Revista **U.D.C.A** Actualidad & Divulgación Científica
Julio-Diciembre 2022-Volumen 25 No. 2:e2334
ISSN: 2619-2551 en línea
ISSN: 0123-4226 impreso

Artículo Científico
http://doi.org/10.31910/rudca.v25.n2.2022.2334

Aplicaciones de aprendizaje automático para el análisis industrial de la provisión azucarera en Matanzas, Cuba

Machine learning application to industrial analysis of the sugar provision in Matanzas, Cuba

Yasmany García-López[1]*; Lourdes Yamen González-Sáez[2]; Juan Alfredo Cabrera-Hernández[3]

[1]Instituto de Investigaciones de la Caña de Azúcar. Matanzas, Cuba; e-mail: yasmanygarcia31@gmail.com

[2]Universidad de Matanzas, Departamento de Química e Ingeniería Química. Matanzas, Cuba; e-mail: lourdesgonzalez71@gmail.com

[3]Universidad de Matanzas, Observatorio Ambiental COSTATENAS. Matanzas, Cuba; e-mail: alfredojuan1956@gmail.com

*autor de correspondencia: yasmanygarcia31@gmail.com

Cómo citar: García-López, Y.; González-Sáez, L.Y.; Cabrera-Hernández, J.A. 2022. Aplicaciones de aprendizaje automático para el análisis industrial de la provisión azucarera en Matanzas, Cuba. Rev. U.D.C.A Act. & Div. Cient. 25(2):e2334. http://doi.org/10.31910/rudca.v25.n2.2022.2334

Publicación oficial de la Universidad de Ciencias Aplicadas y Ambientales U.D.C.A, Institución de Educación Superior Acreditada de Alta Calidad por el Ministerio de Educación Nacional.

Recibido: septiembre 8 de 2022 **Aceptado:** noviembre 24 de 2022 **Editado por:** Helber Adrián Arévalo Maldonado

RESUMEN

El análisis de los servicios ecosistémicos puede aportar conocimientos importantes sobre cómo se procesan y se obtienen los bienes del sistema agroindustrial azucarero. Para este trabajo, se recopilaron 346 datos del procesamiento industrial de la caña de azúcar en tres zafras, en la agroindustria del municipio Calimete, Provincia Matanzas (Cuba), con el objetivo de emplear algoritmos de aprendizaje automáticos, para predicciones relacionadas a datos biofísicos y económicos. Se analizaron siete predictores y mediante best subset selection, se identificó la combinación de rendimiento potencial en caña y pérdidas industriales totales, para predecir el servicio de provisión azucarera, mediante la regresión lineal múltiple. Se ajustó, también, un segundo modelo, que predice el efecto económico de las pérdidas industriales. En ambos modelos, se logró explicar por encima del 70 % de la variabilidad observada, en las variables dependientes, con un test F significativo (p-value: < 0,05), además de cumplirse con las condiciones de diagnóstico y validación.

Palabras clave: Industria; Predicción; Modelo; Provisión; Azúcar.

ABSTRACT

The analysis of ecosystem services can provide important insights into how goods are processed and obtained from the sugar agro-industrial system. For this work, 346 data were collected on the industrial processing of sugarcane in three harvest, in the agroindustry of the Calimete municipality, Matanzas Province (Cuba), with the objective to use the machine learning algorithm, to predict both, biophysical and economic data. Seven predictors were analyzed and by best subset selection, it was identified both the potential yield in sugarcane and the total industrial losses combination to predict the sugar provision service, by multiple linear regression. In addition, it was adjusted a second model to predict the economic effect of the industrial losses. In both models were able to explain over 70 % of the variability observed, in the dependent variables, with a significant F test (p-value: <0.05), also the diagnostic and validation conditions were met.

Keywords: Industry: Prediction; Model; Provision; Sugar.

INTRODUCCIÓN

El término "provisión" es una categoría dentro del enfoque de los servicios ecosistémicos y se refiere a los productos, como comida, agua, madera y fibra, que son obtenidos de los ecosistemas y agroecosistemas (Grunewald *et al.* 2015). Este enfoque es investigado por diferentes autores, como Gaba *et al.* (2015) y Waweru Wangai *et al.* (2016) y se está convirtiendo en la piedra angular del pensamiento contemporáneo sostenible (Bull *et al.* 2016). Su estudio, así como sus aplicaciones en la toma de decisiones, es un área creciente, con amplias perspectivas, para proveer soluciones viables a numerosos desafíos sociales y ambientales, como cambio climático, prevención de la desertificación y gestión del agua (Liquete *et al.* 2016).

El análisis de los servicios ecosistémicos es un enfoque hacia los beneficios producidos p or el capital natural y su relación con los manejos para potenciarlos. Este enfoque, comprende aquellos beneficios que se perciben directamente por sus precios comercializables, como el azúcar y los que soportan su generación, como pueden ser la formación de suelo y los ciclos de nutrientes.

Las producciones agrícolas son definidas como servicios de provisión (Gaba *et al.* 2015). En un contexto agroindustrial azucarero, el servicio de provisión puede ser entendido en términos de toneladas de caña producida en los campos o de azúcar posterior a un procesamiento industrial. La caña de azúcar es influenciada por múltiples factores de carácter natural; también necesita de un adecuado manejo de recursos, para llegar a producciones sostenibles (Bhatt, 2020), que pueden ser analizados mediante herramientas estadísticas. En el trabajo de Everingham *et al.* (2016), para una predicción del rendimiento de la caña de azúcar, se consideró a las variables basadas en simulación de biomasa, datos de precipitaciones, radiación y temperaturas máximas y mínimas. Para Kaup (2015), el rendimiento por hectárea depende fuertemente de la región de cultivo, debido a los manejos específicos de cada campo de caña de azúcar y su relación con el tipo de suelo y las características del cultivo (Pérez Iglesias *et al.* 2015).

