RECIBIDO/RECEIVED:03 JUNIO 2023 | ACEPTADO/ACCEPTED: 15 AGOSTO 2023 | PP. 41 - 60

Técnicas de machine learning aplicadas a la producción de bioetanol a partir de la caracterización de biomasa lignocelulósico (*Theobroma cacao L.*)

Machine learning techniques applied to the production of bioethanol from the characterization of lignocellulosic biomass (*Theobroma cacao L.*)

Greta Valeria Basauri Romero
Universidad de Lima
20160142@aloe.ulima.edu.pe
ORCID: 0000-0003-4492-9324

Melanie Evelyn Velarde Herrera
Universidad de Lima
20161515@aloe.ulima.edu.pe
ORCID: 00000-0003-0113-2921

Yvan Jesús García López
Universidad de Lima
ygarcia@ulima.edu.pe
ORCID: 0000-0001-9577-4188

Revista de Biotecnología Vol.1 Nº1

Versión electrónica
https://investigacion.utmachala.edu.ec/revistas/index.php/biotecnologia

RESUMEN

En los últimos años, la demanda de combustibles fósiles ha ido en aumento y esto ha generado una escasez en las reservas mundiales lo que limita el crecimiento económico; ejemplo de ello es Tocache, una de las provincias más aisladas y pobres del Perú. En este estudio se aborda el uso de residuos lignocelulósicos como la cáscara de cacao para generar biocombustible, cuyo objetivo es comparar los resultados experimentales con los obtenidos de la simulación. Asimismo, se realizaron los procesos de pretratamiento, hidrólisis enzimática y fermentación en la Universidad Técnica de Machala. El Machine Learning se realizó con el software Orange, el cual se basó en los datos y las variables experimentales halladas previamente. El mejor resultado corresponde al método Random Forest, con el que se obtuvo una precisión con el R2 (0.83). Por consiguiente, la glucosa predicha fue 1.04 g/L y la cantidad óptima de alcohol etílico fue 5.34 g / L. Los resultados demuestran que el alcohol etílico simulado se aproxima al hallado experimentalmente (7.1 g/L) y a otros estudios realizados previamente. Finalmente, el uso de Machine Learning es menos costoso y los resultados se pueden obtener en el menor tiempo posible en comparación con los procedimientos experimentales.

Palabras clave: Bioetanol, cáscara de cacao, fermentación, hidrólisis Enzimática, aprendizaje automático.

ABSTRACT

In recent years, the demand for fossil fuels has been increasing and this has generated a shortage in world reserves, which limits economic growth; An example of this is Tocache, one of the most isolated and poorest provinces in Peru. This study deals with the use of lignocellulosic residues such as cocoa husks to generate biofuel, whose objective is to compare the experimental results with those obtained from the simulation. Furthermore, the pretreatment, enzymatic hydrolysis and fermentation processes were carried out at the Technical University of Machala. The Machine Learning was carried out with the orange software, which was based on the data and the experimental variables previously found. The best result corresponds to the Random Forest method, with which a precision with R2 (0.83) was obtained. Therefore, the predicted glucose was 1.04 g/L, and the optimal amount of ethyl alcohol was 5.34 g/L. The results show that the simulated ethyl alcohol is close to that found experimentally (7.1 g/L) and to previous studies. Finally, the use of Machine Learning is less expensive, and the results can be obtained in the shortest possible time compared to experimental procedures.

Keywords: Bioethanol, cocoa pod husk, fermentation, enzymatic hydrolysis, machine learning.

INTRODUCCIÓN

Chohan et al. (2020) señala que “más del 80% de la energía global se produce utilizando combustibles fósiles” (p. 1031). Sin embargo, se ha generado un desabastecimiento de las reservas mundiales de los combustibles fósiles debido principalmente a la alta demanda de estos, causado por la expansión de la población y el uso de estos en los sectores residencial, industrial, de servicios y de calefacción (Marques et al.,2018).

En la literatura analizada, el desabastecimiento también se da por la falta de diversificación en otras fuentes de energía debido a los altos costos en producción, infraestructura subdesarrollada y falta de economías de escala en comercialización (Kochtcheeva,2016). En otro artículo, podemos encontrar que existe una desigualdad en la introducción de energías renovables debido a la escasez de políticas públicas y regulaciones para su implementación, además de la falta de apoyo financiero y fondos del gobierno (Ang et al.,2022).

Sobre la base explicada anteriormente, el Perú no es la excepción, ya que necesita un suministro mayor de energia para sostener el aumento de la demanda, debido a que según el Comité de Operación Económica del Sistema Interconectado Nacional (COES, 2019), se espera una tasa de crecimiento promedio de la demanda de energía de 6.4% con respecto al periodo 2019-2024. “La tasa de cobertura eléctrica global en el Perú es relativamente alta, no obstante, el abastecimiento de electricidad en las zonas rurales es una constante preocupación” (Falcón-Roque et al.,2017, p. 065903-2). Este es el caso de San Martín, que, según el Organismo Supervisor de la Inversión en Energía y Minería (Osinergmin, 2019), su consumo per cápita semestral de combustibles líquidos en 2019 fue de 0.57 barriles por habitante en comparación de los 4.03 barriles consumidos en Moquegua y Madre de Dios; siendo Tocache, provincia de San Martin, afectada por el déficit de combustibles.

Por todo lo anterior, en la Figura 1 muestra el árbol de problemas para resumir los aspectos más destacados del análisis realizado previamente.

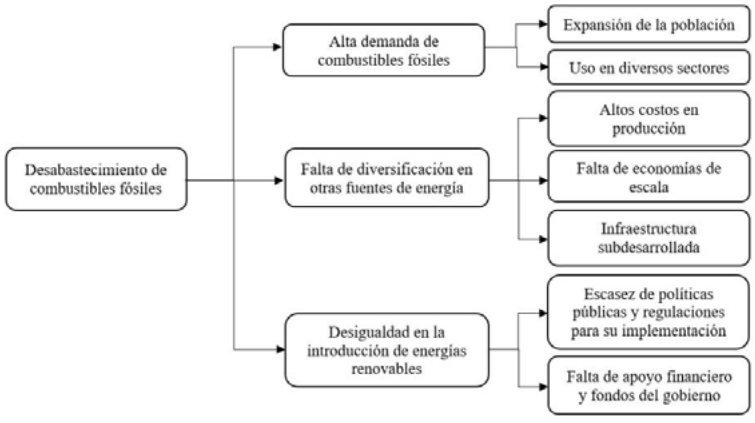

Figura 1. Árbol de Problemas

En orden de brindar una solución a este problema, en los últimos años, los biocombustibles se han vuelto una alternativa de gran interés para resolver este déficit de combustibles. Los principales sectores industriales en la producción de biocombustibles son el bioetanol y biodiesel, siendo Estados Unidos y Brasil los mayores productores a nivel global, gracias a que ambos han recibido un inmenso apoyo a través de medidas como exenciones fiscales y acuerdos regulatorios (Zentou et al.,2019). Asimismo, la producción de estos podría seguir aumentando con el paso de los años, ya que Torroba et.al (2022) señala que "en la última década (2012-2021) el consumo mundial de biocombustibles líquidos tuvo un crecimiento acumulado de 41%" (para.1).

