nombre_emuls_inicial	Promedio_inicial	min_inicial	max_inicial	Promedio_final	min_final	max_final	dif_cantidad_media	dif_cantidad_min	dif_cantidad_max
ABAPH1HZ02	520	115	2.050	291	78	1.242	-44,09%	-31,67%	-39,40%
ABAPH3HZ01	492	142	1.412	388	80	1.622	-21,22%	-43,65%	14,93%
ABAPH3HZ06	765	150	2.454	559	67	2.062	-26,97%	-55,25%	-15,98%
ABAPH7HZ03	494	140	2.066	523	119	2.109	5,76%	-14,90%	2,07%
ABAPH8HZ09	738	84	2.961	580	72	2.466	-21,37%	-15,03%	-16,71%
JAZAC03	602	60	2.511	458	180	1.318	-23,86%	198,88%	-47,50%
JAZCR03	567	123	905	489	225	1.051	-13,77%	83,21%	16,20%
MORAG05	873	174	2.556	582	110	1.210	-33,33%	-36,99%	-52,65%
MORAH204	398	25	2.021	276	104	1.359	-30,68%	322,03%	-32,77%
MORAO01	581	46	2.880	368	59	2.293	-36,72%	28,68%	-20,38%
MORE7	302	65	768	650	58	2.502	115,26%	-12,12%	225,69%
MORT04	864	151	2.805	559	71	2.535	-35,27%	-52,65%	-9,62%

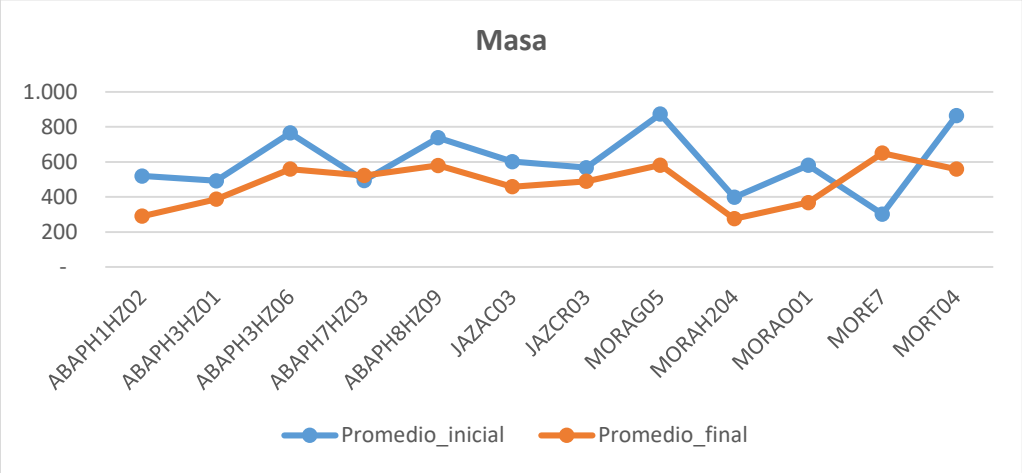

Figura 18: El promedio de la masa de las partículas en una fase inicial y en una fase final %BSW.

Se encuentra una tendencia que indica un mayor porcentaje de brillo promedio en una fase inicial de la emulsión, teniendo un menor brillo en un %BSW final (Figura 18).

Figura 19: Distribución de la masa (Brillo) para cada una las partículas en una fase inicial y en una fase final del %BSW.

Hay una disminución en el brillo entre una fase inicial y una final, siendo desplazando la distribución hacia la izquierda y estando mucho más comprimida en comparación a una fase final que se comporta mucho más “aplanada” (Figura 19).

Excentricidad

Una excentricidad de 0 es un disco perfectamente circular, mientras que una gran excentricidad es una característica muy alargada (Allan, Caswell, Keim, Van der Wel, & Verweij, 2021). Encontrando para el promedio en una fase inicial como final una diferencia tanto positiva como negativa (Tabla 6), notando que cuando se presenta porcentajes positivos se va teniendo

figuras más alargadas, disminuyendo en una menor proporción su forma circular en el %BSW final (Figura 20).

Tabla 6: Evolución de la Excentricidad en una fase inicial y final promedio de cada una de las muestras.

Nombre_emulsión	% BSW Inicial			% BSW Final			%Diferencia		
	Promedio_inicial	min_inicial	max_inicial	Promedio_final	min_final	max_final	dif_cantidad_media	dif_cantidad_min	dif_cantidad_max
JAZCR03	28,4%	1,8%	62,4%	28,3%	3,8%	70,3%	-0,5%	112,6%	12,8%
MORAO01	24,7%	3,9%	77,9%	28,1%	2,5%	66,2%	13,8%	-35,5%	-15,1%
MORAH204	26,7%	0,8%	74,2%	26,8%	1,6%	59,8%	0,3%	92,6%	-19,4%
ABAPH3HZ06	20,1%	1,7%	60,9%	25,6%	2,0%	62,8%	27,1%	12,3%	3,1%
JAZAC03	15,6%	0,5%	58,8%	24,9%	3,3%	58,2%	59,8%	553,7%	-1,0%
ABAPH1HZ02	23,9%	1,3%	56,6%	23,8%	1,0%	64,2%	-0,4%	-21,2%	13,4%
ABAPH7HZ03	22,9%	2,1%	55,4%	23,0%	2,5%	55,4%	0,4%	19,3%	0,1%
ABAPH8HZ09	17,0%	0,9%	67,5%	22,9%	1,6%	59,7%	34,2%	90,7%	-11,6%
MORT04	20,2%	3,1%	65,8%	21,3%	0,2%	68,4%	5,8%	-93,2%	4,0%
MORE7	22,8%	1,8%	64,3%	19,9%	2,3%	54,3%	-12,8%	28,5%	-15,6%
ABAPH3HZ01	23,2%	1,9%	68,1%	19,3%	1,3%	58,8%	-16,8%	-29,0%	-13,7%
MORAG05	20,5%	2,0%	63,3%	18,2%	1,2%	56,1%	-11,2%	-37,6%	-11,4%

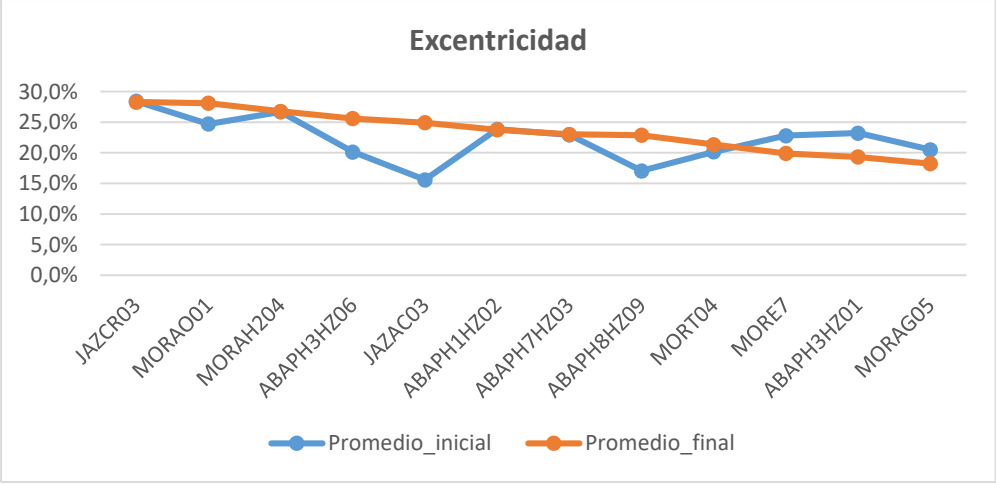

Figura 20: El promedio de la excentricidad de las partículas en una fase inicial y en una fase final promedio del %BSW.

Se identifica un mayor porcentaje de excentricidad promedio en una fase final debido a que la línea se encuentra en el 66% de las muestras sobrepone la línea promedio inicial, sin embargo,

solo en tres casos los valores de excentricidad son mayores por una menor proporción en comparación a una fase final (Figura 20).

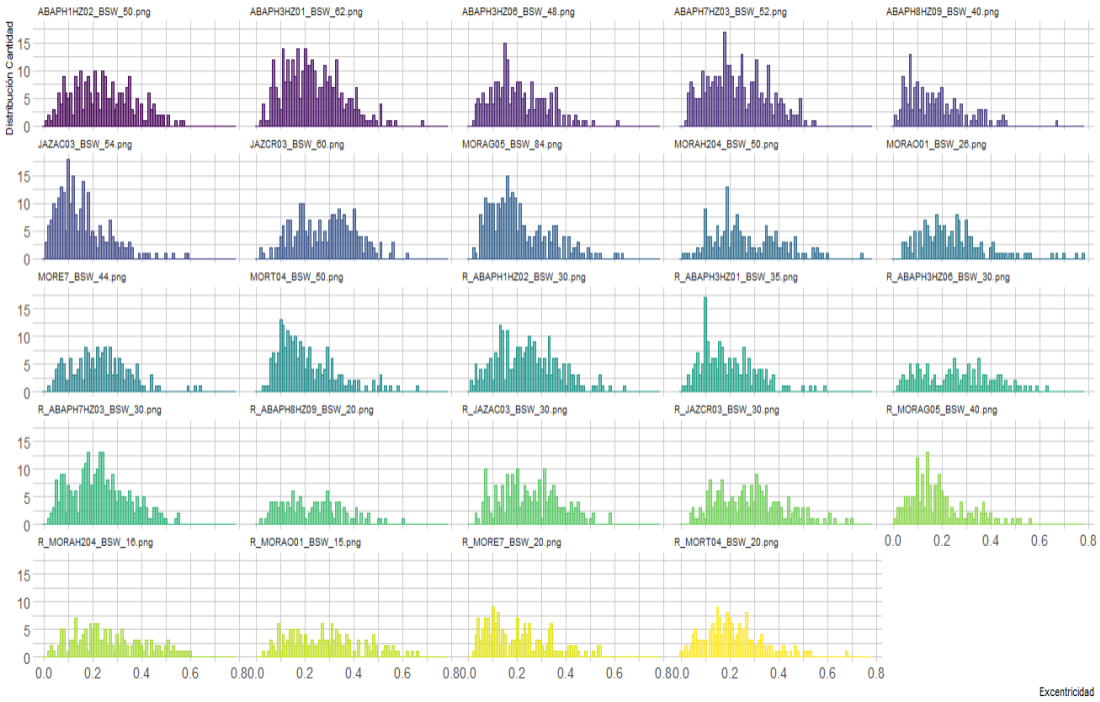

Figura 21: Distribución de la excentricidad para cada una las partículas en una fase inicial y en una fase final del %BSW

En la Figura 21 con respecto a la excentricidad, la distribución de las partículas está mucho más concentrada y desplazada hacia la izquierda del gráfico, en ambos casos gran parte de la población de partículas se concentran por debajo del 0.30.

Tamaño (Radio de giro del brillo)

Es una medida del tamaño de la característica, calculada promediando el cuadrado de la distancia a la posición central, ponderada por la intensidad de píxel (Trackpy Contributors, Trackpy

Documentation, 2015). Tanto el promedio de la fase inicial como en la fase final se encuentran entre un rango de 1.7 y 2.2 de tamaño (Tabla 7).

Tabla 7: Evolución del tamaño en una fase inicial y final promedio de cada una de las muestras.