Aunque mayores rendimientos de caña de azúcar pueden implicar mayor cantidad de azúcar final, el incremento por tonelada de caña molida, necesita, además, de mayor eficiencia industrial. Los procesos agrícolas e industriales promueven las transformaciones del ecosistema en campos manejados, para el incremento del servicio de provisión, pero considerar el enfoque de los servicios ecosistémicos en la de toma de decisión, requiere de metodologías robustas, que incluyan el mapeo de su presencia (Vang Rasmussen *et al.* 2016) y analicen colecciones de datos, para incluir aspectos ambientales, dentro del planeamiento económico local (Keith *et al.* 2016; Sunderland & Butterworth, 2016); sin embargo, identificar correctamente conflictos y potencialidades, estimar impactos simultáneos, no es una tarea simple (Villasante *et al.* 2016). El empleo de colecciones de datos y el uso de algoritmos de aprendizaje automáticos pueden ser de ayuda en el ajuste de modelos para el enfoque de los servicios ecosistémicos y la estimación de flujos asociados (Willcock *et al.* 2018). Los algoritmos de aprendizaje automáticos (del inglés, *machine learning algorithm*) es el sub-campo de la inteligencia artificial, diseñado para aplicar las técnicas estadísticas y aprender de los datos recopilados (Nwanganga & Chapple, 2020).

Dentro de un contexto agroindustrial azucarero, se pueden encontrar múltiples ejemplos de usos de algoritmos de aprendizaje automáticos y modelos, como son análisis de regresión lineal, para evaluar relaciones con el rendimiento del cultivo (Rahman & Robson, 2016); modelo de polinomio, para evaluar la relación de factores con la reducción de rendimiento en azúcar (Nashiruddin *et al.* 2020); las combinaciones de redes neuronales y algoritmos genéticos, para predecir características del jugo de la caña de azúcar (Tarafdar *et al.* 2020); los algoritmos de random forest, boosting y máquinas de soporte vectorial, para predecir el rendimiento agrícola (Natarajan *et al.* 2016; Hammer *et al.* 2019). Shahzad *et al.* (2017) correlacionan el recuperado azucarero con diferentes rasgos morfológicos de la caña de azúcar. También, se utilizan métodos *stepwise regression*, para la selección de variables significativas (Kumar Verma *et al.* 2020). La selección de variables, mediante criterios estadísticos, permite elegir el subgrupo de predictores y establecer una relación adecuada de bias-varianza, además de evitar sobre o bajo ajuste del modelo (Ramasubramanian & Singh, 2019; Zimmerman, 2020).

El análisis predictivo, mediante un modelo basado en la relación causa-efecto con uno o más predictores, permite predecir una variable objetivo en función de un conjunto de variables de entrada (Contreras Juárez *et al.* 2016; Andrade Saltos & Flores M., 2018); sin embargo, cada modelo tiene diferentes niveles de flexibilidad y de restricciones, así como facilidad o dificultad para su interpretación (James *et al.* 2013).

Además de esto, el análisis del servicio de provisión azucarera (SPA) en esta agroindustria, enfrenta grandes retos con respecto a su predicción y a las variabilidades espacio-temporales, donde influyen múltiples factores y se requiere la recopilación de distintos indicadores biofísicos y económicos. El SPA puede ser expresado en términos de rendimiento agrícola (t ha^{-1}), producción de azúcar (t), así como una combinación de las etapas agrícolas e industrial, o sea, cantidad de azúcar (kg) por toneladas de caña molida (kg $_{azúcar}$ t $_{caña}{}^{-1}$). Es de resaltar que cada indicador explica un comportamiento en un espacio-tiempo y como tal puede ser analizado y predicho, con las tecnologías y los conocimientos adecuados. Así, los algoritmos de aprendizaje y el ajuste de un modelo predictivo, podrían proporcionar importantes elementos sobre las relaciones con diferentes factores y cómo estos producen incrementos o no, de azúcar.

Por ello, el presente trabajo tiene la finalidad de comprobar si los datos de la agroindustria evaluada pueden ser analizados, mediante algoritmos de aprendizaje automáticos y servir para las predicciones asociadas al servicio de provisión azucarera, con valores biofísicos y económicos.

MATERIALES Y MÉTODOS

Descripción del estudio de caso. El estudio, se realizó en la agroindustria azucarera del municipio Calimete, provincia de Matanzas, entre las coordenadas 22º 25' 41" N a 22º 36' 22" N y 81º 11' 14" W a 80º 48' 25" W. La misma, se compone por un central azucarero, que permite el procesamiento de la caña de azúcar y la obtención de productos requeridos por la sociedad, así como la generación de residuos. El suministro de caña de azúcar (materia prima principal) procedió de unidades de producción agrícola, las cuales, se subdividen en bloques y estos, a su vez, en campos (unidad mínima de manejo). La concepción de unidades, bloques y campos es una estructura creada en la agroindustria azucarera, que establece límites en el manejo de recursos y la gestión humana, además de una interrelación con las estructuras, los procesos y las funciones ecosistémicas, en un espacio-tiempo determinado.

Análisis del servicio de provisión. Se recurrió a un indicador de SPA en kilogramos de azúcar por tonelada de caña molida ($kg_{azúcar}\ t_{caña}^{-1}$); de esta forma, el análisis se centró en la eficiencia vinculada al servicio de provisión azucarera y su relación con predictores importantes. Para ello, se confeccionó una base de datos, que contó con 340 días del registro histórico. Los resultados, se comprendieron entre las zafras de 2014, 2015 y 2020, de la agroindustria azucarera mencionada. Se utilizaron diferentes indicadores relacionados al procesamiento de la caña de azúcar, en función del objetivo de estudio. En correspondencia, se buscaron los elementos de entrada en la industria, como el rendimiento potencial, contenido en los tallos de caña de azúcar, Brix, Pol, fibra y materia extraña. También, se consideraron los indicadores relacionados al comportamiento dentro de la industria, como las pérdidas totales, la pureza de miel final y el recobrado, además del aprovechamiento de la norma potencial. En total, los indicadores elegidos dentro de la base de los registros históricos, fueron:

- Rendimiento potencial en caña [RPC) (%)]
- Aprovechamiento del RPC [A_RPC (%)]
- Pérdidas totales [PERD_T (%)]
- Aprovechamiento de la norma potencial [ANP (%)]
- Materia extraña (ME)
- Pureza de miel final [P_Miel_Final (%)]
- Recobrado [RECB (%)]

El uso de algoritmos de aprendizaje jugó un papel fundamental dentro del proceso de análisis, con la selección de los predictores importantes, que permitieron el ajuste de los modelos utilizados, con el cumplimiento de los requisitos necesarios de validación. Para estos análisis, se utilizó el lenguaje de programación (R), software, versión 3.6.1 (R Core Team, 2019).