En el Perú, para promover la demanda de producción y consumo de biocombustibles se emitió el Decreto Supremo 021-2007 que establece los parámetros necesarios de biocombustibles en la mezcla de gasolina, esta debe contener al menos 7.8% de bioetanol y por su parte, el combustible diésel un mínimo 5% de biodiésel (Ramírez-Triana,2017).

Banerjee y Tomar (2019) señalan que "la producción de bioetanol a nivel mundial puede reemplazar el 2% aproximadamente, del consumo total de gasolina" (p.509). Diversos países, como EE.UU., Brasil, China, Canadá y varios paises de la UE se han comprometido en programas de producción de bioetanol para resolver el gran problema de desabastecimiento que existe actualmente, aunque solo EE. UU y Brasil han mostrado mayores propuestas hasta el momento (Zabed et al.,2017).

De igual manera, en Perú la producción de bioetanol ha tenido variación positiva durante el periodo de 2012-2021 de 51% y en consumo, de 65%, teniendo como materia prima la caña de azúcar (Torroba et.al.,2022).

Este estudio se centrará en los biocombustibles de segunda generación. Estos son productos de la conversión de biomasa lignocelulósica a partir de los residuos de la industria agrícola, como es el caso de la cascarilla de cacao, abundante en Tocache, donde se produce la mayor cantidad de cacao en grano. Además, esta materia prima es una alternativa a la cáscara de yuca, la mazorca de maíz y el bagazo de caña de azúcar, que son materias primas que se suelen utilizar para producir bioetanol. "Debido principalmente a su abundancia, alta composición de carbohidratos y contenido mínimo de lignina" (Taherzadeh & Karim,2008, citado en Jugwanth et al.,2020, p.116553). Asimismo, la producción de bioetanol implica diferentes pasos del proceso que incluyen la caracterización de la biomasa, el pretratamiento de la biomasa, la hidrólisis y la fermentación (Jannah & Asip, 2015).

En ese sentido, se utilizará la cascarilla de cacao, "la cual contiene una mezcla de celulosa, hemicelulosa, lignina, pectina y fibra cruda y, por lo tanto, sirve como una fuente potencial de sustratos de biomasa para la producción bioquímica" (Adjin-Tetteh et al., 2018, p.305). Esta biomasa es compuesta principalmente por celulosa (16.9%), hemicelulosa (4%) y lignina (69%) (Shet et al., 2018). Por otra parte, según Laconi y Jayanegara (2015), la composición porcentual estuvo constituida por hemicelulosa (6%), celulosa (35.3%) y lignina (38.8%). Estos estudios obtuvieron valores muy diferentes debido

a que “esta variabilidad se relaciona con la variedad, el riego, fertilizantes utilizados, el transporte, almacenamiento, suelo y otros” (Han & Bao,2018, citado en Alvarez-Barreto et al.,2021, p.1490).

Para Wei et al. (2017, citado en Alvarez-Barreto et al.,2021) “esta biomasa lignocelulósica tiene una estructura compleja porque la celulosa está envuelta en una matriz de hemicelulosa que, a su vez, está rodeada por paredes de lignina” (p.1490). Por este motivo, la implementación de un pretratamiento de biomasa es muy importante, “ya que su tarea es destruir la matriz de lignina de celulosa para reducir la cristalinidad de la celulosa para que actúen enzimas y microorganismos” (Jaramillo & Sanchez, 2018, p.151). Existen pretratamientos físicos, químicos, biológicos y combinados, en varios estudios se realiza la molienda como pretratamiento mediante un mortero de madera o un molino para triturar la cascarilla de cacao y tamizarla, reduciendo la cristalinidad de la celulosa. Asimismo, los pretratamientos ácidos y alcalinos son los más comunes, debido a que se puede obtener una mayor concentración de celulosa (Akhtar et al.,2017).

Una vez que la biomasa sea pretratada debe pasar por un proceso de hidrólisis que permita la conversión de la celulosa en azúcares fermentables. En ese sentido, a lo largo de los años las enzimas celulolíticas han sido la principal área de interés para poder hidrolizar diversas biomasas. “El repertorio general de enzimas celulolíticas consistió en endo-glucanasa, exo-glucanasa y b-glucosidasa, generalmente actúan de manera colegiada para mejorar la hidrólisis de materiales celulósicos” (Akhtar et al.,2017, p.134).

En referencia a la fermentación, en la mayoría de los estudios se emplea el microorganismo Saccharomyces cerevisiae. En la actualidad, se están construyendo nuevas plantas para etanol de segunda generación; debido a esto, el uso del S. cerevisiae se maximizará ya que permite la fermentación de hidrolizados lignocelulósicos de residuos agrícolas y cultivos energéticos (Jansen et al., 2017).

En los últimos años se han utilizado métodos de aprendizaje automático para producir bioetanol a partir de biomasa, como redes neuronales artificiales y bosques aleatorios. Por ello, en nuestro estudio buscaremos obtener bioetanol a partir de cáscaras de cacao aplicando técnicas de Machine Learning.

Las redes neuronales artificiales han demostrado su eficacia como herramienta de modelado debido a las funciones de procesamiento de información que poseen; además de los bosques aleatorios que permiten manejar conjuntos de datos con una gran cantidad de variables predictoras y son fáciles de emplear debido a sus modelos simples (Speiser et al., 2019). Con la implementación de estos algoritmos, en los resultados obtenidos de la experimentación, será posible predecir la concentración de este biocombustible (Smuga-Kogut et al., 2021). Por lo tanto, el presente estudio tiene como objetivo realizar la comparación de los resultados experimentales junto con los obtenidos de la simulación con respecto al rendimiento de la producción de bioetanol a partir de la caracterización de la cáscara de cacao.

MATERIALES Y MÉTODOS

La metodología utilizada en el estudio fue cuantitativa, ya que se estudiaron y analizaron mediante procedimientos estadísticos datos numéricos recogidos de experimentos realizados en un laboratorio para la obtención de bioetanol. El proceso de investigación inició con la recolección de datos experimentales, donde se determinaron las variables de decisión del proceso de hidrólisis y fermentación alcohólica con base en la investigación de Romero et al., (2016), donde se realizó un estudio cinético de producción de bioetanol a partir de residuos agroindustriales.