	% BSW Inicial			% BSW Final			%Diferencia		
nombre_emulsión	Promedio_inicial	min_inicial	max_inicial	Promedio_final	min_final	max_final	dif_cantidad_media	dif_cantidad_min	dif_cantidad_max
ABAPH1HZ02	2,1	1,2	2,7	2,1	1,1	2,6	-2,3%	-8,1%	-3,0%
ABAPH3HZ01	2,1	1,3	2,7	1,9	1,2	2,7	-8,6%	-6,6%	-2,2%
ABAPH3HZ06	2,0	1,1	2,6	2,0	1,0	2,7	-1,0%	-12,1%	4,5%
ABAPH7HZ03	2,2	1,2	2,7	2,1	1,2	2,7	-2,7%	-1,6%	1,4%
ABAPH8HZ09	2,0	1,1	2,6	2,0	1,0	2,7	-0,5%	-12,4%	1,8%
JAZAC03	2,0	1,4	2,6	2,1	1,3	2,7	5,6%	-5,5%	4,0%
JAZCR03	2,0	1,2	2,6	2,2	1,4	2,5	5,3%	15,2%	-2,9%
MORAG05	1,8	1,2	2,5	2,0	1,2	2,7	15,9%	3,8%	5,5%
MORAH204	1,9	1,0	2,6	2,1	1,1	2,6	10,9%	19,4%	0,7%
MORAO01	1,9	1,0	2,8	2,1	1,0	2,7	12,0%	-4,1%	-4,0%
MORE7	2,1	1,2	2,6	1,9	1,0	3,3	-8,7%	-18,0%	28,3%
MORT04	1,7	1,3	2,6	1,7	1,0	2,9	-1,2%	-23,5%	12,1%

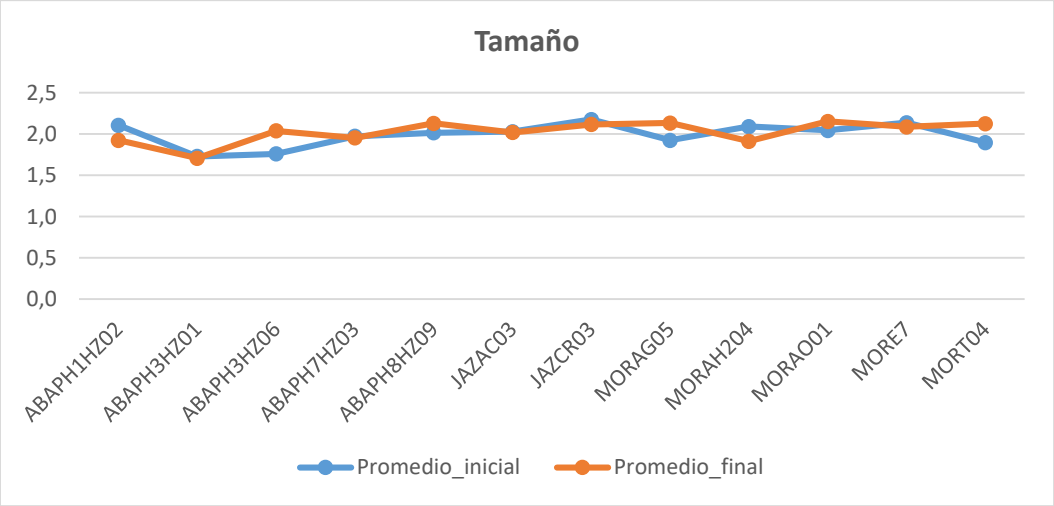

Figura 22: El promedio del tamaño de las partículas en una fase inicial y en una fase final promedio del %BSW.

No hay un cambio muy marcado en el tamaño, es decir luego de una fase final el tamaño promedio de las partículas tiene valores similares (Figura 22).

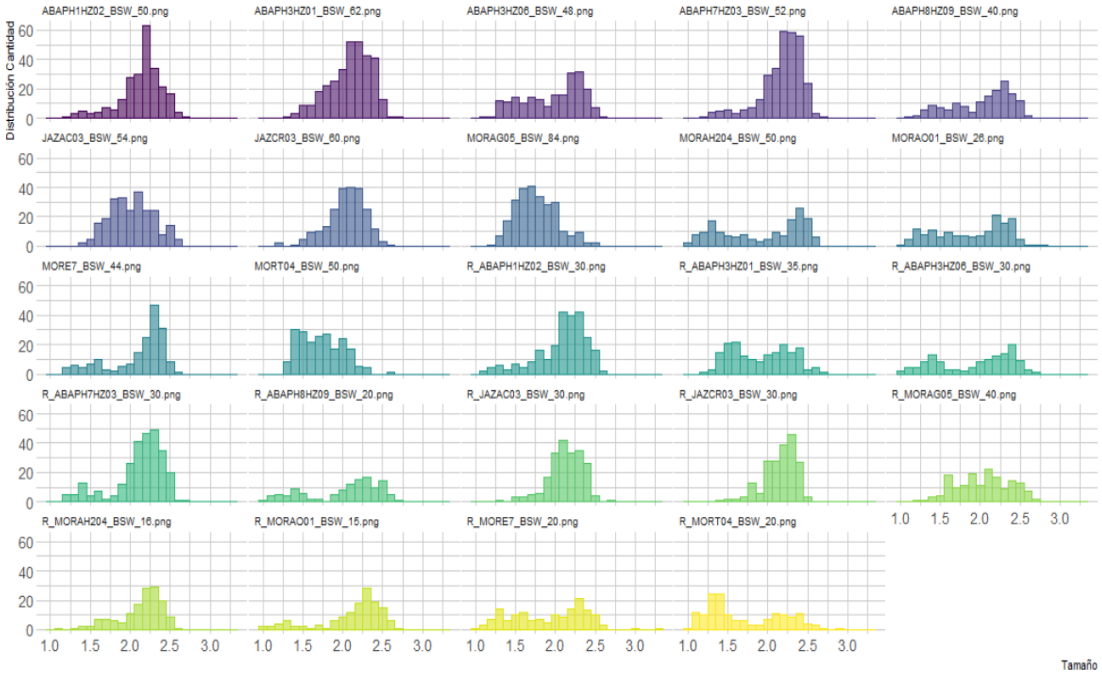

Figura 23: Distribución del tamaño para cada una las partículas en una fase inicial y en una fase final %BSW.

Se encuentran en varios casos que, en la fase inicial, las partículas están concentradas en una sola moda, mientras que una fase final el tamaño de las partículas se expande teniendo varias modas, indicando que luego de un proceso de demulsificación hay una mayor variabilidad en el tipo de tamaño de las partículas dentro de un eje x con un rango de 1.7 a 2.2 (Figura 23).

Señal

El número de fotones detectados siendo una medida relacionada con el contraste (Trackpy Contributors, Trackpy Documentation, 2015). Hay una disminución en la fase final promedio en la mayoría de los casos con respecto a la señal promedio inicial, únicamente las muestras en las que la señal aumento fue en MORE7 y ABAPH7HZ03 (Tabla 8).

Tabla 8: Evolución de la señal en una fase inicial y final promedio de cada una de las muestras.

	% BSW Inicial			% BSW Final			%Diferencia		
nombre_emulsión	Promedio_inicial	min_inicial	max_inicial	Promedio_final	min_final	max_final	dif_cantidad_media	dif_cantidad_min	dif_cantidad_max
ABAPH1HZ02	34,1	17,2	86,2	20,8	7,2	96,9	-39,0%	-58,1%	12,4%
ABAPH3HZ01	34,5	15,3	88,5	31,6	7,7	93,7	-8,5%	-49,4%	6,0%
ABAPH3HZ06	54,1	17,7	121,7	42,0	11,0	103,3	-22,4%	-38,1%	-15,1%
ABAPH7HZ03	30,8	12,0	108,7	35,0	14,1	99,9	13,6%	17,7%	-8,1%
ABAPH8HZ09	47,0	6,4	136,9	40,3	7,9	117,9	-14,3%	21,9%	-13,9%
JAZAC03	36,3	5,5	107,5	29,9	13,9	98,4	-17,4%	153,4%	-8,5%
JAZCR03	45,7	29,2	69,2	34,3	20,2	57,8	-24,9%	-30,9%	-16,5%
MORAG05	77,2	18,7	132,5	38,6	15,3	63,8	-50,0%	-18,3%	-51,8%
MORAH204	39,5	4,7	150,3	19,0	7,4	48,1	-51,8%	56,0%	-68,0%
MORAO01	50,5	7,3	155,4	27,4	6,1	119,9	-45,8%	-16,4%	-22,8%
MORE7	21,7	9,9	66,5	46,9	8,5	114,6	115,9%	-13,8%	72,3%
MORT04	84,2	15,7	154,0	53,9	11,3	110,5	-36,0%	-28,3%	-28,3%

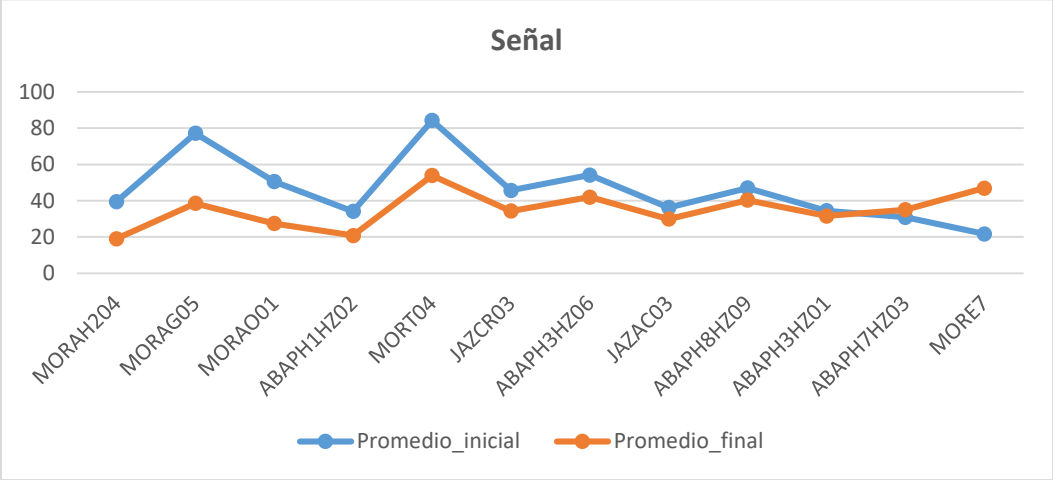

Figura 24: El promedio de la señal de las partículas en una fase inicial y en una fase final promedio del %BSW.

Se evidencia que es más controlado el promedio del %BSW en una fase final para la señal iniciando desde 19 hasta 53 (un rango de 34), mientras que el promedio del %BSW en una fase inicial está dado por una señal desde 21 hasta 84 (un rango de 63) existiendo mucha más variabilidad dentro de las muestras en el estado inicial del %BSW (Figura 24).

Figura 25: Distribución de la señal para cada una las partículas en una fase inicial y en una fase final.

Se encuentra varios casos en los que en una fase inicial las partículas se encuentran concentradas en una sola moda, teniendo un “pico” muy marcado, mientras que en una fase final el tamaño de las partículas se distribuye mucho más, siendo menor el “pico” de partículas que se pueden tener concentradas en un mismo rango (Figura 25).

Épsilon

La incertidumbre estimada en la posición de la característica (Trackpy Contributors, Trackpy Documentation, 2015). Hay una variabilidad muy alta en diferentes fases, existiendo

cambios en las posiciones de las características desde una fase inicial a una final y teniendo en el 75% de las muestras una disminución del épsilon del superior al 23% (Tabla 9).

Tabla 9: Evolución del Épsilon en una fase inicial y final promedio de cada una de las muestras.

Nombre_emulsión	% BSW Inicial			% BSW Final			%Diferencia		
	Promedio_inicial	min_inicial	max_inicial	Promedio_final	min_final	max_final	dif_cantidad_media	dif_cantidad_min	dif_cantidad_max
ABAPH1HZ02	0,3	-27,4	11,3	0,6	-109,4	65,4	107%	299%	480%
ABAPH3HZ01	0,2	0,1	3,9	-0,4	-118,5	9,3	-283%	-216498%	137%
ABAPH3HZ06	0,1	-25,9	1,3	-0,2	-9,7	5,6	-309%	-63%	331%
ABAPH7HZ03	-0,3	-6,7	17,2	-0,1	-2,0	24,6	-69%	-71%	43%
ABAPH8HZ09	0,5	-92,4	74,9	0,1	-85,4	86,9	-68%	-8%	16%
JAZAC03	-0,4	-113,3	7,5	-0,2	-2,5	3,4	-37%	-98%	-55%
JAZCR03	0,6	-8,7	17,3	1,1	-4,5	69,7	92%	-48%	302%
MORAG05	0,2	-11,5	6,3	0,1	-22,0	5,1	-23%	92%	-19%
MORAH204	0,3	-45,2	70,4	0,5	-21,8	76,5	104%	-52%	9%
MORAO01	0,1	-22,0	23,7	-0,1	-2,3	4,0	-179%	-90%	-83%
MORE7	0,3	-3,4	7,5	0,2	-6,1	10,3	-35%	79%	38%
MORT04	1,2	-36,8	127,6	-1,6	-101,7	40,1	-232%	176%	-69%

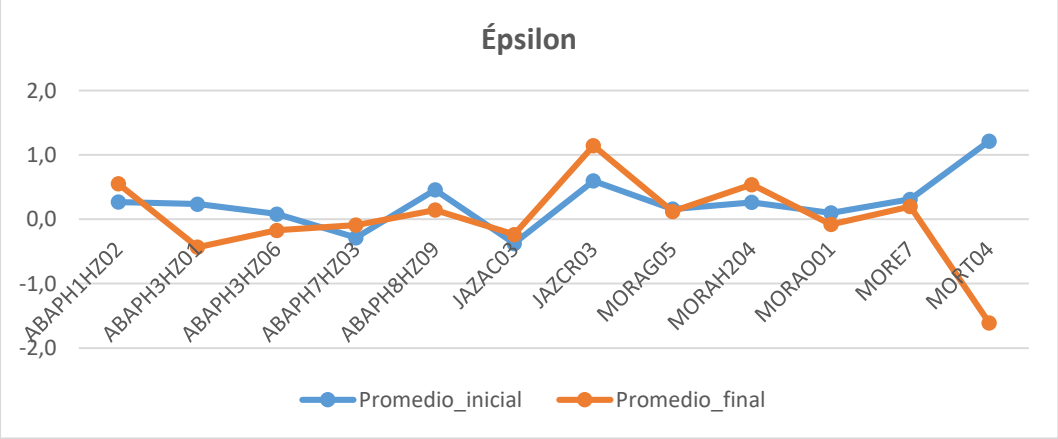

Figura 26: Distribución del Épsilon para cada una las partículas en una fase inicial y en una fase final del %BSW.