Construcción del modelo de regresión lineal. Para los datos biofísicos, se utilizó un modelo de regresión lineal múltiple, extensión del modelo de regresión lineal simple a p, variables independientes, según la ecuación 1:

$$Y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + e \qquad \text{ecuación 1}$$

Donde,

Y: fue la variable respuesta o indicador específico de provisión azucarera

x1, x2, xp: fueron las variables independientes seleccionadas

β0, β1, β2, βp: fueron los coeficientes de regresión

e: los errores aleatorios con distribución normal, media cero y varianza σ^2

Para el análisis de la relación entre los predictores y la variable respuesta, se realizó una prueba de hipótesis, donde un *p-value* < valor calculado implicó el rechazo de la hipótesis nula, de que la pendiente es igual a cero (James *et al.* 2013; Carrasquilla-Batista *et al.* 2016). El modelo seleccionado requirió solo dos predictores, por lo que pudo ser representado, mediante un gráfico de superficie de respuesta.

Selección de predictores importantes. Las medidas de bondad de ajuste que se emplearon fueron: la suma de cuadrados de los residuos al (RSS) y el coeficiente de determinación (R^2) (ecuación 2), que describió la proporción de variabilidad observada en la variable dependiente (Y), explicada por el modelo respecto a la variabilidad total. Se consideró, además, el criterio Mallow (Cp) (ecuación 3); esencialmente, la estadística de Cp agrega una penalización 2d $\hat{\sigma}2$ al RSS (James *et al.* 2013).

$$R^2 = 1 - \frac{RSS}{RSS} \qquad \text{ecuación 2}$$

$$Cp = \frac{1}{n}\,(RSS + 2d\hat{\sigma}^2) \qquad \text{ecuación 3}$$

Donde: TSS es la suma de cuadrados totales.

Como R^2 aumenta con la inclusión más variables (p) (James *et al.* 2013), se calculó también el R^2 ajustado (ecuación 4):

$$R^2\,\text{ajustado} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} \qquad \text{ecuación 4}$$

Se utilizó el método de mejor modelo de selección (*best subset selection*). Según James *et al.* (2013), este sufre limitaciones computacionales, para cantidades de predictores mayores a 40; sin embargo, en este caso, la cantidad de predictores estuvo por debajo de ese valor y esquema general de la selección consistió en:

-Creación de un conjunto de modelos, todos los posibles (*best subset*), mediante diferentes combinaciones de los predictores disponibles.

-Para cada posible tamaño de modelo (1 predictor, 2 predictores...), se seleccionó el mejor, basado en el RSS de los datos de entrenamiento.

-Los modelos se compararon entre ellos, para identificar el mejor, con base en la estimación de diferentes criterios (Cp, R^2 ajustado).

Ajuste del modelo de regresión lineal. Se utilizó, para el ajuste del modelo, los criterios R^2 y Error estándar residual (RSE), de acuerdo con lo expuesto por James *et al.* (2013); mediante RSEm se midió la desviación promedio de los puntos estimados por el modelo, respecto a la recta de regresión (ecuación 5).

$$RSE = \sqrt{\frac{1}{n-p-1}}\ RSS \qquad \text{ecuación 5}$$

Grados de libertad (Gl) = número observaciones (n)-número predictores (p)-1.

Diagnóstico del modelo lineal. En el diagnóstico del modelo, se observó cómo se ajustó a los datos de entrenamiento y se comprobaron los principales criterios considerados por diferentes autores: linealidad, distribución normal de los residuos, varianza de residuos constante (homocedasticidad), valores atípicos y de alta influencia, independencia y factor de Inflación de la Varianza (James *et al.* 2013; Ramasubramanian & Singh, 2019).

Validación del modelo lineal. Con la validación cruzada, se estimó el error de predicción del modelo. Para ello, a la base de datos de 340 observaciones, se le realizó la extracción aleatoria de 200 datos de entrenamiento del modelo y el resto para la validación y estimación del error de predicción (en un set de validación simple). Al tratarse de una variable continua, se empleó el error cuadrático medio (MSE), que consistió en la división del RSS entre el número de observaciones (n) (ecuación 6):

$$MSE = \frac{RSS}{n} \quad \text{(Para datos de validación)} \qquad \text{ecuación 6}$$

Análisis de econometría. En el contexto de variabilidad temporal, se realizó un análisis de econometría, con la utilización de la ecuación 7, la que permitió determinar las diferencias entre potencial azucarero de la caña que es molida (P_{Ai}) (ecuación 8) y azúcar B-96 (P_{AR}) (ecuación 9).

$$DPR = (P_{Ai} - P_{AR}) \qquad \text{ecuación 7}$$

$$P_{A_i} = Cmi * \frac{RPC}{100} \qquad \text{ecuación 8}$$

$$P_{A_R} = Cmi * \frac{RB_{96}}{100} \qquad \text{ecuación 9}$$

Donde:

DPR: Diferencias entre potencial y real (t)

Cmi: Caña de azúcar molida ($t_{caña}$)

RPC: Rendimiento potencial en caña (%)

P_{Ai}: potencial azucarero de la caña molida ($t_{azúcar}$)

P_{AR}: azúcar B-96 obtenido ($t_{azúcar}$)

RB96: Rendimiento industrial en base 96 (%).

Las diferencias obtenidas entre potencial y real producido, se multiplicó por un precio (P) del azúcar, de 227,9 USD $t_{azúcar}^{-1}$ (Azcuba, 2020) y se dividió entre las toneladas de caña molida (Cmi) (ecuación 10).

$$DPRE = \frac{DPR*P}{Cmi} \qquad \text{ecuación 10}$$

A las diferencias entre potencial y real obtenido en términos económicos (DPRE), se les ajustó un modelo de regresión lineal, que las relacionó con las pérdidas industriales totales; se trabajó con 333 observaciones, de la base de datos. La diferencia respondió a valores eliminados por comportamiento, que ejercieron efectos negativos en el modelo. Al igual que en el caso biofísico, se utilizaron aleatoriamente 200 datos, para entrenamiento del modelo, mientras que el resto fue para la validación y la estimación del error de predicción, donde se siguieron los criterios que fueron expuestos previamente.

RESULTADOS Y DISCUSIÓN

Las variabilidades temporales presentes en el servicio de provisión azucarera son consistentes con un crecimiento de inicio a mediados y un decrecimiento hacia el final de zafra; sin embargo, al disminuir la escala de análisis, además de este comportamiento, existen variabilidades dentro de cada mes, como se puede apreciar en la zafra 2020 (Figura 1).