Posterior a ello, se realizó una preparación de datos a través de la técnica de "Data augmentation" (aumento de datos) que se refiere a una técnica utilizada en el campo del machine learning y la inteligencia artificial para mejorar el rendimiento de un modelo mediante la generación de nuevas instancias de entrenamiento a partir de las instancias originales. Esta técnica es comúnmente empleada en problemas de clasificación de imágenes, aunque también puede aplicarse en otros tipos de datos (Shorten & Khoshgoftaar, 2019).

La idea detrás de la data augmentation es crear datos originales introduciendo pequeñas transformaciones mientras se mantiene la etiqueta o clase asociada. Al introducir estas variaciones, se le proporciona al modelo más diversidad y generalización, ayudando a prevenir el sobreajuste y mejorando la capacidad del modelo para reconocer patrones en datos nuevos.

Se preparo 956 experimentos de datos que a continuación se obtiene la siguiente Tabla 1.

Tabla 1. Análisis de la data experimental

	Glucosa [g/L]	ph	Brix	Sacarosa gr/100gr	Do[g/L]	DQO [g/L]
Exp	956	956	956	956	956	956
Media	0,85	4,75	1,30	1,30	14,56	0,61
Desy. Estandar	136,01	0,19	0,06	0,06	0,65	27,74
Min	0,61	4,33	1,20	1,20	13,46	0,56
25%	0,73	4,59	1,25	1,25	13,99	0,58
50%	0,86	4,76	1,30	1,30	14,55	0,61
75%	0.96	4,93	1,35	1,35	15,14	0,63
Max	1.09	5,08	1,40	1,40	15,70	0,65

La data augmentation es una estrategia valiosa para mejorar el rendimiento de los modelos de Machine Learning, especialmente cuando el conjunto de datos de entrenamiento es limitado.

Luego estos datos se introdujeron al software Orange para la simulación en los diferentes modelos de Machine Learning para conseguir la cantidad óptima de bioetanol.

Materia prima

Las cascarillas de cacao (Theobroma cacao L) fueron obtenidas en el mercado mayorista de frutas "Las flores" de la ciudad de Lima. Estas fueron desechadas en el descascarillado, provenientes de diversos productores de la provincia de Tocache, San Martín.

Pretratamiento, hidrólisis enzimática y fermentación alcohólica

Estos procedimientos se efectuaron en el laboratorio de la Universidad Técnica de Machala, Ecuador. En el pretratamiento, se utilizó un molino de laboratorio para reducir el tamaño de la cascarilla de cacao. Posteriormente, se realizó el método de campos eléctricos pulsados, donde se empleó corriente alterna para conducir electricidad al equipo que genera pulsos eléctricos mediante dos varillas de metal. Luego, se preparó un vaso de precipitado que tenía agua y cáscara de cacao; y se lo expuso a pulsos eléctricos por 1 hora a 60 V para hidrolizar la celulosa, la hemicelulosa y lignina; y así esté libre para convertirla en glucosa. La hidrólisis enzimática se realizó en biorreactores de 2 a 50 g/L de polvo de cascarilla de cacao; después se añadieron 0.2 ml de la enzima beta-glucosidasa, previamente elegida para romper los enlaces glucosídicos, luego se agitó durante 1 hora a 120 rpm y para finalizar se tomó una muestra. En este proceso se midieron nuestras variables de decisión, que son pH, °Bx (grados Brix), oxígeno disuelto y demanda química de oxígeno. Para la cuantificación de la glucosa se empleó el método del ácido 3,5-dinitro-salicílico (DNS), utilizando un colorímetro Hach DR / 870 para generar una curva de calibración con soluciones de concentraciones de glucosa (Romero Bonilla, et al., 2016). Posteriormente, la fermentación alcohólica se realizó con 100 ml de solución, donde se controló el pH y se agregó 800 ufc/mL de levadura seca comercial (Saccharomyces cerevisiae), previamente activada. Se cuantificó la concentración de etanol mediante cromatografía de gases a través de muestras del jarabe glucosado. Cabe recalcar que las variables de decisión también se controlaron en la fermentación.

Machine learning

Machine learning o aprendizaje automático es una disciplina compuesta por un conjunto de técnicas que permiten a los computadores aprender desde los datos, generando predicciones a partir de algoritmos, adquirir conocimiento sobre un dominio y facilitando la toma de futuras decisiones (Gramajo et al., 2020). En el proceso de aprendizaje, los algoritmos pueden clasificarse en cuatro categorías: aprendizaje supervisado, no supervisado, semi supervisado y por refuerzo. Esto se muestra en la Figura 2.

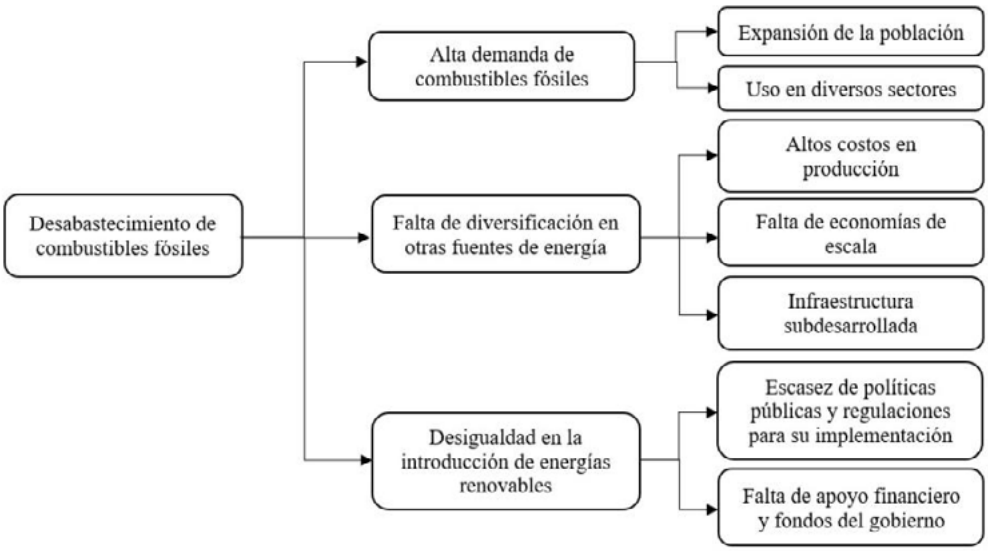

Figura 2. Clasificación de algoritmos de aprendizaje automático

Nota. Adaptada de "Seizing requirements engineering issues through supervised learning techniques" por M. G. Gramajo, 2020, IEEE Latin America Transactions, 18(7), p.1164–1184 (https://latamt.ieeer9.org/index.php/transactions/article/view/54)

En el caso del aprendizaje sin supervisión, los algoritmos cuentan con datos con estructura desconocida y sin etiquetado, con el objetivo de reestructurar los datos de entrada con nuevas funciones. Por parte de la semi supervisada, tiene la capacidad de usar datos sin etiqueta para mejorar la funcionalidad del aprendizaje supervisado. Mientras que, "el aprendizaje por refuerzo se basa en la retroalimentación obtenida del entorno producto de distintas interacciones, siendo su objetivo mejorar su desempeño a base de recompensas" (Gramajo et al., 2020, p. 1165).