En la Figura 26 se encuentra como el 75% de las muestras de emulsiones tienen un épsilon que oscilan entre un rango de -1 y 1, indicando una mayor incertidumbre en la posición de algunas partículas que corresponden a las muestras de MORT04 y JAZCR03 (Savin & Doyle, 2005).

Coordenadas X y la Y

Ubicación en píxeles donde es detectada la partícula, a partir de una estimación del centro de masa de una característica brillante (Allan, Caswell, Keim, Van der Wel, & Verweij, 2021). Las distribuciones de las coordenadas X y la Y (Figura 27 y Figura 28) presentan una distribución uniforme para cada una de las muestras, principalmente entre los valores de 50 a 250 para la coordenada X y de 50 a 200 para la coordenada Y, concentrándose buen porcentaje de las partículas en el centro de la imagen y no hay sesgos destacables en los extremos de las distribuciones. (Trackpy Contributors, 2021).

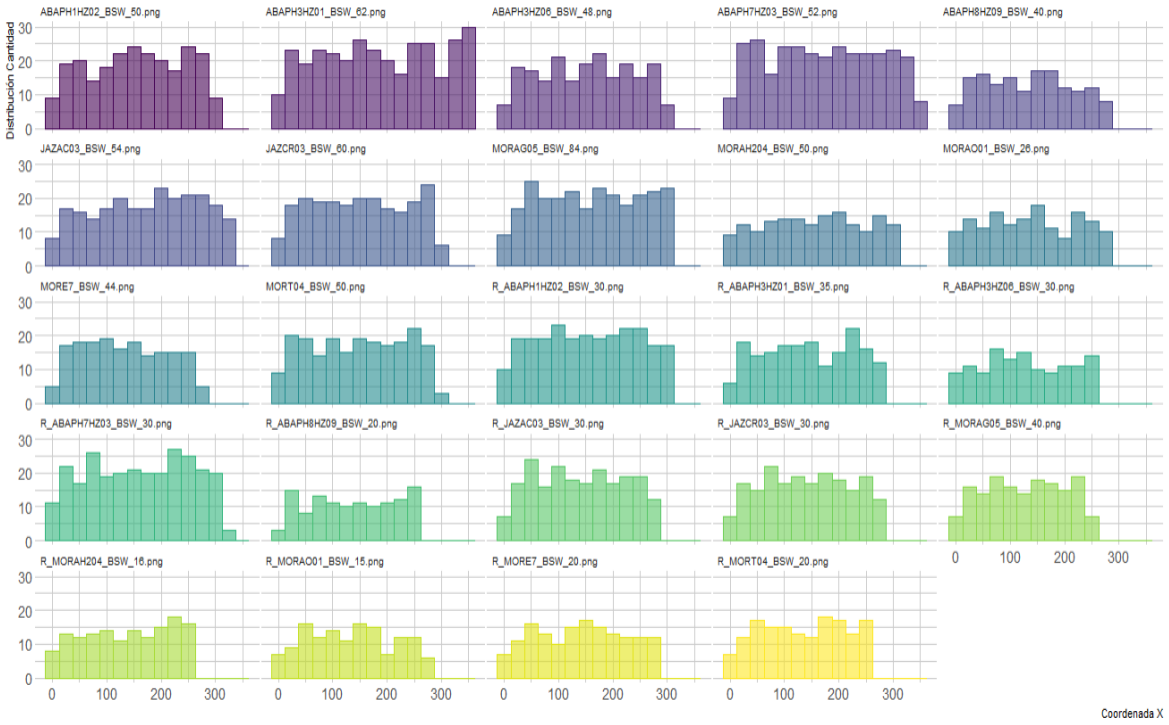

Figura 27: Distribución de la coordenada X para cada una las partículas en una fase inicial y en una fase final del %BSW.

Figura 28: Distribución de la coordenada Y para cada una las partículas en una fase inicial y en una fase final del %BSW.

%BSW

Corresponde a la variable respuesta a predecir (Tabla 10) y es el porcentaje de agua contenido dentro de la emulsión de un crudo (Utria Robinson, 2017), siendo agrupados por 5 categorías con el fin de disminuir la variabilidad en la predicción del valor correcto del %BSW y que los valores contenidos por categoría estén balanceados (Kumar, Bhatnagar, Gaur, & Bhatnagar, 2021).

Tabla 10: Porcentaje del %BSW agrupado en cinco categorías (Unnikrishnan, Donovan, Macpherson, & Tormey, 2020).

Categoría BSW%	[0% al 20%]	(20% al 30%]	(30% al 40%]	(40% al 50%]	(>50%)	Total
Cantidad	712	1.206	497	996	1.357	4.768
%	15%	25%	10%	21%	28%	100%

4.2.2 Correlación entre las variables

El análisis de correlación es empleado para determinar si dos variables están relacionadas o no. El resultado del análisis es un coeficiente de correlación que puede tomar valores entre -1 y +1, cuyo signo indica el tipo de correlación entre las dos variables, siendo una relación positiva entre las dos variables; es decir, cuando la magnitud de una incrementa, la otra también o un signo negativo indica que existe una relación negativa entre las dos variables. (Vinuesa, 2016).

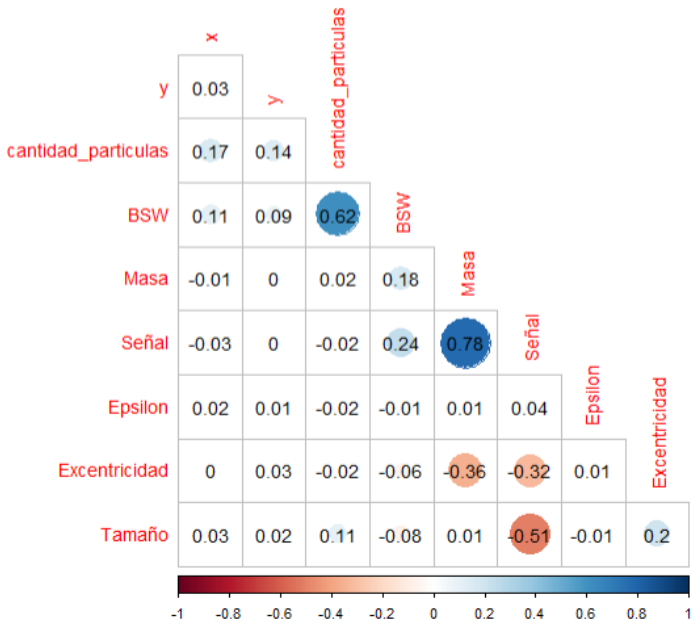

Figura 29: Análisis de correlación por las variables principales por el método del Trackpy.

Mientras los valores de una incrementan, los de la segunda variable disminuyen. Si dos variables son independientes, el coeficiente de correlación es de magnitud cero. La fuerza de la relación lineal incrementa a medida que el coeficiente de correlación se aproxima a -1 o a +1, encontrando una mayor correlación entre la señal y la masa (Figura 29), es decir, cuando la cantidad fotones detectados es mayor la masa o el brillo de la intensidad también se encuentra en aumento, este comportamiento se identifica en la mayoría de las muestras (Figura 30) donde sus comportamientos son crecientes, pero no en su completitud debido a que llegan a un punto de la masa donde la señal comienza a disminuir. También hay una correlación positiva entre el %BSW y la cantidad de partículas, indicando que cuando hay un alto contenido de agua en la emulsión, la detección de partículas también es mayor.

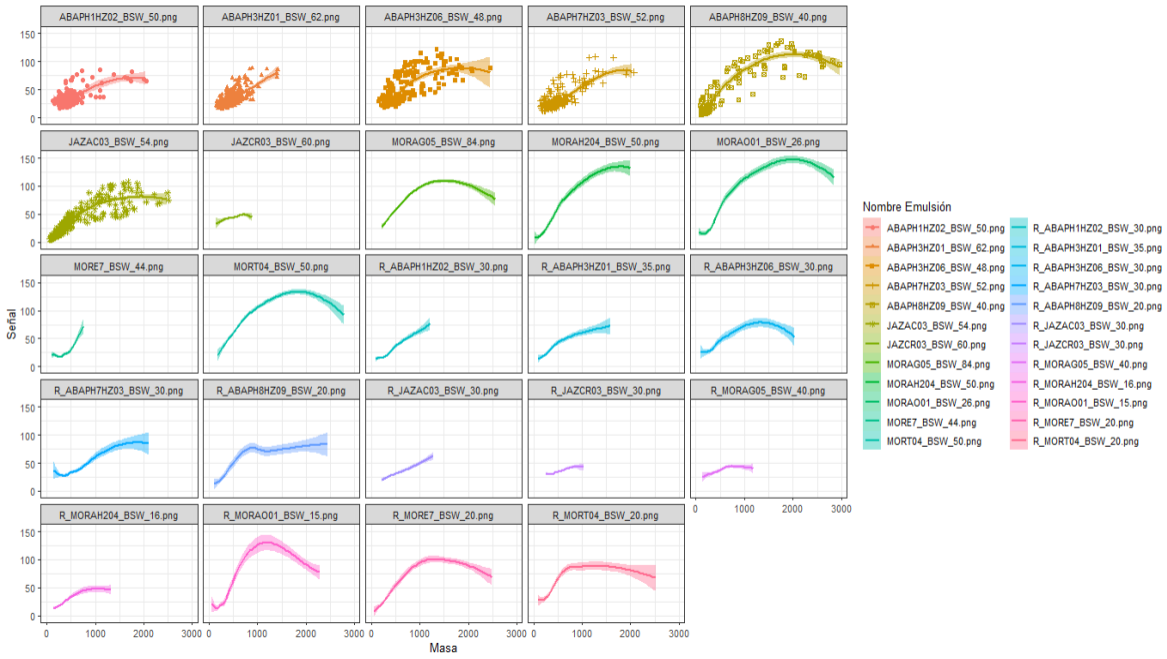

Figura 30: Análisis de correlación por muestras con respecto a la señal y la masa.

4.2.3 Componentes principales

Los componentes principales o PCA son empleados esencialmente para reducir la cantidad de variables, pero aun describiendo los mismos datos, usando PCA, podemos estudiar la relación de varianza acumulada explicada de estas características para comprender qué características explican la mayor variación en los datos (Dunn, 2023). Mediante la Figura 31 se puede identificar la proporción de varianza explicada por cada componente principal y la proporción acumulada de varianza explicada el conjunto de componentes. Indicando que con 5 componentes se puede resumir el 80% de toda la varianza explicada acumulada, encontrando que en la primera componente principal se explica el 27% de la variabilidad.

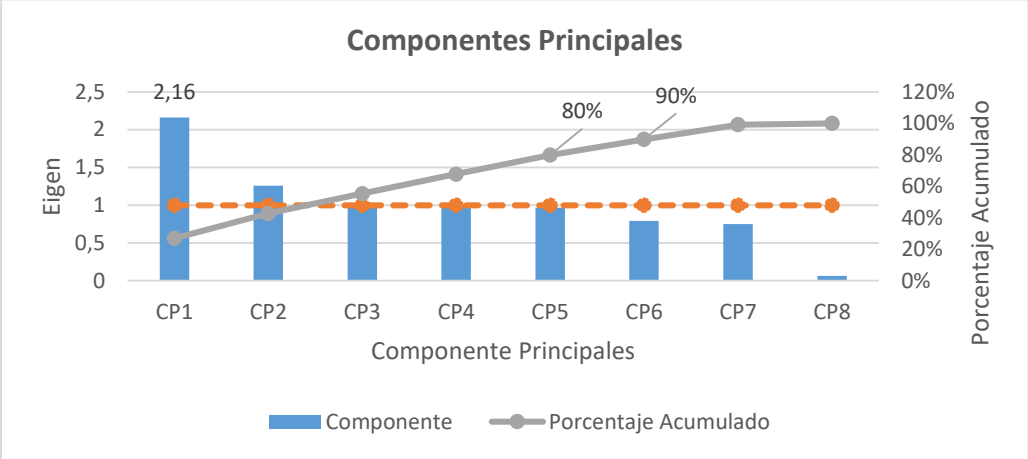

Figura 31: Componentes Principales para las variables que componen a la emulsión.