Las complejidades de la agroindustria azucarera están comprendidas en diferentes escalas espaciales y temporales. Así, el servicio de provisión azucarera en una zafra, no presenta un comportamiento homogéneo para cada uno de los días que la componen, como tampoco es similar en todos los campos de producción del cultivo de la caña de azúcar, debido a diferentes sitios y necesidades específicas de elementos esenciales. Las etapas de este sistema, agrícola e industrial, están sometidas a diversos impulsores de cambio, que fomentan beneficios o los perjudican y se pueden relacionar con el manejo del suelo, la calidad de la caña de azúcar y la máxima eficiencia industrial para procesarla.

Múltiples indicadores dentro del proceso industrial se pueden correlacionar con SPA; sin embargo, también se pueden correlacionar entre ellos, lo que origina información redundante en la construcción del modelo. Por lo que el mejor modelo no es el que posee más predictores, sino el que incluya los más representativos. Según el método de selección (*best subset selection*), los predictores de mayor significación, para el caso evaluado, fueron RPC y PERD_T. El RPC es resultado de factores de afuera de la industria, como la variedad, edad, época del año y condiciones del

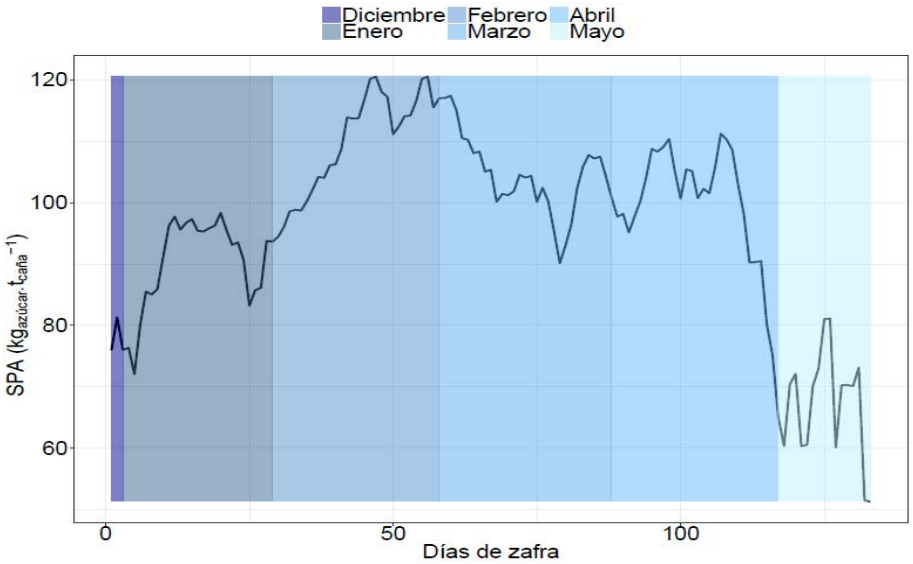

Figura 1. Variabilidad temporal del servicio de provisión azucarera.

Tabla 1. Coeficientes del modelo de regresión lineal múltiple.

	Estimación	Error	valor T	Pr(>|t|)	Sig.
Intercepto	60,30	3,66	16,37	$< 2e^{-16}$	***
RPC	6,64	0,30	22,99	$< 2e^{-16}$	***
PERD_T	-2,06	0,07	-26,09	$< 2e^{-16}$	***

Significancia: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1.
Error estándar residual: 3,5 con 197 grados de libertad; R^2 múltiple: 0,85; R^2 ajustado: 0,85; Estadígrafo F: 569,1 en 2 y p-valor: $< 2{,}2e^{-16}$.

cultivo (Martínez Pérez & De León Benítez, 2012), mientras que las pérdidas industriales, se ven influenciadas por factores dentro de la industria, como condiciones de la fábrica, eficiencia y disciplina del proceso industrial (Martínez Pérez & De León Benítez, 2012). Autores, como Roy & Chandra (2020), también destacan que el tiempo después de cosechado el cultivo puede incidir en bajas o altas cuantías de pérdidas de azúcar, mientras que Navarro Hernández & Rostgaard Beltrán (2014), explican que las materias extrañas pueden incrementar las pérdidas en bagazo, así como una disminución de su poder calórico, reducir la pureza del jugo mezclado, aumentar las pérdidas en la miel final y, de forma general, disminuir la producción de azúcar. En resumen, las condiciones de la caña de azúcar a la entrada del central y su procesamiento afectan el servicio de provisión azucarera. Por tal motivo, se concuerda con los indicadores RPC y PERD_T, como predictores en un modelo regresión lineal, para describir la variabilidad temporal del servicio de provisión azucarera. En la tabla 1, se recoge la significancia y los coeficientes del modelo. Los valores denotan una marcada influencia de los predictores seleccionados en el comportamiento de la variable dependiente.

Al ser dos predictores los seleccionados, el modelo se puede representar en un gráfico de superficie respuesta (Figura 2). El modelo toma la expresión de:

"SPA = 60,30 + 6,64 * RPC - 2,06 * PERD_T"

En la validación cruzada el cuadrado medio del error (MSE) obtenido fue de 14,4, mientras que en la figura 2b, se observa la relación entre los valores observados y los valores predichos del modelo con los datos de prueba, los cuales, se ajustan a una línea recta. Este modelo es capaz de explicar el 85 % de la variabilidad observada en el servicio de provisión azucarera (R^2 de 0,85). El test F muestra que es significativo (*p-value*: $< 2{,}2\ e^{-16}$); sin embargo, para que el modelo pueda ser utilizado deben cumplir, además, con otros criterios estadísticos. El diagnóstico del modelo es realizado mediante pruebas estadísticas y comportamientos gráficos. La prueba de *Lilliefors* arrojó un *p-value* de 0,62, mientras que la prueba de *Breusch-Pagan*, para residuos estandarizados, el *p-value* fue de 0,34. En la prueba de Durbin-Watson, no se encontraron evidencias de autocorrelación, donde el estadístico (d) fue de 2,11 y el *p-value*, de 0,36. También, se obtienen valores VIF (1,003), para ambos predictores, por debajo de 5, por lo que no existen problemas de colinealidad, ni presentan una inflación de varianza marcada, por lo que no se encuentran evidencias en el comportamiento de los residuos para rechazar el modelo. Además de lo expuesto, en la figura 3, se muestran los gráficos que corroboran los test estadísticos.