Por último, el aprendizaje supervisado, predice los resultados futuros gracias a un entrenado modelo con entradas y salidas ya conocidos, que son monitoreados constantemente (Harippriya et al., 2022). A continuación, la Figura 3 muestra el diagrama de flujo, a partir de un modelo de datos dividido en capacitación y pruebas.

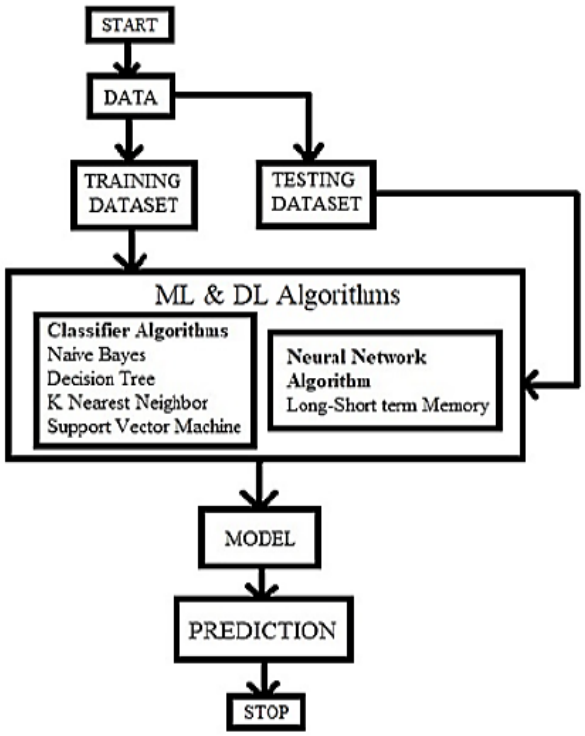

Figura 3. Diagrama de Flujo de un modelo de datos

Nota. Adaptada de "Battery management system to estimate battery aging using deep learning and machine learning algorithms", por S. Harippriya, E.E Vigneswaran & S.Jayanthy, 2022, Journal of Physics: Conference Series, 2325(1), p.012004 (https://doi:10.1088/1742-6596/2325/1/012004).

Como se muestra en la Figura 3 los datos son recolectados en los algoritmos de aprendizaje supervisado y no supervisado, estos se encuentran divididos en 80% y 20% de los datos de entrenamiento y prueba. Para finalmente obtener el modelo e iniciar el proceso predicción, en base a lo aprendido.

Para la creación del modelo de aprendizaje automático en su primera fase se usó el software Orange Data Mining que es un software libre para planificar como seria el modelo propuesto y posterior a ello en la segunda fase el modelo fue comprobado usando el lenguaje de programación Python (Romero, E., et al., 2023); los datos recolectados fueron obtenidos en la prueba piloto de hidrolisis enzimática del laboratorio de la Universidad Técnica de Machala, Ecuador. A continuación, en la figura 4a y 4b se describe las variables numéricas que fueron evaluadas y su comportamiento dentro de una distribución normal

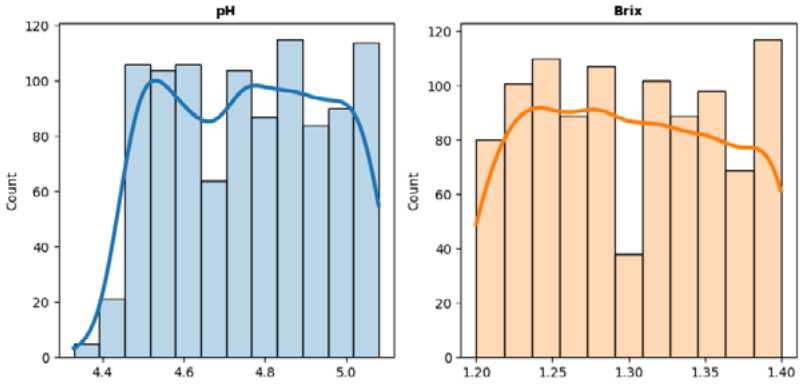

Figura 4a. Diagrama de distribución normal de pH y Grados Brix

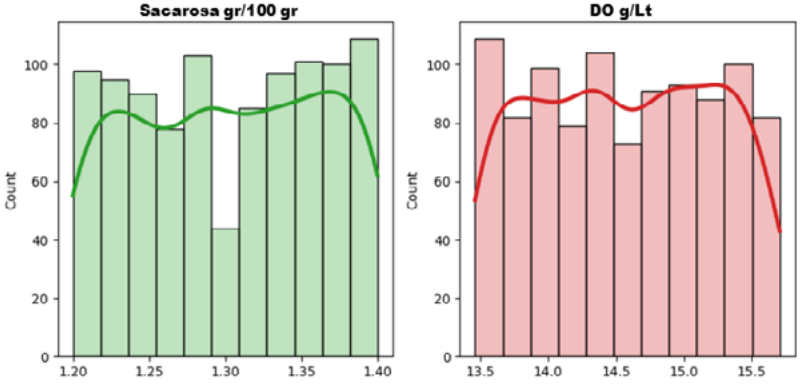

Figura 4b. Diagrama de distribución normal de sacarosa y Demanda de Oxígeno Disuelto

Nota. Extraído del modelo Python desarrollado en el análisis exploratorio de datos de la muestra experimental

Métodos de Machine Learning utilizados para predecir bioetanol

Random Forest

En el modelo basado en la metodología de los árboles de decisión se da un algoritmo entre N -el número de casos a prueba y M - el número de variables en el clasificador- asimismo las variables de entrada m al desarrollar un conjunto de entrenamiento por cada nodo del árbol, elige aleatoriamente en qué basar su decisión, lo cual calcula la mejor participación por conjunto desde estas variables (Vallejos-Romero et al, 2022).

Este método es derivado del árbol de clasificación y regresión que incluye resistencia al ruido y facilidad de ajuste, además se debe aplicar este método debido a que cada árbol es un proceso de aprendizaje, y eso le permite seleccionar muestras al azar, obteniendo un promedio de las predicciones de todos los árboles para la predicción final; evitando el sobreajuste y el efecto de muestras ruidosas (Xing et al., 2019).