El cuadro de correlación (Figura 32 a.), muestran las variables que están más fuertemente correlacionadas con las dos primeras componentes principales (CP). Podemos considerar más correladas aquellas que se encuentren fuera del cuadrado $[-0.5,0.5]x[-0.5,0.5]$ que corresponde a la señal, la masa y la cantidad de partículas, mientras que en el círculo de correlación (Figura 32 b.) indica aquellas variables que están cerca de la periferia del círculo de radio 1 las que más contribuyen a las dos primeras componentes que son la masa y la señal hacia la misma orientación y la cantidad de partículas con lados opuestos y siendo ortogonales indicando que no están correlacionados (Gonzalez Rojas, Conde Arango, & Ochoa Muñoz, 2021).

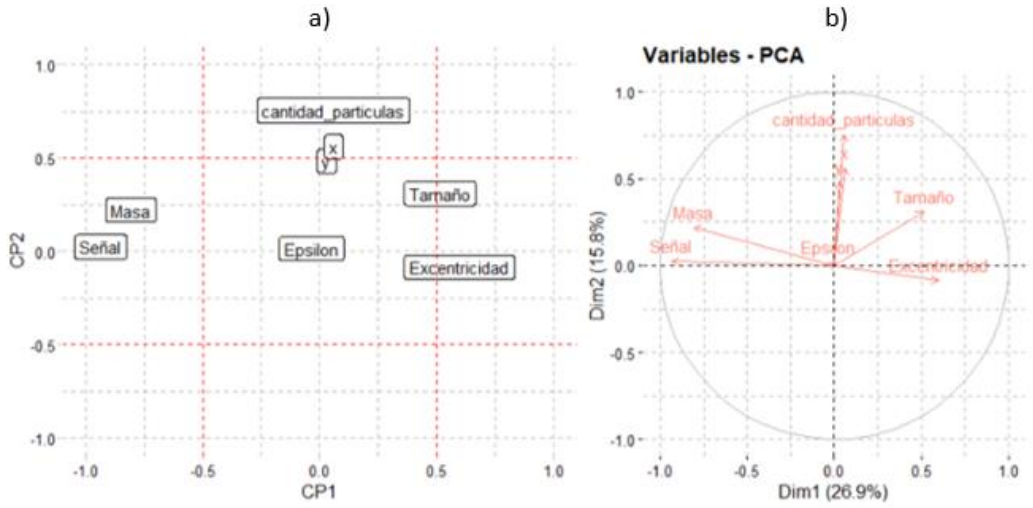

Figura 32: Componentes Principales, a) cuadro de correlación, b) círculo de correlación.

Se emplea la función de correlación (Figura 33 a.) con el fin de resaltar las variables que más contribuyen a cada CP – dimensión, siendo la masa y la señal las que más aportan para la primera componente principal, donde cada cuadrado es un valor y círculos más oscuros y más grandes corresponden a valores más altos al igual que en el caso de la contribución total para Cp1 y Cp2 (Figura 33 b.) siendo la señal y la masa.

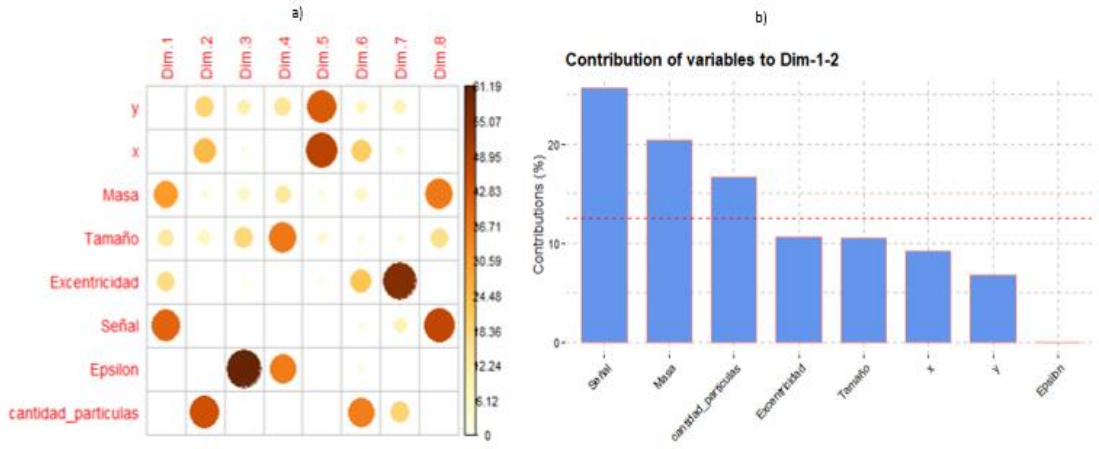

Figura 33: Contribución por componente principal, a) función de correlación, b) contribución Total a Cp1 y Cp2.

4.3 Método para la predicción del contenido de agua en una emulsión de un crudo pesado empleando técnicas de aprendizaje de máquinas supervisado

De acuerdo al análisis anterior realizado dividimos nuestros datos para darle a los modelos empleados la mayor cantidad de datos posible para el proceso de entrenamiento, sin embargo, también se requiere tener la seguridad de contar con los suficientes datos para que el modelo se pruebe a sí mismo (Lun Chao, 2011). En general, a medida que aumenta el número de filas en el conjunto de datos, más datos podemos dar al conjunto de entrenamiento, de manera queda dada la cantidad de datos disponibles y el estándar definido para los modelos de aprendizaje de máquinas supervisado se opta por una división de 80% de entrenamiento y 20 % de prueba (Hung Nguyen, Bang Ly, Al-Ansari, & Thai Pham, 2021), garantizando que tanto el conjunto de entrenamiento como la prueba tengan la misma proporción para cada una de las clases del %BSW del conjunto de datos original.

Para la elaboración del modelo se tuvieron en cuenta todas las variables anteriormente mencionadas, se probaron los siguientes modelos tanto las variables con su distribución original como el método anterior aplicado de PCA cuya estandarización de los datos controla el hecho de que diferentes variables se miden en diferentes escalas para que cada predictor no influya de acuerdo a su distribución y poder decidir el grado de importancia para cada modelo, también para que los modelos de aprendizaje funcionen mejor o convergen más rápido cuando las características están en una escala relativamente similar y/o cerca, (Hale, 2019), sin embargo, para todos los modelos desarrollados se encontró que los datos que brindaban una mejor capacidad de predicción son los datos con sus distribuciones originales, esto se identificó con una validación previa con las métricas de rendimiento.

Se emplean los métodos de aprendizaje de máquina sobre una variable de respuesta continua que está simplificada en 5 intervalos, esto con el fin de brindar una interpretación más fácil de acuerdo al cálculo del porcentaje del %BSW se comporta mejor la predicción ya sea un porcentaje con un bajo contenido de agua en el crudo o uno alto, obteniendo cuando se realiza la predicción de un intervalo que nos indica donde se encuentra el valor y la probabilidad de que corresponda a ese intervalo.

Dentro de la configuración de las opciones de control de entrenamiento, se especifica para los cuatro modelos una validación cruzada repetida (más específicamente, 10 repeticiones de una validación cruzada de 10 veces), estos son los parámetros iniciales con los cuales se ajusta y se evalúa cada modelo, determinando las variables más importantes, siendo entendidas como las que tienen un mayor efecto sobre la predicción de la variable respuesta, siendo transformadas en una

escala porcentual con el fin de entender cuál es el porcentaje de aporte a la predicción con respecto a la variable respuesta %BSW.

También se determina la matriz de confusión con los indicadores de precisión, sensibilidad, especificidad, precisión del balance, son ajustados los parámetros del modelo mediante una optimización de parámetros (Bagui & Mink, 2022).

Todos los modelos son ejecutados desde un computador con una memoria RAM 8 GB y un procesador de 2.61 GHZ.

4.3.1 Regresión Multinomial

La Regresión Logística Multinomial tuvo un tiempo de ejecución de 1.85 minutos para el proceso de entrenamiento, teniendo como variables más importantes la excentricidad, el tamaño y la cantidad de partículas, teniendo las tres variables un efecto sobre la variable respuesta del 99%(Figura 34 c.), se iteró sobre el modelo cada una de las variables, encontrando una mejor precisión para el modelo sin la variable Masa, La precisión de entrenamiento es del 57% y para prueba del 58%, realizando la optimización de los parámetros, se indica que el mejor ajuste se obtiene con menos de 2 variables para lograr la mayor predicción (Figura 34 c.).

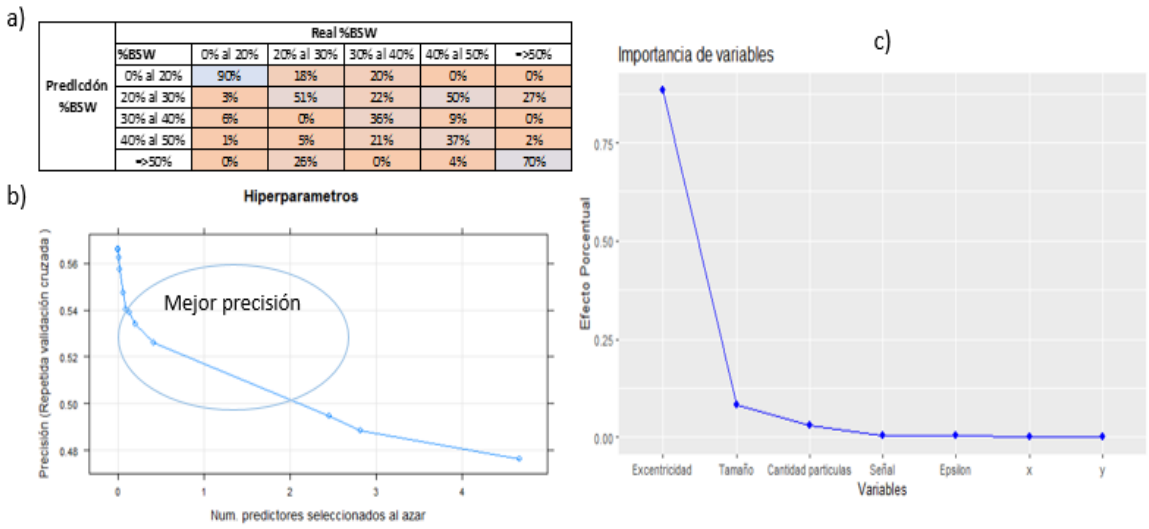

Figura 34: Resultados modelo de Regresión Multinomial, a) matriz de confusión (Real vs. predicho), b) Hiperparámetros, c) Variables importantes.

De acuerdo al resultado de la matriz de confusión (Figura 34 c.) en tres casos la sensibilidad se encuentra inferior del 50%, la clase en donde hay una mejor clasificación es cuando él %BSW se encuentra en un intervalo de “[0% -20%]” al igual cuando él %BSW es mayor de “(>50%)”. En el caso de la especificidad en las categorías es superior del 71%, siendo la clase “(30-40%]” donde se encuentra clasificados en verdaderos negativos que se clasifican como negativas es el 96% y con una precisión equilibrada del 72% promedio para las 5 categorías.

4.3.2 Árboles Aleatorios.

El modelo de árboles aleatorios, tuvo un tiempo de ejecución de 3 minutos para el proceso de entrenamiento, siendo las variables más importantes la cantidad de partículas, épsilon y la señal teniendo un efecto sobre la variable respuesta del 93% (Figura 35 c.), obteniendo una precisión de

entrenamiento del 97% y para prueba del 98%. Realizando la optimización de los parámetros el mejor ajuste se obtiene con 7 variables para lograr la mayor predicción (Figura 35 b.).

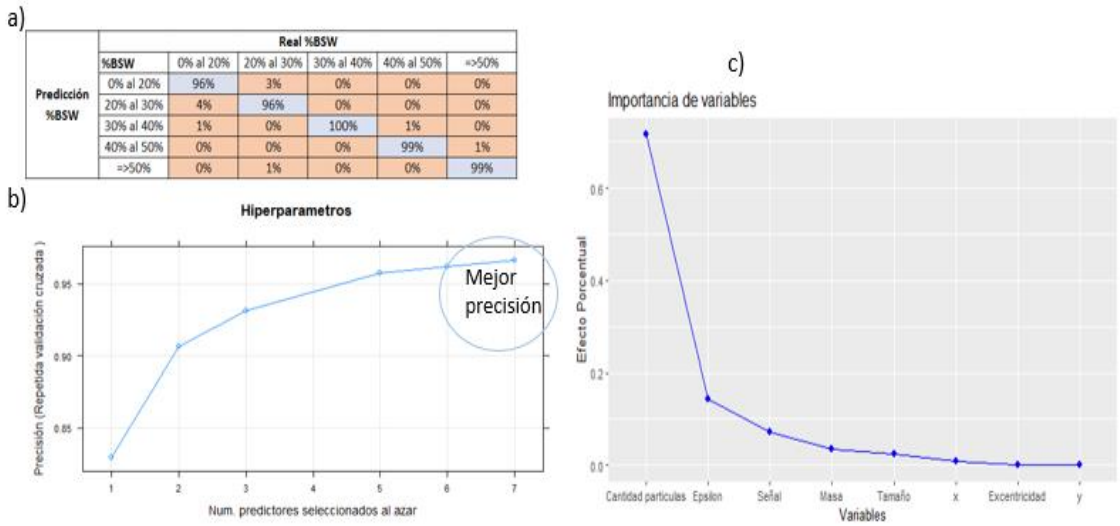

Figura 35: Resultados modelo de Bosques Aleatorios, a) matriz de confusión (Real vs. predicho), b) Hiperparámetros, c) Variables importantes.