En la figura 3a, no observó ningún patrón, lo cual, es indicativo de que no existen heterocedasticidad; en la figura 3b, se diagnosticó la normalidad y los puntos se encontraron cerca de la diagonal y la figura 3c, también es indicativo para la heterocedasticidad y lo ideal es encontrar una pendiente nula (Ramasubramanian & Singh, 2019), mientras que la figura 3d permite detectar valores atípicos o influyentes. En este, se representó los residuos estandarizados en función del valor de influencia o *leverage*, así como la distancia

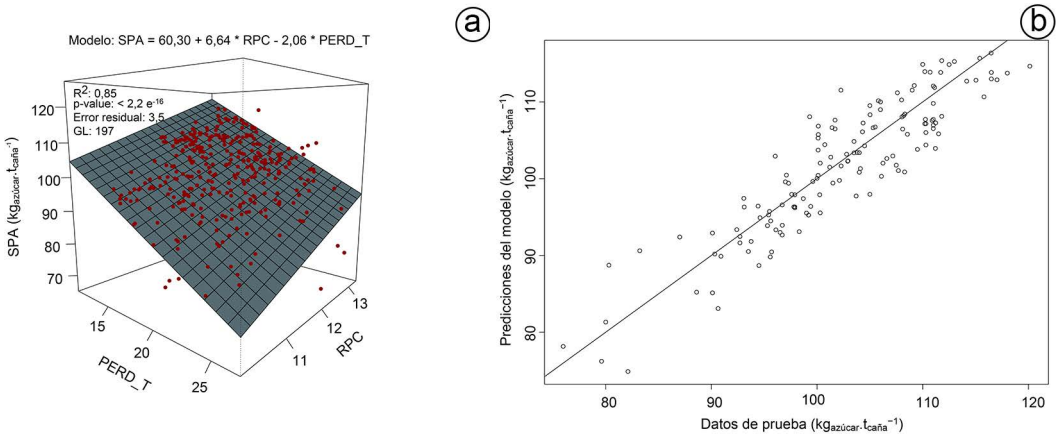

Figura 2. a) Superficie respuesta; b) Validación con datos de prueba.

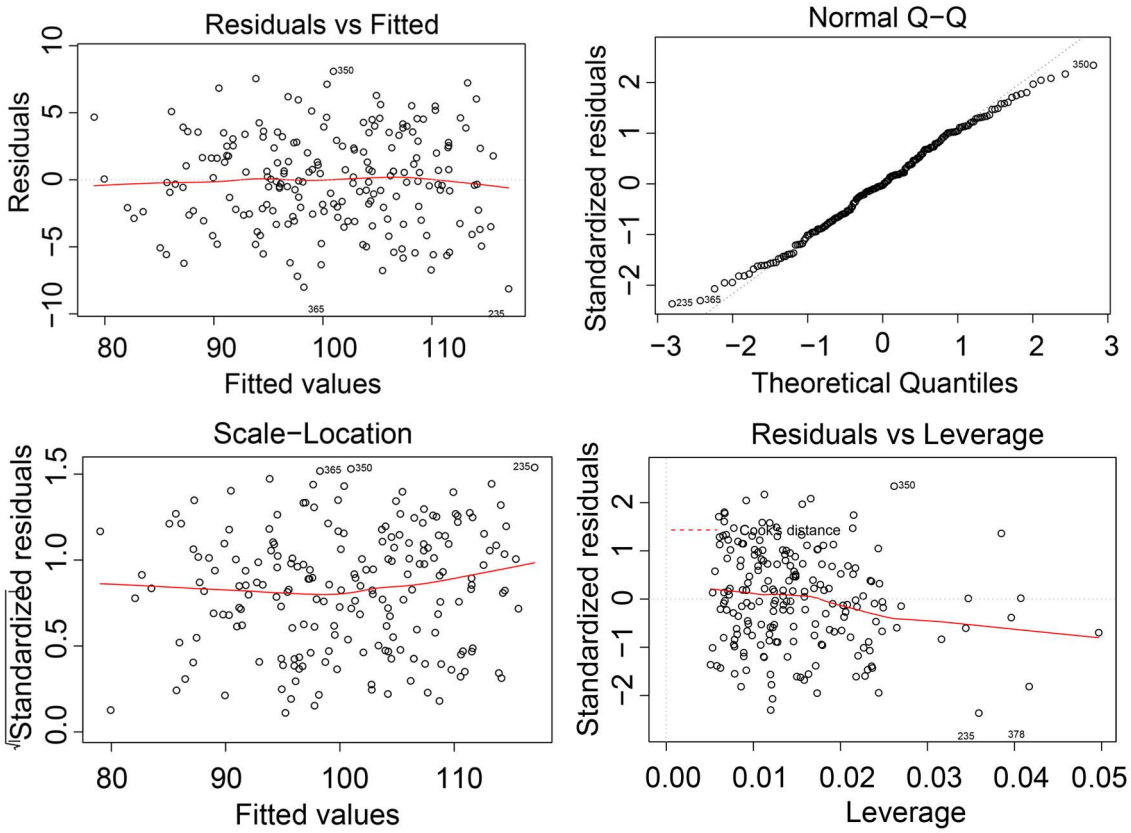

Figura 3. Diagnóstico del modelo de regresión lineal múltiple.

Cook para valores influyentes que, al no poseer valores elevados, no existen evidencias de un impacto regular en la línea de la regresión estimada (Nwanganga & Chapple, 2020). Se satisfacen, entonces, las condiciones de normalidad, de autocorrelación y no se presenta una inflación de varianza marcada para los predictores. Tampoco, se encontraron evidencias contra la homogeneidad de varianza, por lo que el servicio de provisión azucarera y su relación con los predictores: rendimiento potencial en caña (RPC) y pérdidas totales en la industria (PERD_T), puede ser modelada por un modelo de regresión lineal múltiple.

Resultados del análisis de econometría. Los análisis biofísicos y económicos pueden ser complementarios y permiten un mejor análisis y comprensión de los servicios ecosistémicos. La cuantificación de la producción real y su alejamiento de un potencial que puede ser alcanzado es una medida de pérdidas presentes en el proceso. Su expresión económica permite razonar la cantidad de beneficio dejado de percibir, además de visualizar un margen financiero que se puede tener, para invertir en acciones de mejoras y disminución de las pérdidas. Los resultados obtenidos en esta investigación permitieron representar, en un modelo lineal (Figura 4), un alejamiento del potencial azucarero, de 0,42 hasta 10 USD tc^{-1}, para los parámetros considerados.