Cabe recalcar que el modelo empleando Random Forest va a presentar siempre un buen rendimiento. El diagrama del algoritmo Random Forest es mostrado en la figura 5.

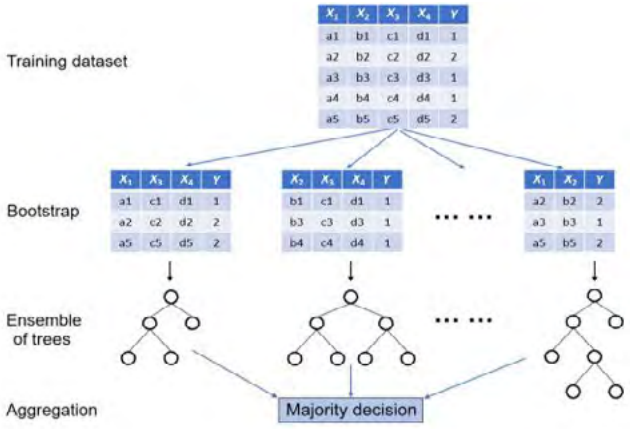

Figura 5. Modelo Random Forest

Nota. Adaptada de "Chapter 9 – Noninvasive fracture characterization based on the classification of sonic wave travel times", por Siddharth, M. y Hao L., 2020, Elsevier, 1, p. 243-287 (https://doi.org/10.1016/B978-0-12-817736-5.00009-0)

AdaBoost

El segundo método es bastante resistente al sobreajuste generado por datos experimentales inexactos, esto se debe a que en este método se construyen clasificadores posteriores, los cuales modifican instancias mal clasificadas por clasificadores anteriores; permitiendo que sus pesos aumenten en comparación con los clasificados correctamente, de modo que el nuevo clasificador se centre más en los ejemplos incorrectos (Sprenger et al., 2017). La figura 6 muestra el modelo Adaboost.

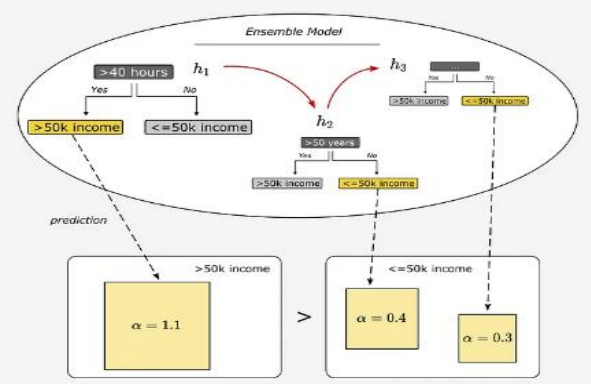

Figura 6. Modelo AdaBoost

Nota. Adaptada de "Chapter 9 – Noninvasive fracture characterization based on the classification of sonic wave travel times", por Siddharth, M. y Hao L., 2020, Elsevier, 1, p. 243-287 (https://doi.org/10.1016/B978-0-12-817736-5.00009-0)

Gradient Boosting

Finalmente, este es el método principal para el aprendizaje de problemas con características heterogéneas, datos ruidosos y dependencias complejas, ya que se compone de un conjunto de árboles de decisión individuales que están entrenados de forma secuencial, lo que permite que cada árbol nuevo pueda mejorar los errores de los árboles anteriores (Dorogush et al., 2018). Estos tres métodos fueron comparados estadísticamente para determinar el mejor método a emplear para predecir la cantidad de glucosa. La figura 7 muestra el modelo Gradiente Boosting.

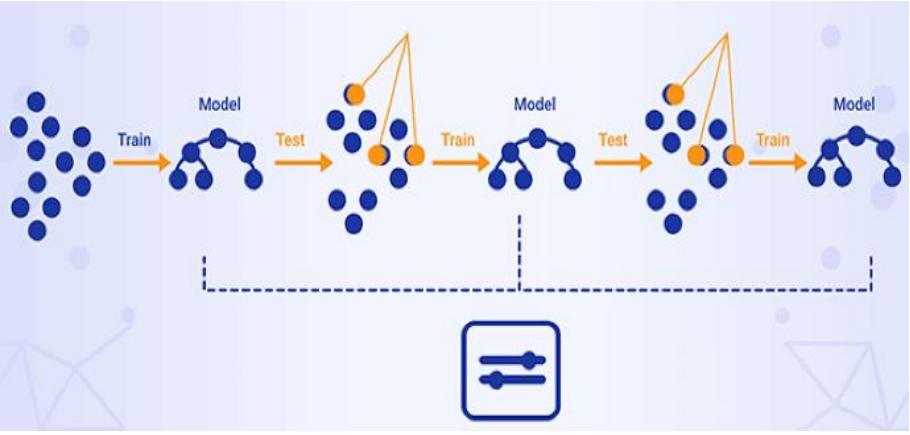

Figura 7. Modelo Gradient Boosting

Nota. Adaptada de: "Chapter 9 – Noninvasive fracture characterization based on the classification of sonic wave travel times", por Siddharth, M. y Hao L., 2020, Elsevier, 1, p. 243-287 (https://doi.org/10.1016/B978-0-12-817736-5.00009-0)

Resultados de la prueba piloto de producción de bioetanol

Se realizo las pruebas de laboratorio en la Universidad de Machala, Ecuador. Las siguientes variables fueron determinadas: el pH, oxígeno disuelto, la demanda química de oxígeno y el contenido de glucosa para producir bioetanol. Además, con ayuda de 0.5 de porcentaje residuo y 0.2 ml de concentración de la enzima usada en la hidrólisis enzimática se realizaron aproximadamente 956 experimentos en 5 meses, donde su análisis estadístico se muestra en la Tabla 2.