De acuerdo al resultado de la matriz de confusión (Figura 35 a.) en todos los casos la sensibilidad se encuentra superior al 96%, la clase en la que hay una mejor clasificación es cuando él %BSW se encuentra en un intervalo de a “(30%-40%]”.

En el caso de la especificidad en la categoría menor e igual a un %BSW del “[0%-20%]” donde se encuentra clasificados en verdaderos negativos y que son clasificados como negativos es el 99%, teniendo una precisión equilibrada del 98% promedio para las 5 categorías.

4.3.3 Potenciación del gradiente

El modelo de Potenciación del gradiente, tuvo un tiempo de ejecución de 9 minutos para el proceso de entrenamiento, siendo las variables más importantes la cantidad de partículas, épsilon y la señal, teniendo estas tres variables un efecto sobre la variable respuesta del 99% (Figura 36 c.), obteniendo una precisión tanto de entrenamiento del 99% y para prueba del 99%, realizando la optimización de los parámetros el mejor ajuste se obtiene con 150 árboles con una profundidad de interacción o complejidad del árbol que sería 3 para lograr la mayor predicción (Figura 36 b.).

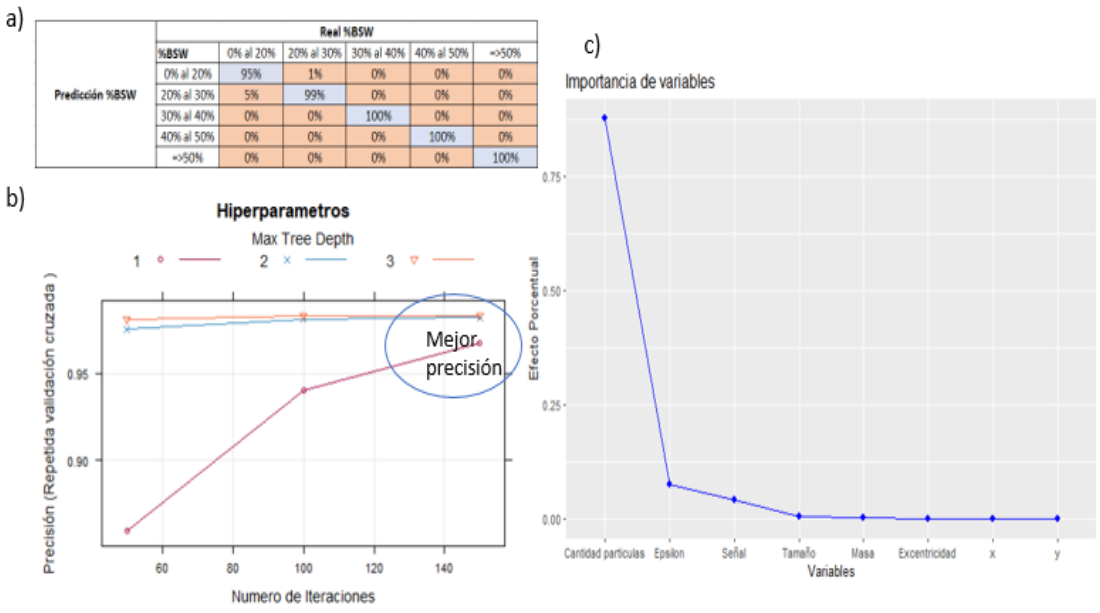

Figura 36: Resultados modelo de Potenciación del Gradiente, a) matriz de confusión (Real vs. predicho), b) Hiperparámetros, c) Variables importantes.

De acuerdo al resultado de la matriz de confusión (Figura 36 a.) en la mayoría de todos los casos la sensibilidad se encuentra superior al 95%, la clase que difiere de las demás porque el valor

es menor es cuando él %BSW se encuentra en un intervalo de a “[0%-20%]”. En el caso de la especificidad en todas las categorías del %BSW donde se encuentra clasificados en verdaderos negativos se clasifican como negativos es el 99% y teniendo una precisión equilibrada del 99% promedio para las 5 categorías.

4.3.4 Red Neuronal

El modelo de Red Neuronal, tuvo un tiempo de ejecución de 7,6 minutos para el proceso de entrenamiento, siendo las variables más importantes el tamaño, la excentricidad, cantidad de partículas teniendo un efecto sobre la variable respuesta del 68% (Figura 37 c.), obteniendo una precisión de entrenamiento del 44% y para prueba del 43%, realizando la optimización para el caso de las redes neuronales de una sola capa se ajustan dos parámetros como los son el número de unidades ocultas y la disminución del peso, siendo el tamaño de 5 y del decaimiento 0.21 los parámetros que permiten lograr la mayor predicción (Figura 37 b.), quedando con una arquitectura de 8-5-5 (8 entradas,5 capas ocultas y 5 salidas con un peso de 0.21).

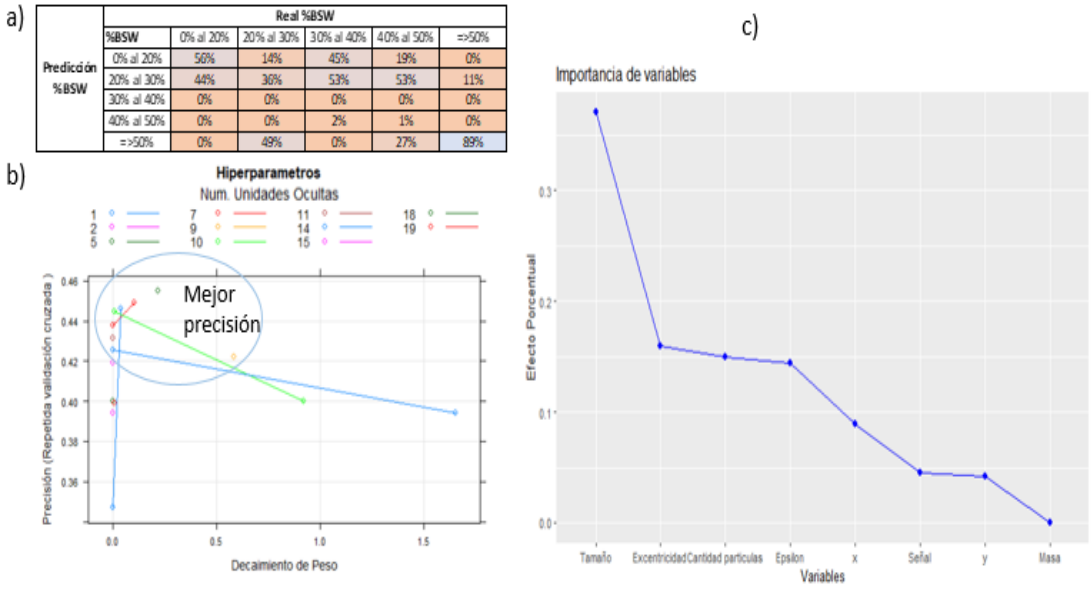

Figura 37: Resultados modelo de Red Neuronal, a) matriz de confusión (Real vs. predicho), b) Hiperparámetros, c) Variables importantes.

De acuerdo al resultado de la matriz de confusión (Figura 37 a.) en la única categoría en la que la sensibilidad se encuentra con un valor significativamente alto del 89% es cuando se tiene un %BSW en un intervalo de “(>50%)”.

En el caso de la especificidad en la categoría del %BSW “[0%-20%]” y “(=50%>)” donde se encuentra la mayor cantidad de datos, los verdaderos negativos que se clasifican como negativos son del 85% y 75%, con una precisión equilibrada del 70% y del 82% para ambas categorías.

4.4 Resultados

Comparando el desempeño de cada uno de los modelos con el conjunto de datos de prueba, se identifica una variabilidad tanto en los tiempos de ejecución de cada método como en el

porcentaje de predicción, corresponde a los Bosques Aleatorios con el 97% y la Potenciación del Gradiente con el 99%, cuyos valores son muy similares en la predicción, pero con una diferencia en tiempos mucho menor para los Bosques Aleatorios, siendo un tiempo menor a tres veces el valor obtenido por la Potenciación del Gradiente (Figura 38).

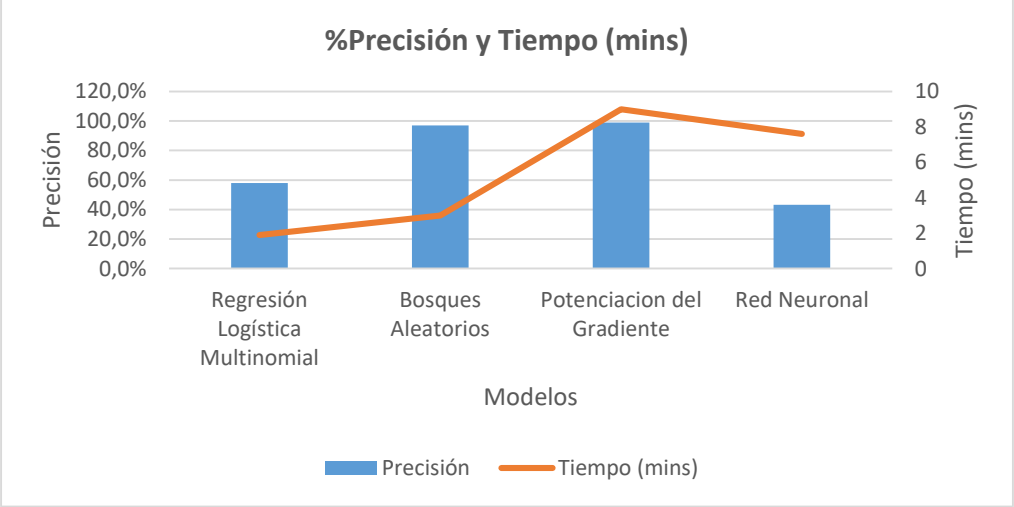

Figura 38: Resultados del porcentaje de precisión y tiempo de ejecución de los modelos supervisados.

En el caso del porcentaje de sensibilidad (Figura 39) los Bosques Aleatorios y la Potenciación del Gradiente se comportan de manera similar clasificando correctamente las muestras positivas que son verdaderas, superior al 96% teniendo mejores desempeños cuando los niveles del %BSW es superior al 30%. Mientras que en el caso de la Regresión Logística Multinomial y la Red Neuronal únicamente tienen una aceptable sensibilidad en los niveles de los extremos del %BSW cuando es “[0%,20%]” o “(>50%)”.

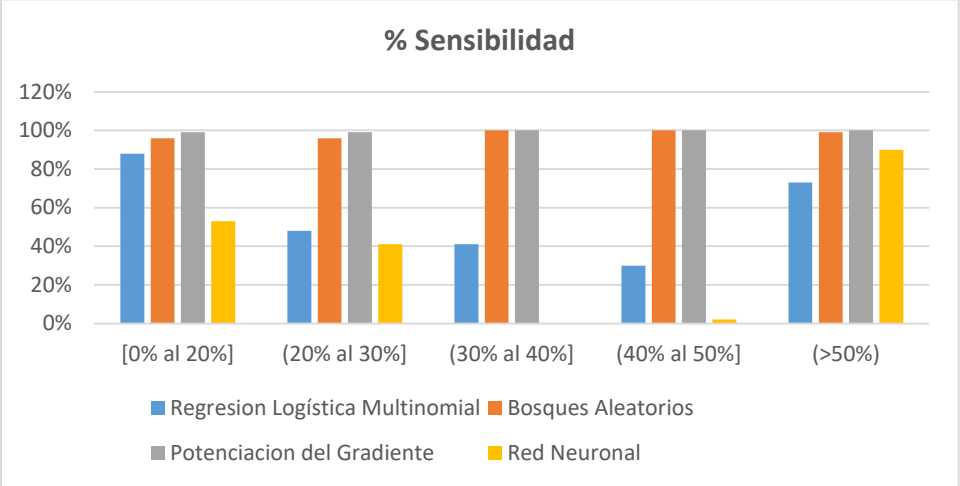

Figura 39: Resultados del porcentaje de sensibilidad de los modelos supervisados.

En el caso de la especificidad, cuando la proporción de los verdaderos negativos se clasifican como negativos (Figura 40), el método de Bosques Aleatorios y la Potenciación del Gradiente son constantes en los resultados superiores al 99%. Mediante la especificidad, los métodos de Regresión Logística y de Red Neuronal indican un rendimiento mucho mejor para la detección (superior del 66%) al ser comparado con la presión y la sensibilidad.