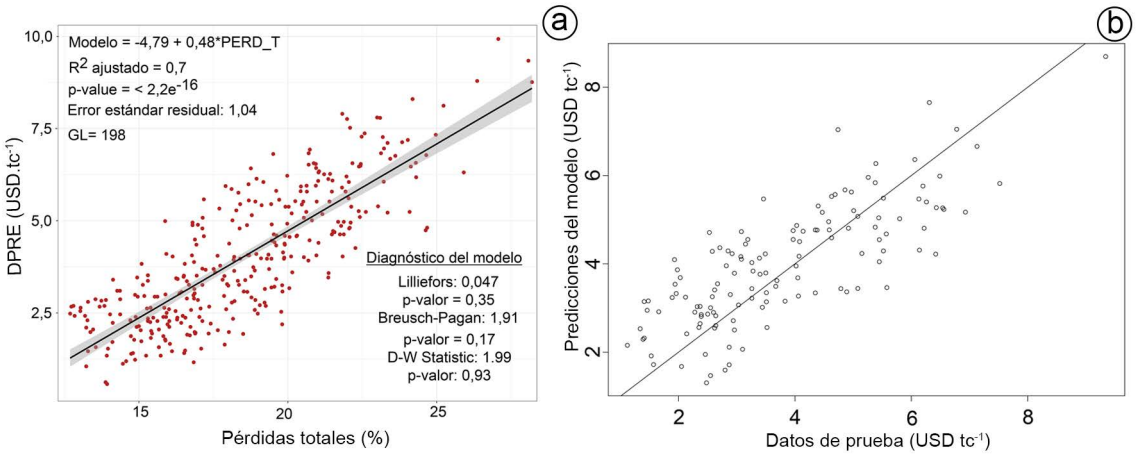

Figura 4. a) Modelo lineal; b) validación con datos de prueba.

Mediante el análisis combinado de parámetros estadísticos calculados y métodos gráficos (Figura 5), se considera que el modelo de regresión lineal simple cumple con los parámetros de diagnósticos y de validación, por lo que se considera adecuado para determinar las DPRE, en función de las pérdidas industriales totales, con un R^2, de 0,7 y un valor de la probabilidad menor que 0,05, mientras que la prueba de *Breusch-Pagan* obtuvo un valor de la probabilidad de 0,17. Además que, en la prueba de Durbin-Watson, no se encontraron evidencias de autocorrelación, donde el estadístico d fue de 1,99 y el *p-value*, de 0,93, además de un valor MSE, de 1,04, mientras que los métodos gráficos (Figura 5) corroboran los resultados, al igual que en la regresión lineal múltiple. Los resultados encontrados, además de considerar a los modelos válidos, denotan que las predicciones conllevan a un proceso complejo de ajuste, identifican errores y validación; sin obviar la incertidumbre, siempre tiene algún efecto en los métodos de análisis (Azadi *et al.* 2021) y que la incorporación de múltiples indicadores convierte al análisis de servicios ecosistémicos en un desafío (Smith *et al.* 2011).

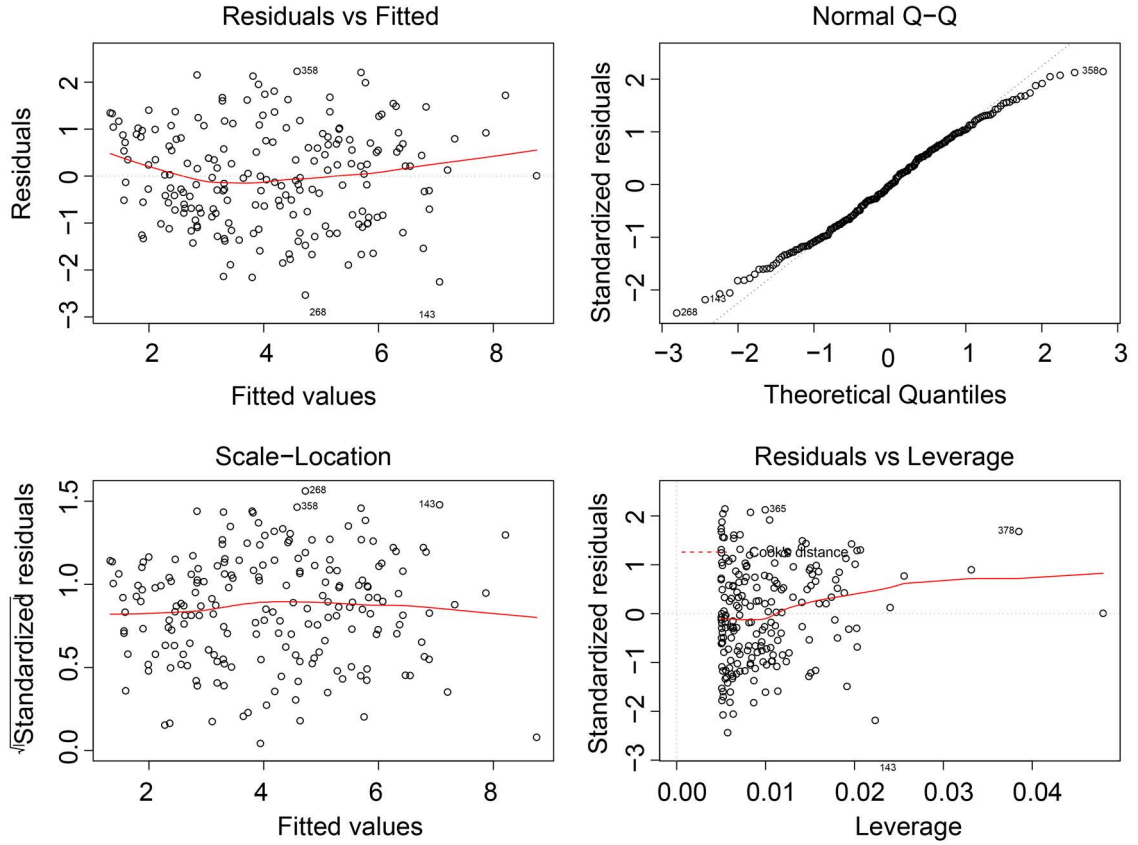

Figura 5. Diagnóstico del modelo de regresión lineal simple.