Tabla 2. Análisis estadístico

	Glucosa [g/L]	% Cacao	Enzima [,l]	pH	DO [g/L]	DO [g/L]
Media	0,85	0,5	0,2	4,76	14,57	0,61
Mínimo	0,61	0,5	0,2	4,33	13,46	0,56
Maximo	1,09	0,5	0,2	5,08	15,71	0,65

Con respecto a la cantidad de etanol a predecir, se halló empleando la Ecuación (2.4.1):

$$Etanol\ \left(\frac{g}{L}\right) = \frac{(X_i \times 92140)}{18015.6} \quad (2.4.1)$$

donde la cantidad de glucosa previamente calculada está representada por Xi donde i ∈ {1, . . ., 956}. Esta ecuación proviene de la reacción química, identificada como Ecuación (2.4.2) (Castro Giraldez et al., 2018).

$$C_6H_{12}O_6 \rightarrow 2CH_3CH_2OH + 2CO_2 \quad (2.4.2)$$

RESULTADOS Y DISCUSIÓN

En este estudio, se utilizó la cáscara de la mazorca de cacao para producir bioetanol y el proceso de hidrólisis enzimática pudo determinar un valor máximo de glucosa experimental y alcohol etílico (Tabla 3). Los 7.1 g/L de alcohol etílico son superiores a los reportados por Shet et al. (2018), quienes encontraron 2 g/L de alcohol etílico experimental, debido a que realizaron una hidrólisis ácida con ácido clorhídrico, lo que generó la formación de productos tóxicos que directamente afectó la fermentación. Esto se justifica ya que en la hidrólisis ácida se obtuvo un valor máximo de glucosa experimental de 2.37 g/L superior al de nuestro estudio; sin embargo, no se pudo obtener una concentración de alcohol etílico adecuada.

En los experimentos que utilizan aprendizaje automático, incluidos los tres métodos para estimar la glucosa predicha, además se obtiene una precisión con el R2 de determinación del modelo que representa la eficiencia de este y es el más alto conseguido hasta el momento, el cual se obtuvo con el método de Random Forest, como se muestra en la Tabla 4.

El valor de precisión del R2 de determinación fue de 0.83 encontrado en este estudio y es mayor al que fue encontrado por Awolu y Oyeyemi (2015), que busca predecir la cantidad de bioetanol a través de un modelo de predicción optimizado, utilizando la metodología de superficie de respuesta (RSM), que resultó un R2 de determinación de 0.5642. Este resultado confirma aún más la precisión del modelo Random Forest en comparación con otras técnicas.

Además, demuestra que el aprendizaje automático es una gran alternativa para la predicción de modelos tan complejos como el caso del bioetanol.

En el presente estudio se realizaron 956 experimentos donde se pudo obtener el error de raíz cuadrática media (RMSE) más bajo en el método Random Forest, permitiendo tener una mayor precisión en el modelo de predicción ya que la glucosa experimental mínima (0.61) no presenta una diferencia grande con respecto a la glucosa predicha mínima (0.64).

Este resultado permite confirmar los hallazgos en el estudio realizado por Speiser et al. (2019), que en similitud con nuestro estudio analiza conjuntos de datos con 956 observaciones a menos, concluyendo que el método de Random Forest tiende a tener tasas de error más bajas permitiendo tener un mejor ajuste en sus modelos.

En la Figura 8, se analiza la correlación entre las variables a través de Pearson, que miden el grado de relación lineal entre cada par de elementos o variables. Los valores de correlación pueden estar entre -1 y +1.

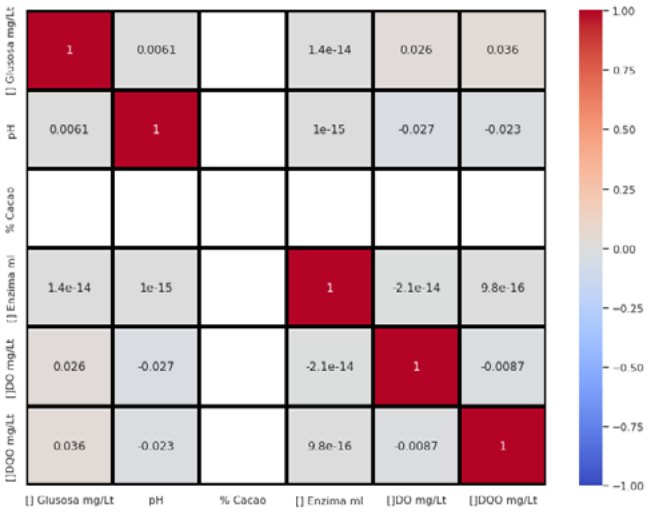

Figura 8. Matriz de correlación de variables

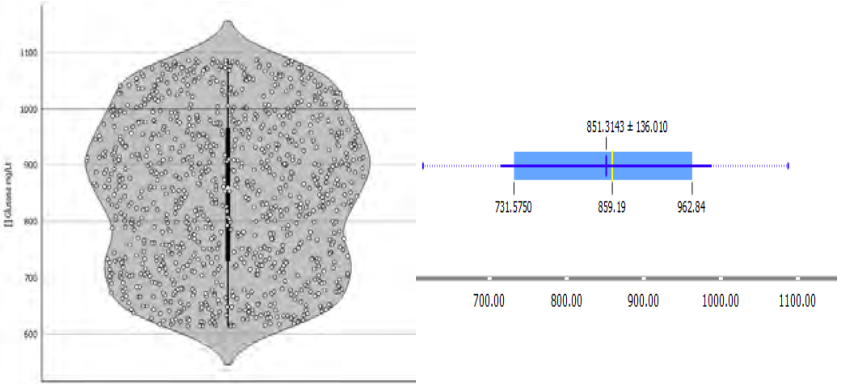

Figura 9. Gráfico de violín y Box-Plot para la concentración de glucosa.

En la figura 9, se muestra el diagrama violín para visualizar la distribución de los datos y su densidad de probabilidad con los datos preparados y suavizados para este análisis de datos. Este gráfico es una combinación de un box-plot junto con un gráfico de densidad girado y colocado a cada lado para mostrar la forma de la distribución de datos. Los valores de la glucosa están en el box-plot en el primer cuartil al 25% es de 731.57 mg/L, para la mediana

o segundo cuartil es 859.19 mg/L, en el tercer cuartil al 75% es de 962.84 mg/L. Se tiene una media de Glucosa de 851.31 mg/L y una desviación estándar de ± 136.

La barra negra gruesa en el centro representa el rango intercuartílico, la barra negra delgada que se extiende desde ella representa los intervalos de confianza del 95 % y el punto blanco en el centro es la mediana.

Además, la comparación de los tres métodos de aprendizaje automático se puede ver en la Tabla 3. En el modelo de aprendizaje automático Random Forest se observa una mayor precisión con un R cuadrado (0.83) con respecto a los demás. Por otro lado, se presenta una raíz del error cuadrático medio bastante bajo RMSE (55.06), esto indica que hay un mejor ajuste y que el modelo tiene una mayor precisión en la predicción de la concentración de glucosa.

Tabla 3. Comparación de los modelos de predicción de la concentración de glucosa

Modelo	Error cuadrático medio (MSE)	Error de raíz cuadrada media (RMSE)	Error absoluto medio (MAE)	R2
Random Forest	3032.11	55.06	44.83	0.830
AdaBoost	17519.62	132.36	114.16	0.027
Gradient Boosting	13002.57	114.03	96.79	0.278

Posteriormente se realiza la simulación para predecir la concentración de glucosa con el método Random Forest (Ecuación 2.4.1), obteniendo los resultados en la Tabla 4.