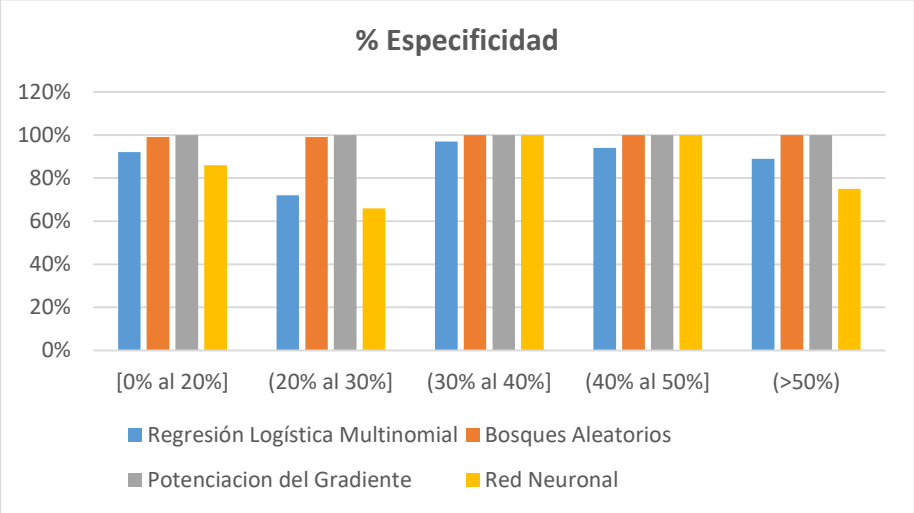

Figura 40: Resultados del porcentaje de especificidad de los modelos supervisados.

La precisión equilibrada indica (Figura 41), en los métodos de Bosques Aleatorios y la Potenciación del Gradiente hay una mejor medida de precisión para cada nivel de %BSW, estando muy bien balanceados cada una de clases al realizarse la modelación. En los métodos de Regresión Logística y Red Neuronal, se mejora el rendimiento debido a un mayor porcentaje en la especificidad, sin embargo, las variabilidades en los porcentajes indican un desequilibrio del modelo para la detección de cada uno de los niveles del %BSW.

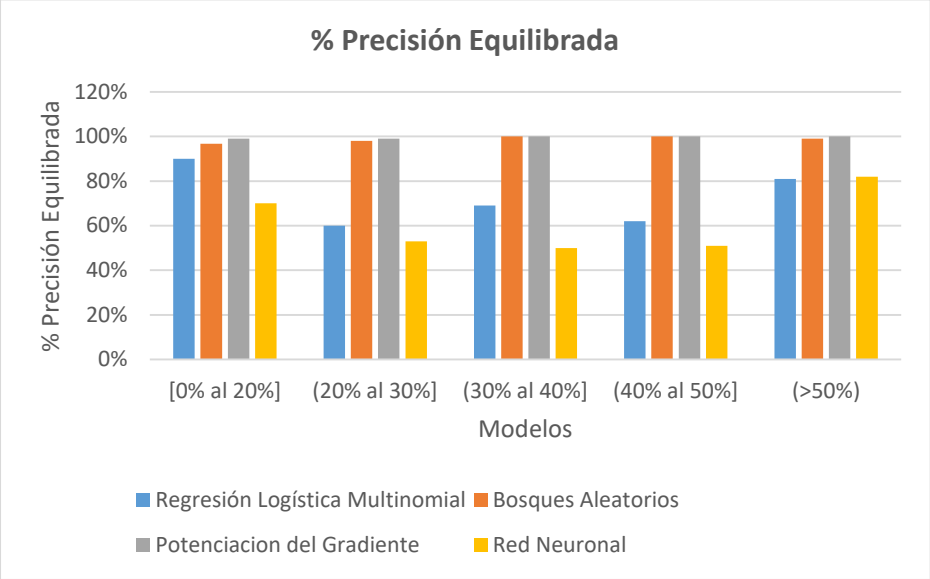

Figura 41: Resultados del porcentaje de Precisión Equilibrada de los modelos supervisados.

La curva de ROC (Figura 42) ilustra la sensibilidad y especificidad para cada uno de los niveles del %BSW. La curva ROC se construye con base en la unión de distintos puntos de corte, correspondiendo el eje Y la sensibilidad y el eje X a (1-especificidad) de cada uno de ellos. Ambos ejes incluyen valores entre 0 y 1 (0% a 100%). A modo de referencia, en todo gráfico de curva ROC se traza una línea desde el punto 0,0 al punto 1,1, llamada diagonal de referencia, notando como en todos los niveles los modelos de Bosques Aleatorios y Potenciador Gradiente tienen resultados muy similares y están muy cerca del ángulo superior- izquierdo que corresponde a una sensibilidad y especificad del 100%. En los casos de la Regresión Logística el rendimiento más adecuado es en el nivel inicial con un %BSW [0%-20%] al estar mucho más alejada de la diagonal de referencia, mientras que el desempeño de la Red Neuronal es inferior a los resultados de los demás modelos.

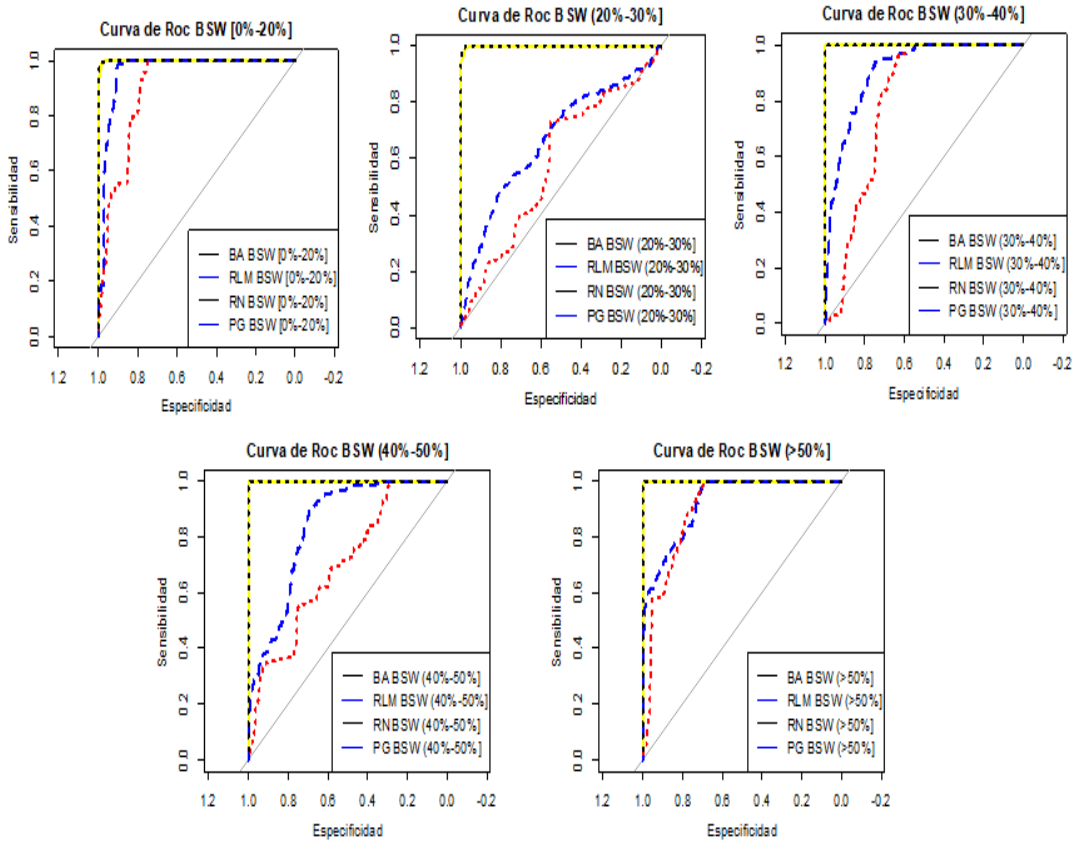

Figura 42: Resultados del área bajo la curva (AUC) para cada uno de los modelos supervisados en las cinco categorías del %BSW.

La importancia de las variables (Tabla 11), en los métodos de Bosques Aleatorios y la Potenciación del Gradiente el mayor porcentaje de impacto se encuentra en la variable de cantidad de partículas, el épsilon y en la señal, en ambos casos está contenido más del 90% del efecto de estos modelos sobre la variable respuesta del %BSW.

Tabla 11: Comparativo de importancia de variables para cada uno de los modelos supervisados.

Variables Importantes	% Importancia Bosques Aleatorios	% Importancia Regresion Logistica Multinomial	% Importancia Red Neuronal	% Importancia Potenciación del Gradiente
Cantidad de particulas	71,6%	3,0%	14,9%	87,6%
Epsilon	14,4%	0,2%	14,4%	7,4%
Señal	7,1%	0,5%	4,5%	4,1%
Masa	3,6%	0,0%	0,0%	0,2%
Tamaño	2,3%	16,6%	37,0%	0,5%
x	0,8%	0,0%	8,9%	0,0%
Excentricidad	0,1%	79,7%	16,0%	0,1%
y	0,0%	0,0%	4,2%	0,0%

4.5. Evaluación de desempeño de los métodos propuestos con respecto a los resultados en la literatura

Considerando como análisis desde la literatura, se identifica un artículo en el cual se exponen imágenes sobre la investigación en emulsiones y el porcentaje de agua identificada por los estudios del autor. Estas imágenes son seleccionadas desde el artículo, se extraen manualmente y se adquieren las diferentes características mediante las técnicas de seguimiento de partículas (Trackpy) y son empleadas las técnicas de aprendizaje de máquina junto con las métricas de validación.

El artículo corresponde a “Emulsiones con crudo pesado en presencia de nanopartículas” (Riaza, Cortés, & Otalvaro, 2014), teniendo tres escenarios (Figura 43 a.) que indican el porcentaje de agua contenido en las imágenes que son (4%,20%,48%) y con algunas variaciones en las características como lo es la concentración de las nanopartículas (el peso) y el tiempo de estabilización de la emulsión (luego de un proceso de agitación).

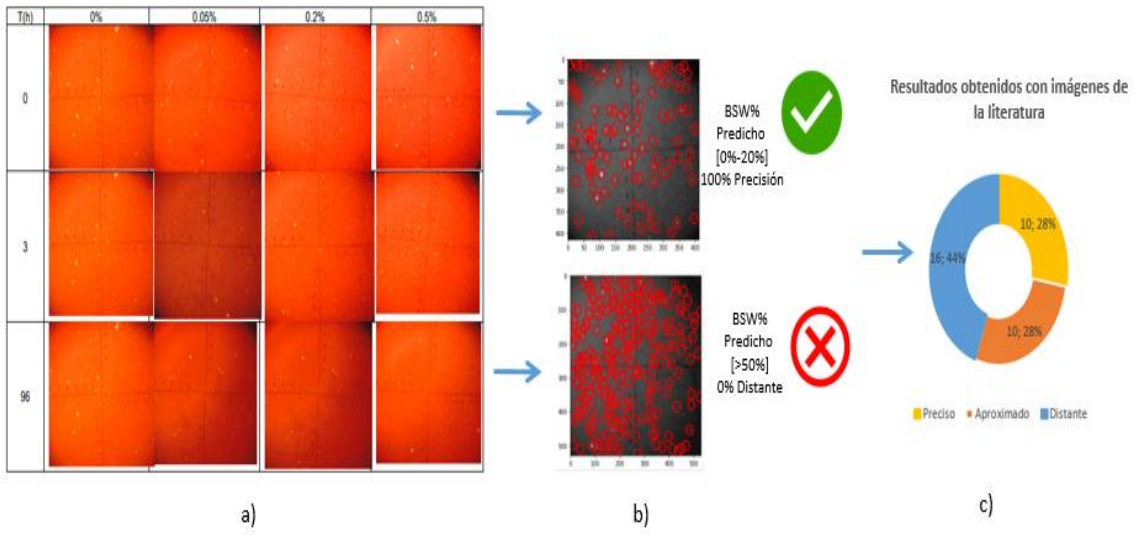

Figura 43: Análisis de las imágenes de la literatura, a) Imágenes del artículo (Riaza, Cortés, & Otalvaro, 2014), b) caracterización y predicción del %BSW, c) Resultados por los métodos aplicados.

Obteniendo un total de 36 imágenes de las cuales fueron validadas por los modelos que tuvieron el mejor rendimiento, siendo los Bosques Aleatorios y la Potenciación del Gradiente.