En función de los resultados del trabajo, se puede expresar que mayores o menores niveles de eficiencia en la agroindustria azucarera afectan de manera positiva o negativa a los servicios ecosistémicos. Entiéndase por eficiencia, la capacidad de acercar la producción real al potencial azucarero, que posee la caña de azúcar. Las pérdidas industriales conllevan a una menor cantidad de azúcar obtenida, lo que implica un menor beneficio económico por tonelada de caña molida. Por ello, un desafío de la agroindustria azucarera es lograr mayor eficiencia de su proceso e implementar estrategias para su análisis y predicción.

Los modelos de regresión lineal que han sido expuestos pueden contribuir a la comprensión de relaciones que se presentan, para el enfoque de servicios ecosistémicos, en un contexto agroindustrial azucarero y en concordancia con Ribas García *et al.* (2016), ser de utilidad en la planificación y en la optimización del uso de los recursos técnicos, humanos, financieros e indicar las variables tecnológicas de mayor peso.

Como conclusiones se pueden indicar, que los algoritmos de aprendizajes automáticos fueron útiles para el análisis del servicio de provisión azucarera y permitieron la identificación de predictores significativos y ajuste de modelos de regresión lineal, en un contexto agroindustrial azucarero.

El uso de valores biofísicos y económicos, de forma complementaria, permitieron una mejor comprensión de la manera en que se procesa y se obtiene el servicio de provisión azucarera.

Con la técnica de *best subset selection*, se identificaron a las pérdidas industriales y el RPC, como la mejor combinación de los predictores analizados, para modelar el servicio de provisión azucarera.

Se obtuvo un modelo de regresión lineal múltiple, para predecir el servicio de provisión azucarera, en función de las pérdidas industriales y el rendimiento potencial en caña, capaz de explicar el 85 % de la variabilidad observada y un test F significativo, menor a 0,05, además de cumplir con las diferentes condiciones de los residuos y la validación cruzada.

Mediante el modelo de regresión lineal simple, se determinó que el aumento de las pérdidas totales en la industria aleja el servicio de provisión azucarera, del potencial que posee el cultivo de la caña de azúcar, con valores entre 0,42 a 10 USD tc-1, para los parámetros considerados.

Conflictos de intereses: el manuscrito fue preparado y revisado con la participación de los autores y se declara que no existe conflicto de intereses que ponga en riesgo la validez de los resultados presentados. Financiación: la recopilación de datos y los análisis han estado comprendidas dentro de las etapas de trabajo del proyecto: "Tecnologías para adecuar los manejos de suelo y cultivo a la variabilidad del sistema agroindustrial azucarero", desarrollado por el Instituto de Investigaciones de la Caña de Azúcar en Matanzas, Cuba.

REFERENCIAS

1. ANDRADE SALTOS, V.A.; FLORES M., P. 2018. Comparativa tiva entre classification trees, random forest y gradient boosting, en la predicción de la satisfacción laboral en Ecuador. Ciencia Digital. 2(4.1):42-54. https://doi.org/10.33262/cienciadigital.v2i4.1..189

2. AZADI, H.; VAN PASSEL, S.; COOLS, J. 2021. Rapid economic valuation of ecosystem services in man and biosphere reserves in Africa: A review. Global Ecology and Conservation. 28:e01697. https://doi.org/10.1016/j.gecco.2021.e01697

3. AZCUBA. 2020. Evaluación diaria para dirigir económicamente. Empresa Azucarera, Matanzas. Informe No. 096. Matanzas, Cuba.

4. BHATT, R. 2020. Resources management for sustainable sugarcane production. In: Kumar, S.; Meena, R.S.; Jhariya, M.K. (eds.), Resources use efficiency in agriculture. Springer. Singapore. p.647-693. https://doi.org/10.1007/978-981-15-6953-1_18

5. BULL, J.W.; JOBSTVOGT, N.; BÖHNKE-HENRICHS, A.; MASCARENHAS, A.; SITAS, N.; BAULCOMB, C.; LAMBINI, C.K.; RAWLINS, M.; BARAL, H.; ZÄHRINGER, J.; CARTER-SILK, E.; BALZAN, M.V.; KENTER, J.O.; HÄYHÄ, T.; PETZ, K.; KOSS, R. 2016. Strengths, Weaknesses, Opportunities and Threats: A SWOT analysis of the ecosystem services framework. Ecosystem Services. 7:99-111. http://dx.doi.org/10.1016/j.ecoser.2015.11.012

6. CARRASQUILLA-BATISTA, A.; CHACÓN-RODRÍGUEZ, A.; NÚÑEZ-MONTERO, K.; GÓMEZ-ESPINOZA, O.; VALVERDE-CERDAS, J.; GUERRERO-BARRANTES, M. 2016. Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. Tecnología en Marcha. Encuentro de Investigación y Extensión. 33-45. https://doi.org/10.18845/tm.v29i8.2983

7. CONTRERAS JUÁREZ, A.; ATZIRY ZUÑIGA, C.; MARTÍNEZ FLORES, J.L.; SÁNCHEZ PARTIDA, D. 2016. Análisis de series de tiempo en el pronóstico de la demanda de almacenamiento de productos perecederos. Estudios Gerenciales. 32(141):387-396. http://dx.doi.org/10.1016/j.estger.2016.11.002

8. EVERINGHAM, Y.; SEXTON, J.; SKOCAJ, D.; INMANBAMBER, G. 2016. Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for Sustainable Development. 36(27):1-9. https://doi.org/10.1007/s13593-016-0364-z

9. GABA, S.; LESCOURRET, F.; BOUDSOCQ, S.; ENJALBERT, J.; HINSINGER, P.; JOURNET, E.-P.; NAVAS, M.-L.; WERY, J.; LOUARN, G.; MALÉZIEUX, E.; PELZER, E.; PRUDENT, M.; OZIER-LAFONTAINE, H. 2015. Multiple cropping systems as drivers for providing multiple ecosystem services: from concepts to design. Agronomy for Sustainable Development. 35:607-623. https://doi.org/10.1007/s13593-014-0272-z

10. GRUNEWALD, K.; BASTIAN, O.; MANNSFELD, K. 2015. Development and Fundamentals of the ES Approach. En: Grunewald, K.; Bastian, O. (eds.) Ecosystem Services - concept, methods and case studies. Springer (Germany). p.13-34. https://doi.org/10.1007/978-3-662-44143-5_2