Tabla 4. Comparación del proceso experimental con la simulación

	Glucosa experimental [g/L]	Glucosa predicha [g/L]	Bioetanol experimental [g/L]	Bioetanol predicho [g/L]
Media	0,85	0,85	-	4,36
Mínimo	0,61	0,64	-	3,28
Maximo	1,09	1,04	7,10	5,34

A continuación, se presenta los datos reales con el modelo Random Forest ajustado.

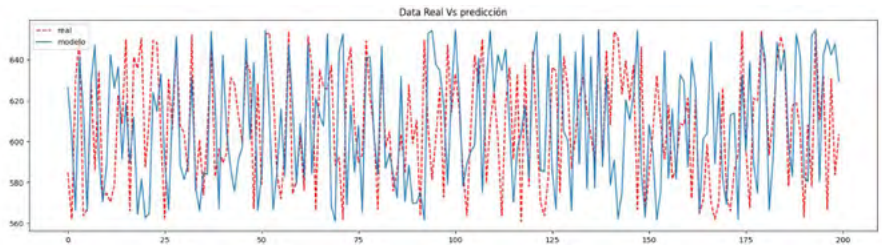

Figura 10. Modelo Random Forest para determinación de la Glucosa con R2:0.830

'

CONCLUSIONES

Los resultados de este estudio indican que el Machine Learning es una buena opción como herramienta para predecir el desempeño del etanol en un proceso de hidrólisis enzimática y fermentación a partir de la biomasa lignocelulósica de la cáscara de la mazorca de cacao, siendo un buen primer paso antes de pasar a gran escala, para lo cual se recomendaría un análisis de prefactibilidad y posteriormente un estudio de factibilidad si el proyecto es viable, para implementarlo en la provincia de Tocache. Además, el Machine Learning sigue demostrando que tiene un gran potencial como técnica de simulación porque se puede aplicar en diferentes campos, aportando precisión y calidad en sus modelos propuestos. Asimismo, su uso es menos costoso y los resultados se pueden obtener en el menor tiempo posible en comparación con los procedimientos experimentales. La calidad de este estudio indica que los resultados de investigaciones adicionales sobre la producción de bioetanol a partir de la cáscara de la mazorca de cacao se pueden utilizar para ampliar y aumentar continuamente la verificación del modelo propuesto.

AGRADECIMIENTOS

Este trabajo fue apoyado por Hugo Romero de la Universidad Técnica de Machala y Alexander Román, estudiante de Ingeniería de Sistemas de la Universidad de Lima.

REFERENCIAS BIBLIOGRÁFICAS

Adjin Tetteh, M., Asiedu, N., Dodoo Arhin, D., Karam, A., & Amaniampong, P. N. (2018). Thermochemical conversion and characterization of cocoa pod husks a potential agricultural waste from Ghana. *Industrial Crops and Products,* 119, 304-312. https://doi.org/10.1016/j.indcrop.2018.02.060.

Akhtar, N., Goyal, D., & Goyal, A. (2017). Characterization of microwave-alkali-acid pre-treated rice straw for optimization of ethanol production via simultaneous saccharification and fermentation (SSF). *Energy Conversion and Management,* 141, 133-144. https://doi.org/10.1016/j.enconman.2016.06.081.

Alvarez-Barreto, J. F., Larrea, F., Pinos, M. C., Benalcázar, J., Oña, D., Andino, C., ... & Almeida-Streitwieser, D. (2021). Chemical pretreatments on residual cocoa pod shell biomass for bioethanol production. Rev. Bionatura., 6, 1490-1500.

Ang, T. Z., Salem, M., Kamarol, M., Das, H. S., Nazari, M. A., & Prabaharan, N. (2022). A comprehensive study of renewable energy sources: Classifications, challenges and suggestions. *Energy Strategy Reviews,* 43, 100939. https://doi.org/10.1016/j.esr.2022.100939

Awolu, O., & Oyeyemi, S. O. (2015) Optimization of bioethanol production from cocoa (Theobroma cacao) bean shell. *International Journal of Current Microbiology and Applied Sciences, 4(4),* 506-514. https://www.ijcmas.com/vol-4-4/Awolu,%20OlugbengaOlufemi%20and%20Oyeyemi,%20Sanjo%20Oyetuji.pdf.

Banerjee, S., Kaushik, S., & Tomar, R. S. (2019). Global scenario of biofuel production: Past, present and future. *Prospects of Renewable Bioprocessing in Future Energy Systems,* 499-518. https://doi.org/10.1007/978-3-030-14463-0_18.

Bonilla, H. R., Balón, C. M., Moreno, A. P., & Pesantez, F. R. (2019). Estudio cinético de la producción de bioetanol a partir de residuos agroindustriales de la cáscara de banano maduro. *Industrial data,* 22(1), 187-202. https://doi.org/10.15381/idata.v22i1.16534.

Castro Giraldez, M., Tomás Egea, J. Á., Ortolá Ortolá, M., & Fito Suñer, P. J. (2018). Comparación de combustibles utilizados en quemadores industriales acoplados a un secador. *Universidad Politécnica de Valencia.* https://riunet.upv.es/bitstream/handle/10251/104147/Castro%3bTom%c3%a1s%3bOrtol%c3%a1%20-%20Comparaci%c3%b3n%20de%20combustibles%20utilizados%20en%20quemadores%20industriales%20acoplados....pdf?sequence=1&isAllowed=y.

Comité de Operación Económica del Sistema Interconectado Nacional (2019), *Informe de Diagnóstico de las Condiciones Operativas del SEIN, Periodo 2021-2030.* http://www.coes.org.pe/portal/browser/download?url=Planificación%2FPlan de 50 Transmision%2FActualización Plan de Transmisión 2021 - 2030%2F02. Informe de Diagnóstico 2021-2030%2F 01. Informe%2FInforme COES-DP 01-2019_COMPLETO.pdf.

Chohan, N. A., Aruwajoye, G. S., Sewsynker-Sukai, Y., & Kana, E. G. (2020). Valorisation of potato peel wastes for bioethanol production using simultaneous saccharification and fermentation: process optimization and kinetic assessment. *Renewable Energy,* 146, 1031-1040. https://doi.org/10.1016/j.renene.2019.07.042.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv,* 1810, 11363. https://arxiv.org/abs/1810.11363.