Los resultados indican que el porcentaje de predecir correctamente el contenido de agua en la imagen de la emulsión es del 28% (10 imágenes) siendo resultados basados en una identificación

adecuada de las partículas en la detección del brillo en la imagen como en la (Figura 43 b.), mientras que en los casos en los que fue aproximado del 28% (10 imágenes), fueron en los casos en los que la predicción estuvo a un nivel de cercanía del indicado por el artículo de donde se extraen las imágenes (Riaza, Cortés, & Otalvaro, 2014). En el escenario en donde la predicción estaba desfasada, corresponde a un 44% (16 imágenes) se debe a que sus resultados son más de un nivel distante del valor real indicado por el artículo, en estos casos el método se vio afectado porque detecto muchas más partículas de las que realmente podría tener la emulsión de acuerdo al contenido del agua que identifico el autor (Figura 43 c.).

Tabla 12: Validación de la predicción con respecto al porcentaje real del contenido del agua indicado por el artículo (Riaza, Cortés, & Otalvaro, 2014)

%BSW	Preciso	Aproximado	Distante
%BSW 4	8 (67%)	0 (0%)	4(33%)
%BSW 20	0 (0%)	3 (25%)	9(75%)
%BSW 48	2 (17%)	7 (58%)	3(25%)

Evidenciando el mayor porcentaje de desfase o distante con respecto al porcentaje de agua indicado desde la literatura es cuando él %BSW es del 20% (Tabla 12), mientras que la mayor precisión se logra cuando se tiene el %BSW del 4%, correspondiendo a un escenario que no es muy afectado por las otras variables que no hicieron parte dentro de la construcción del método de aprendizaje de máquinas.

5. Aplicación de la solución implementada

Se construye un aplicativo cuyo objetivo es facilitar el acceso a la evaluación de emulsiones, propiciando información sobre las características, detección y predicción del porcentaje de agua contenida %BSW.

Para ello, se ha desarrollado una interfaz de interacción con el usuario de fácil manejo, y sus valores de salida sean fácilmente interpretables (Figura 44). El programa está dividido en cuatro etapas, donde el usuario interactúa y puede tener intervención o modificación con las diferentes opciones de la solución, cuyos pasos para ser utilizado son los siguientes:

1) El usuario identifica una emulsión de interés de la cual quiere adquirir la mayor información, cabe mencionar que la calidad de la imagen es un factor determinante para lograr tener una mayor comprensión sobre la emulsión de estudio (Figura 44 a), es recomendable mantener el mismo formato de las imágenes que han sido analizadas en el presente trabajo.

2) El usuario almacena la imagen en una carpeta del aplicativo "Google Drive" (Nolledo, 2020) para guardar la imagen de interés junto con las imágenes ya existentes con las cuales fueron entrenados los modelos previos (Figura 44 b).

3) El usuario deberá contar con acceso a "Google Colaboratory" (Google, 2017) el cual es un aplicativo de código libre en un servidor propio de Google cuya interfaz es "Jupyter Notebook" que es un cuaderno de código abierto para computación interactiva en todos los lenguajes de programación (Jupyter, 2021), en este programa se encuentra la solución completa del análisis de las emulsiones desde una fase de lectura, procesamiento, modelación e ingestas de resultados. El usuario que estará haciendo uso del aplicativo lo único que deberá de hacer es

ejecutar completamente esta interfaz de código (Figura 44 c). Después de una ejecución de varios minutos tendrá un archivo en formato “txt” con la información sobre cada paso ejecutado y la finalización exitosa del proceso (Figura 44 d).

4) Luego de que el usuario valide la finalización del proceso, ingresa en la interfaz libre de Google llamada “Looker Studio” (Looker Studio, 2022), en este aplicativo podrá filtrar las imágenes que son de interés para analizar las partículas de agua detectadas, la distribución de los atributos que la caracterizan como es el tamaño, la excentricidad, la señal, la cantidad de partículas, y la probabilidad asignada a cada una de las categorías del %BSW (Figura 44 e).

5) Ingresando en este link se puede tener acceso a la interfaz: https://lookerstudio.google.com/reporting/cf804b3b-ce8a-4525-98f2-152bcb63ec73

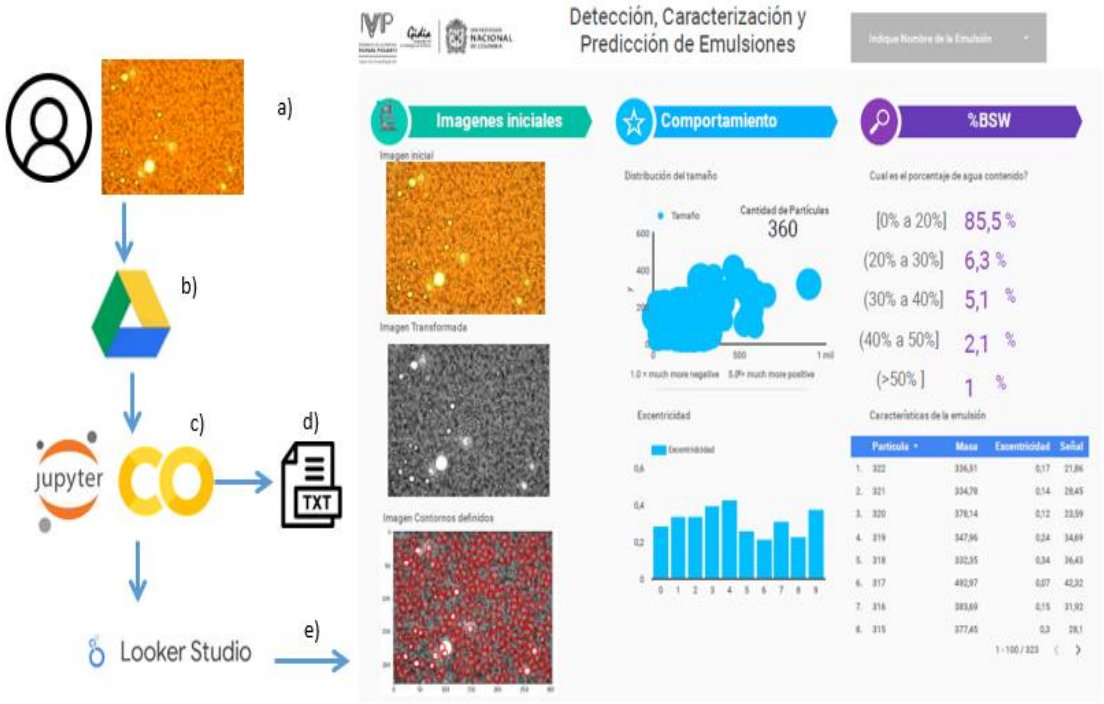

Figura 44: Aplicativo implementado, a) Selección de la imagen de estudio**,** b) Guardado o almacenamiento de la imagen en "Google Drive" (Nolledo, 2020), c) Ejecución de todo el análisis desde "Google Colaboratory" (Google, 2017), d) Visualización de los resultados desde Google "Looker Studio" (Looker Studio, 2022).

6. Conclusiones

La caracterización de emulsiones de agua en aceite tiene una importancia fundamental dentro de la industria de hidrocarburos, dada la frecuencia con la que presentan estas condiciones en el crudo. Lo anterior complica la producción de petróleo, dado que este debe ser separado del agua en el menor tiempo posible antes de llegar al proceso de su refinación. Por tanto, se estudiaron variables como el tamaño, la masa, la cantidad de partículas, la señal, las coordenadas de las partículas, la excentricidad, épsilon, cada una aportando en el entendimiento del %BSW desde que se recibe la muestra en el laboratorio hasta que se aplica los métodos de demulsificación.

El análisis comparativo entre una fase inicial y final del %BSW permitió identificar una diferencia entre la cantidad de gotas detectadas de una fase a otra y como esto hace que se encuentre menos partículas de gotas, también como se ve disminuido el brillo dentro de las partículas; por otro lado, el cambio en la forma de las partículas en varias muestras vinculado a la perdida de la figura esférica definida en una fase inicial, adicionando el cambio en el tamaño de las partículas al pasar de una distribución concentrada en una misma modal a una distribución con varias modas.

Por parte de los métodos de aprendizaje de máquina supervisados, el desempeño tanto en las métricas de rendimiento como de tiempo indican un mejor ajuste de los datos al método de Bosques Aleatorios, parte de este resultado se debe a la importancia del efecto de las características de la cantidad de partículas, épsilon y la señal, con respecto al %BSW.

Si bien el método de Bosques Aleatorios y el de Potenciación del Gradiente tuvieron los mejores resultados con los conjuntos de datos tanto de entrenamiento y prueba, se requiere de una

constante mejora de todo el flujo de análisis, iniciando desde una correcta adquisición de imágenes, depuración de atributos inválidos, modelación con parámetros eficientes en rendimientos y tiempos, al igual que ir poblando el repositorio de imágenes de emulsiones cada vez más. Esto se identifica debido a los resultados obtenidos con las imágenes de la literatura, dado el reto que significo al tratarse de un formato diferente de micrografías y de nuevas variables que son el nivel de concentración de nanopartículas y tiempo de estabilización.

Para este estudio se analizó un conjunto de imágenes de emulsiones de agua en aceite (W/O) adquiridas mediante microscopia óptica, se identificaron las características por medio de seguimiento de partículas permitiendo crear un conjunto de datos que sintetiza las propiedades de las muestras, aplicando métodos estadísticos con el fin de entender el comportamiento o distribución de cada una de las variables para finalizar implementado y evaluando los métodos de aprendizaje de máquina supervisado para detectar el contenido de agua en una emulsión para finalizar con el desarrollo de una herramienta automatizada para el procesamiento de dicha imágenes, siendo un método alternativo mediante el marco de la ingeniería de sistemas que permite generar información significativa en esta área de investigación desde la metodología del análisis de imágenes, la implementación de los métodos de aprendizaje de máquinas, la automatización de todo el proceso, de tal forma de que se logre llegar a una respuesta similar a la obtenida mediante los métodos químicos con un enfoque que da beneficios desde el aspecto de precisión para el análisis de grandes cantidades de datos e imágenes, automatización en el proceso de detección, eficiencia en el tiempo de la respuesta, adaptabilidad ante los diferentes tipos de muestras de agua en aceite, reducción de costos al eliminar la necesidad de equipos costosos y/o materiales consumibles o desechables que son utilizados dentro de los métodos convencionales de medición.

En particular se identifica con respecto a las investigaciones realizadas por los autores, de que la solución propuesta en el presente trabajo integra toda una solución completa en la fase de análisis, es decir, luego de la adquisición de las micrografías, se realiza tanto el análisis de las imágenes como la implementación de las técnicas de aprendizaje de máquina desde un solo aplicativo, también la integración de otros tipos de variables que no son tenidas en cuenta por parte de los autores y desde el presente trabajo no se tuvieron presente como lo es el diámetro de Feret o escenarios de evolución en el tiempo de la imagen, se considera de igual forma por parte de los autores la evolución en la cantidad de partículas en comparación de las variables de análisis, si son influidos ya sea disminuyendo o aumentando una característica específica. Se realiza de igual forma el submuestreo propuesto por los autores y se adiciona desde el presente trabajo la búsqueda de los hiperparámetros mediante una búsqueda aleatoria y se validan los resultados mediante las métricas de rendimiento para cada uno de los intervalos donde se predice el porcentaje del agua %BSW, cabe mencionar que los autores dan una aproximación en los métodos de aprendizaje de máquinas mediante las redes convolucionales lo cual será abordado de igual forma en trabajos futuros. Se complementa finalmente implementando una interfaz de usuario de consulta, este tipo de recurso no se vio que fuera implementado desde la literatura por parte los autores.

7. Trabajo futuro

Dando uso de las técnicas y procedimiento en el marco del desarrollo del presente trabajo, tenemos la intención de analizar más técnicas que ponga a competir el desempeño del método de seguimiento de partículas con el fin de robustecer la toma de decisiones integrando nuevas características mediante segmentación de imágenes, lo cual proporciona una gran cantidad de rutinas que se ajustan a los retos aquí presentados (Gómez, Maška, Kotrbová, Pospíchalová, & Matula, 2019).

Durante el desarrollo se identificó un conjunto de imágenes de micrografías limitadas, un objetivo futuro es aumentar la cantidad de datos con el fin de mejorar la capacidad de entrenamiento y predicción de los modelos (Tsaregorodtseva & Belagiannis, 2021), al igual que con la caracterización de nuevos escenarios en las imágenes y volumetría de los datos y también para evitar el procesamiento excesivo de emulsiones desde el laboratorio, lo que conlleva a prácticas de análisis inteligentes y sostenibles.

8. Referencias

Albert, A., & Méndes, E. (2014). Evaluación de una Emulsión de Aceite en Agua (O/W) con Surfactantes no Iónicos: Efecto del Coque de Petróleo. Universidad de Carabobo Facultad de Ingeniería Escuela de Ingeniería Química Trabajo Especial de Grado.

Allan, D., Caswell, T., Keim, N., Van der Wel, C., & Verweij, R. (2021). Zenodo. Obtenido de https://zenodo.org/record/4682814#.Y9fZW3bMLIV

Ametek Spectro Scientific. (s.f.). Obtenido de https://www.spectrosci.com/knowledge-center/test-parameters/measuring-water-in-oil#:~:text=Pure%20water%20absorbs%20infrared%20light,environment%20around%20that%20water%20molecule.

Aranberri, I., Binks, B., Clint, J., & Fletcher, P. (2006). Elaboracion y Caracterización de Emulsiones Estabilizadas por Polimeros y Agentes Tensioactivos. Revista Iberoamericana de Polímeros Volumen 7(3).

Araujo, A. M., Santos, M., L., Fortuny, & Montserrat. (2008). Evaluation of Water Content and Average Droplet Size in Water-in-Crude Oil Emulsions by Means of Near-Infrared Spectroscopy. Energy & Fuels, 3450–3458.

Bagnato, J. (24 de Diciembre de 2020). Aprendemachinelearning. Obtenido de Aprendemachinelearning: https://www.aprendemachinelearning.com/aprendizaje-por-refuerzo/

Bagui, S., & Mink, D. (2022). Detecting Reconnaissance and Discovery Tactics from the MITRE ATT&CK Framework in Zeek Conn Logs Using Spark's Machine Learning in the Big Data Framework. mdpi. Obtenido de https://en.wikipedia.org/wiki/Precision_and_recall#cite_note-OlsonDelen-24

Balabin, R., Lomakina, E., & Safieva, R. (2011). Neural Network (ANN) Approach to Biodiesel Analysis: Analysis of Biodiesel Density, Kinematic Viscosity, Methanol and Water Contents Using Near Infrared (NIR) Spectroscopy. Fuel, 2007-2015.

Bampi, M., P. Scheer, A., & Castilhos, F. (2013). Application of Near Infrared Spectroscopy to Predict the Average Droplet Size and Water Content in Biodiesel Emulsions. Fuel, 546-552.

Blanco, M., & Peguero, A. (2008). An Expeditious Method for Determining Particle Size Distribution by Near Infrared Spectroscopy: Comparison of PLS2 and ANN Models. Talanta, 647-651.

Brownlee, J. (2016). Machine Learning Mastery. Obtenido de A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning:

https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

Cañadas, R. (2021). Abdatum. Obtenido de https://abdatum.com/tecnologia/redes-neuronales-recurrentes

Cardellino, F. (2021). freecodecamp.org. Obtenido de https://www.freecodecamp.org/espanol/news/random-forest-classifier-tutorial-how-to-use-tree-based-algorithms-for-machine-learning/

Colucci, D., Morra, L., Zhang, X., Fissore, D., & Lamberti, F. (2020). An Automatic Computer Vision Pipeline for the in-Line Monitoring of Freeze-Drying Processes. Computers in Industry.

Crocker, J., & Grier, D. (1995). Methods of Digital Video Microscopy for Colloidal Studies. Journal of Colloid and Interface Science.

Di Sipio, R. (2021). Medium. Obtenido de https://towardsdatascience.com/a-quick-guide-to-auc-roc-in-machine-learning-models-f0aedb78fbad

Digital Guide Ionos. (2020). Ionos. Obtenido de https://www.ionos.es/digitalguide/online-marketing/marketing-para-motores-de-busqueda/que-es-una-neural-network/

Dunn, K. (2023). Process Improvement Using Data.

Fafalios, S., Charonyktakis, P., & Tsamardinos, I. (2020). Gradient Boosting Trees. Gnosis Data Analysis PC.

Forero, J. E., Ortíz, O. P., Nariño, F. A., Díaz, J., & Peña, H. (2008). Diseño y Desarrollo de un Tanque de Alta Eficiencia para Deshidratación de Crudo (I). CTF Cienc. Tecnol. Futuro Vol.3 No.4 Bucaramanga.

García, V., Mollineda, R., & Sánchez, J. (2009). Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions. 441-448.

Gershgorn, D. (2022). Quartz. Obtenido de Quartz: https://qz.com/1046350/the-quartz-guide-to-artificial-intelligence-what-is-it-why-is-it-important-and-should-we-be-afraid/

Gómez, E., Maška, M., Kotrbová, A., Pospíchalová, V., & Matula, P. (2019). Deep Learning Based Segmentation of Small Extracellular Vesicles in Transmission Electron Microscopy Images. Scientific Reports.

Gonzalez Rojas, V. M., Conde Arango, G., & Ochoa Muñoz, A. F. (2021). Análisis de Componentes Principales en Presencia de Datos Faltantes: El Principio de Datos Disponibles. Scientia et Technica Año XXVI, Vol. 26, No. 02.

Google. (2017). Research google. Obtenido de https://research.google.com/colaboratory/faq.html

Hale, J. (2019). Medium. Obtenido de Scale, Standardize, or Normalize with Scikit-Learn: https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02

Henríquez, C. (2009). W/O Emulsions : Formulation, Characterization and Destabilization.

Hung Nguyen, Q., Bang Ly, H., Al-Ansari, N., & Thai Pham, B. (2021). Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. Hindawi.

Jupyter, E. d. (2021). Jupyter. Obtenido de https://docs.jupyter.org/en/latest/

Kallevik, H., Brunsgaard Hansen, S., Sæther, Ø., Kvalheim, O., & Sjöblom, J. (2000). Crude Oil Model Emulsion Characterised by Means of Near Infrared Spectroscopy and Multivariate Techniques. Journal of Dispersion Science and Technology, 245-262.

Kokal, S. (2005). Crude Oil Emulsions: A State-Of-The-Art Review. SPE Production & Facilities, 5-13.

Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. (2021). Classification of Imbalanced Data:Review of Methods and Applications. IOP Conference Series: Materials Science and Engineering.

Langevin, D., Poteau, S., & Argillier, J. (2004). Crude Oil Emulsion Properties and Their Application to Heavy Oil Transportation. Oil & Gas Science and Technology-revue De L Institut Francais Du Petrole, 511-521.

Lendínez Gris, M. (2015). Estudio de Emulsiones Altamente Concentradas de Tipo W/O: Relación entre Tamaño de Gota y Propiedades. Consejo Superior de Investigaciones Cientificas.

Leoca, D. (2017). Trabajo de Fin de Grado Estudio de Simulación Estocástica para el Análisis de las Propiedades de los Estimadores Random Forests Frente a Árboles Individuales. Madrid: Industriales etsii upm.

Looker Studio. (2022). Google Cloud. Obtenido de https://cloud.google.com/looker-studio?hl=es-419

López, C. (2017). Análisis de Datos Categóricos Regresión Multinomial. Obtenido de http://tarwi.lamolina.edu.pe/~clopez/Categoricos/Regresion_Multinomial.pdf

Lun Chao, W. (2011). Machine Learning Tutorial. Taiwan.

Mahdi Jafari, S., & Bhandaric, B. (2008). Re-coalescence of Emulsion Droplets During High-Energy Emulsification. Food Hydrocolloids,Volume 22, Issue 7, Pages 1191-1202.

Maisa, W. G. (Junio de 2019). Análisis de Objetos Translúcidos Usando Técnicas de Visión por Computador. Universidad Técnica de Ambato. Obtenido de https://repositorio.uta.edu.ec/: chrome-

extension://efaidnbmnnnibpcajpcglclefindmkaj/https://repositorio.uta.edu.ec/bitstream/123456789/29955/1/Tesis_1604masc.PDF

Michael Polanyi, G. (s.f.). Facminas. Obtenido de https://facminas-unalmed.demo.elsevierpure.com/en/organisations/grupo-de-investigaci%C3%B3n-fen%C3%B3menos-de-superficie-michael-polanyi

Muhammad Shahani, N., & Muhammad, K. (2021). Application of Gradient Boosting Machine Learning Algorithms to Predict Uniaxial Compressive Strength of Soft Sedimentary Rocks at Thar Coalfield. Hindawi.

Neves Miranda, M. (2019). Particle Tracking in Microscopy Sequences for Microrheology Studies. Faculdade de Engenharia da Universidade do Porto.

Noboa, G., Márquez, L., & López, J. C. (2017). Tamaño de Gota: Factor Determinante sobre la Velocidad de Clarificación de una Emulsión O/W. Ciencia e Ingeniería, vol. 38, núm. 3.

Nolledo, M. (2020). Businessinsider. Obtenido de https://www.businessinsider.com/guides/tech/what-is-google-drive-guide#:~:text=Google%20Drive%20is%20a%20cloud,edit%20and%20collaborate%20on%20files.

Ocampo, M. (2018). Inteligencia Artificial. INCyTU.

Olalekan S, A., Mahmoud, M., Al Shehri, D., & Sultan, A. (2021). Rapid Determination of Emulsion Stability Using Turbidity Measurement Incorporating Artificial Neural Network (ANN): Experimental Validation Using Video/Optical Microscopy and Kinetic Modeling. American Chemical Society.

Omer, A., & Pal, R. (2010). Pipeline Flow Behavior of Water-in-Oil Emulsions with and without a Polymeric Additive in the Aqueous Phase. Chemical Engineering & Technology, 983-992.

Pájaro, M., & Álvarez, O. (2014). Análisis y Evaluación de la Deshidratación de Emulsiones Concentradas de Agua en Crudo Pesado Mediante Tratamiento Químico.

Pando Fernández, V., & San Martín Fernández, R. (2004). Regresión Logística Multinomial. Sociedad Española de Ciencias Forestales.

Pettersen, B., & Sjöblom, J. (2012). Flow Properties of Water-in-North Sea Heavy Crude Oil Emulsions. Journal of Petroleum Science and Engineering, 14-23.

Riaza, S., Cortés, F. B., & Otalvaro, J. (2014). Emulsiones con Crudo Pesado en Presencia de Nanopartículas. Boletín de Ciencias de la Tierra.

Ribeiro, M., Guimarães, M., Madureira, C., & Cruz Pinto, J. (2004). Non-Invasive System and Procedures for the Characterization. Chemical Engineering Journal, 173-182.

Rouhiainen, L. (2018). Inteligencia Artificial 101 Cosas que Debes Saber hoy Sobre Nuestro Futuro. Alienta Editorial.

Savin, T., & Doyle, P. (2005). Static and Dynamic Errors in Particle Tracking Microrheology. Biophysical Journal.

Schindelin, J. (2012). Fiji: An Open Source Platform for Biological Image Analysis. Focus on Bioimage Informat.

Seos-project.eu. (s.f.). Obtenido de https://seos-project.eu/marinepollution/marinepollution-c02-s14-p01.html

Servicios Científicos Técnicos, Universidad de Oviedo. (s.f.). Obtenido de sct.uniovi: https://www.sct.uniovi.es/unidades/analisis-biologico/microscopia/equipos

SPE INTERNATIONAL. (2015). Petrowiki Spe Org. Obtenido de https://petrowiki.spe.org/Oil_emulsions

Statistics, S. (22 de 03 de 2021). IBM. Obtenido de https://www.ibm.com/docs/es/spss-statistics/25.0.0?topic=regression-multinomial-logistic

TIBCO. (s.f.). Obtenido de https://www.tibco.com/reference-center/what-is-a-random-forest

Tolosa, L. (2016). Emulsiones Estabilizadas con Particulas (Emulsiones de Pickering). FIRP.

Toquica Cáceres, H. (2020). Estado de Arte: Visión de Máquina y Técnicas de Inteligencia Artificial para la Detección de Infracciones de Tránsito por Alta Velocidad. ResearchGate.

Trackpy Contributors. (2015). Trackpy Documentation.

Trackpy Contributors. (2021). Obtenido de http://soft-matter.github.io/trackpy/dev/tutorial/walkthrough.html

Tsaregorodtseva, A., & Belagiannis, V. (2021). ParticleAugment: Sampling-based Data Augmentation. Elsevier.

Unnikrishnan, S., Donovan, J., Macpherson, R., & Tormey, D. (2020). An Integrated Histogram-Based Vision and Machine-Learning Classification Model for Industrial Emulsion Processing. IEEE Transactions on Industrial Informatics, 5948-5955.

Utria Robinson, L. (2017). Caracterización Fisicoquímica y Evaluación de un Rompedor de Emulsiones Agua/Aceite para el Tratamiento Químico de Aplicación en Campo Petrolero.

Vinuesa, P. (2016). CCG-UNAM. Obtenido de Tema 8 - Correlación: Teoría y Práctica: https://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema8_correlacion.pdf

Wikipedia. (s.f.). Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Multinomial_logistic_regression

Wong, S., Lim, J., & Dol, S. (2015). Crude Oil Emulsion: A Review on Formation, Classification and Stability of Water-in-Oil Emulsions. Journal of Petroleum Science and Engineering, Pages 498-504.

Yudo Wardhono, E., Permana Pinem, M., Wahyudi, H., & Agustina, S. (2019). Calorimetry Technique for Observing the Evolution of Dispersed Droplets of Concentrated Water-in-Oil (W/O) Emulsion during Preparation, Storage and Destabilization. mdpi.