Falcón-Roque, E. J., Marcos Martín, F., Pascual Castaño, C., Domínguez-Dafauce, L. C., & Bastante Flores, F. J. (2017). Energy planning model with renewable energy using optimization multicriteria techniques for isolated rural communities: Cajamarca province, Peru. *Journal of Renewable and Sustainable Energy,* 9(6). https://doi.org/10.1063/1.4989574

Gramajo, M. G. (2020). Seizing Requirements Engineering Issues through Machine learning: A Systematic Mapping Study. *IEEE Latin America Transactions,* 18(7), 1164–1184. https://latamt.ieeer9.org/index.php/transactions/article/view/54

Harippriya, S., Esakki Vigneswaran, E., & Jayanthy, S. (2022). Battery management system to estimate battery aging using deep learning and machine learning algorithms. *Journal of Physics. Conference Series,* 2325(1), 012004. https://doi.org/10.1088/1742-6596/2325/1/012004

Jannah, A. M., & Asip, F. (2015). Bioethanol production from coconut fiber using alkaline pretreatment and acid hydrolysis method. *International Journal on Advanced Science, Engineering and Information Technology,* 5(5), 320-322. https://core.ac.uk/reader/296919080.

Jansen, M. L., Bracher, J. M., Papapetridis, I., Verhoeven, M. D., de Bruijn, H., de Waal, P. P., ... & Pronk, J. T. (2017). Saccharomyces cerevisiae strains for second-generation ethanol production: from academic exploration to industrial implementation. *FEMS yeast research,* 17(5), fox044.

Jaramillo, I., & Sanchez, A. (2018). Identification of Mass Flow Dynamics in a Pretreatment Continuous Tubular Reactor. Computer Aided Chemical Engineering. *Elsevier,* 43, 151-156. https://doi.org/10.1016/B978-0-444-64235-6.50028-0.

Jugwanth, Y., Sewsynker, Y., & Kana E. G. (2020). Valorization of sugarcane bagasse for bioethanol production through simultaneous saccharification and fermentation: Optimization and kinetic studies. *Fuel,* 262, 116552. https://doi.org/10.1016/j.fuel.2019.116552.

Kochtcheeva, L. V. (2016). Renewable energy: global challenges. *E-International Relations,* 27. https://www.e-ir.info/2016/05/27/renewable-energy-global-challenges/.

Laconi, E. B., & Jayanegara, A. (2015). Improving nutritional quality of cocoa pod (Theobroma cacao) through chemical and biological treatments for ruminant feeding: in vitro and in vivo evaluation. *Asian-Australasian journal of animal sciences,* 28(3), 343-350. https://doi.org/10.5713/ajas.13.0798.

Marques, A. C., Fuinhas, J. A., & Pereira, D. A. (2018). Have fossil fuels been substituted by renewables? An empirical assessment for 10 European countries. *Energy policy,* 116, 257-265. https://doi.org/10.1016/j.enpol.2018.02.021.

Osinergmin. (2019). *Reporte Semestral de Monitoreo del Mercado de Hidrocarburos (informe nro.14).*https://www.osinergmin.gob.pe/seccion/centro_documental/Institucional/Estudios_Economicos/Reportes_de_Mercado/Osinergmin-RSMMH-I-2019.pdf

Ramírez-Triana, C. A. (2017). Biofuels in the world and the Latin America (LAC) region. *Catálogo editorial,* 61-88. https://journal.poligran.edu.co/index.php/libros/article/view/1991/1909.

Romero, E., Garcia-Lopez, Y. (2023). Techniques of Machine learning applied to reduce employee turnover in a company cleaning and disinfection. Proceedings of the 3rd Indian International Conference on Industrial Engineering and Operations Management New Delhi, India, November 2-4, 2023. ISSN / E-ISSN: 2169-8767. https://index.ieomsociecity.org

Shet, V. B., Bhat, M., Naik, M., Mascarenhas, L. N., Goveas, L. C., Rao, C. V., ... & Aparna, A. (2018). Acid hydrolysis optimization of cocoa pod shell using response surface methodology approach toward ethanol production. *Agriculture and Natural Resources,* 52(6), 581-587.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data,* 6(1), 1-48. https://doi.org/10.1186/s40537-019-0197-0.

Siddharth, M., & Hao, L. (2020). Noninvasive fracture characterization based on the classification of sonic wave travel times. *Elsevier,* 243 - 287. https://doi.org/10.1016/B978-0-12-817736-5.00009-0

Smuga-Kogut, M., Kogut, T., Markiewicz, R., & Słowik, A. (2021). Use of Machine Learning Methods for Predicting Amount of Bioethanol Obtained from Lignocellulosic Biomass with the Use of Ionic Liquids for Pretreatment. *Energies,* 14(1), 243. https://doi.org/10.3390/en14010243.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications,* 134, 93-101. https://doi.org/10.1016/j.eswa.2019.05.028.

Sprenger, M., Schemm, S., Oechslin, R., & Jenkner, J. (2017). Nowcasting foehn wind events using the adaboost machine learning algorithm. *Weather and Forecasting,* 32(3), 1079-1099. https://doi.org/10.1175/WAF-D-16-0208.1.

Torroba, A., Orozco Montoya, R. A., & Productivo, D. (2022). *Atlas de los biocombustibles líquidos 2021-2022.* (2ª ed). Instituto Interamericano de Cooperación para la Agricultura. https://repositorio.iica.int/handle/11324/21328.

Torroba, A., & Brenes Porras, C. (2022). *Estado de los biocombustibles líquidos en las Américas.* Instituto Interamericano de Cooperación para la Agricultura. https://repositorio.iica.int/handle/11324/21279.

Vallejos-Romero, D., Deudor-Fernandez, C., Garcia-Lopez, Y. (2022). Use of machine learning to predict profit in LPG distribution in metropolitan Lima. *DYNA New Technologies, Jan-Dec.* 2022, vol. 9, no. 1, [9P.]

Xing, J., Wang, H., Luo, K., Wang, S., Bai, Y., & Fan, J. (2019). Predictive single-step kinetic model of biomass devolatilization for CFD applications: A comparison study of empirical correlations (EC), artificial neural networks (ANN) and random forest (RF). *Renewable Energy,* 136, 104-114. https://doi.org/10.1016/j.renene.2018.12.088.

Zabed, H., Sahu, J. N., Suely, A., Boyce, A. N., & Faruq, G. (2017). Bioethanol production from renewable sources: Current perspectives and technological progress. *Renewable and Sustainable Energy Reviews,* 71, 475-501. https://doi.org/10.1016/j.rser.2016.12.076.

Zentou, H., Rosli, N. S., Wen, C. H., Abdul Azeez, K., & Gomes, C. (2019). The viability of biofuels in developing countries: Successes, failures, and challenges. *Iranian Journal of Chemistry and Chemical Engineering,* 38(4), 173-182.https://www.ijcce.ac.ir/article_35879_a06ce3ef12267266e4812bbe1a4d7a4d.pdf.