11. HAMMER, R.G.; SENTELHAS, P.C.; MARIANO, J.C.Q. 2019. Sugarcane yield prediction through data mining and crop simulation models. Sugar Tech. 22:216-225. https://doi.org/10.1007/s12355-019-00776-z

12. JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. 2013. An Introduction to Statistical. With Applications in R. Springer (New York). 426p. https://doi.org/10.1007/978-1-4614-7138-7

13. KAUP, F. 2015. The sugarcane complex in Brazil. The role of innovation in a dynamic sector on its path towards sustainability. Contributions to Economics. Springer (Switzerland). 280p. https://doi.org/10.1007/978-3-319-16583-7

14. KEITH, A.M.; SCHMIDT, O.; MCMAHON, B.J. 2016. Soil stewardship as a nexus between Ecosystem Services and One Health. Ecosystem Services. 17:40-42. http://dx.doi.org/10.1016/j.ecoser.2015.11.008

15. KUMAR VERMA, A.; KUMAR GARG, P.; HARI PRASAD, K. S.; KUMAR DADHWAL, V.; KUMAR DUBEY, S.; KUMAR, A. 2020. Sugarcane yield forecasting model based on weather parameters. Sugar Tech. 23:158-166. https://doi.org/10.1007/s12355-020-00900-4

16. LIQUETE, C.; UDIAS, A.; CONTE, G.; GRIZZETTI, B.; MASI, F. 2016. Integrated valuation of a nature-based solution for water pollution control. Highlighting hidden benefits. Ecosystem Services. 22:392-401. http://dx.doi.org/10.1016/j.ecoser.2016.09.011

17. MARTÍNEZ PÉREZ, C.M.; DE LEÓN BENÍTEZ, J.B. 2012. Influencia de la calidad de la materia prima en el proceso tecnológico, calidad del producto final, y el rendimiento industrial en una fábrica de azúcar. Revista Centro Azúcar. 39(3):28-34.

18. NASHIRUDDIN, N.I.; FADZIYANA MANSOR, A.; RAHMAN, R.A.; ILIAS, R.M.D.; WAN YUSSOF, H. 2020. Process parameter optimization of pretreated pineapple leaves fiber for enhancement of sugar recovery. Industrial Crops and Products. 152:112514. https://doi.org/10.1016/j.indcrop.2020.112514

19. NATARAJAN, R.; SUBRAMANIAN, J.; PAPAGEORGIOU, E. I. 2016. Hybrid learning of fuzzy cognitive maps for sugarcane yield classification. Computers and Electronics in Agriculture. 127:147-157. http://dx.doi.org/10.1016/j.compag.2016.05.016

20. NAVARRO HERNÁNDEZ, H.; ROSTGAARD BELTRÁN, L. 2014. Impacto de la materia extraña en la calidad de los jugos de caña y en los indicadores de eficiencia de un central azucarero. Revista Centro Azúcar. 41:44-54.

21. NWANGANGA, F.; CHAPPLE, M. 2020. Practical Machine Learning in R. John Wiley and Sons (Indiana). 464p. https://doi.org/10.1002/9781119591542

22. PÉREZ IGLESIAS, H.; SANTANA AGUILAR, I.; RODRÍGUEZ DELGADO, I. 2015. Manejo sostenible de tierras en la producción de caña de azúcar. Ediciones UTMACH (Ecuador). 188p. Disponible desde Internet en: http://repositorio.utmachala.edu.ec/bitstream/48000/6649/1/16%20MANEJO%20SOSTENIBLE%20DE%20LA%20TIERRA%20EN%20LA%20PRODUCCION%20DE%20CA%C3%91A%20DE%20AZUCAR%20VOL%20II.pdf

23. R CORE TEAM. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria. Disponible desde Internet en: https://www.R-project.org/

24. RAHMAN, M.M.; ROBSON, A.J. 2016. A novel approach for sugarcane yield prediction using landsat time series imagery: A case study on bundaberg region. Advances in Remote Sensing. 5(2):93-102. http://dx.doi.org/10.4236/ars.2016.52008

25. RAMASUBRAMANIAN, K.; SINGH, A. 2019. Machine learning Using R: With time series and industry-based use cases in R. Second Edition. Springer. New York. 724p. https://doi.org/10.1007/978-1-4842-4215-5

26. RIBAS GARCÍA, M.; CONSUEGRA DEL REY, R.; ALFONSO ALFONSO, M. 2016. Análisis de los factores que más inciden sobre el rendimiento industrial azucarero. Revista Centro Azúcar. 43(1):51-60.

27. ROY, M.M.; CHANDRA, A. 2020. Optimizing sugar recovery in India: Need for an integrated approach. Acta Scientific Agriculture. 4(3):1-6. https://doi.org/10.31080/ASAG.2020.04.0806

28. SHAHZAD, S.; SHOKAT, S.; FIAZ, N.; HAMEED, A. 2017. Impact of yield and quality-related traits of sugarcane on sugar recovery. Journal of Crop Science and Biotechnology. 20:1-7. https://doi.org/10.1007/s12892-016-0048-2

29. SMITH, R.I.; DICK, J.; SCOTT, E.M. 2011. The role of statistics in the analysis of ecosystem services. Environmetrics. 22(5):608-617. https://doi.org/10.1002/env.1107

30. SUNDERLAND, T.; BUTTERWORTH, T. 2016. Meeting local economic decision-maker's demand for environmental evidence: The local environment and economic development (LEED) toolkit. Ecosystem Services. 17:197-207. http://dx.doi.org/10.1016/j.ecoser.2015.12.007

31. TARAFDAR, A.; KAUR, B.P.; NEMA, P.K.; BABAR, O.A.; KUMAR, D. 2020. Using a combined neural network - genetic algorithm approach for predicting the complex rheological characteristics of microfluidized sugarcane juice. LWT. 123:109058. https://doi.org/10.1016/j.lwt.2020.109058

32. VANG RASMUSSEN, L.; MERTZ, O.; CHRISTENSEN, A. E.; DANIELSEN, F.; DAWSON, N.; XAYDONGVANH, P. 2016. A combination of methods needed to assess the actual use of provisioning ecosystem services. Ecosystem Services. 17:75-86. http://dx.doi.org/10.1016/j.ecoser.2015.11.005

33. VILLASANTE, S.; LOPES, P.F.M.; COLL, M. 2016. The role of marine ecosystem services for human well-being: Disentangling synergies and trade-offs at multiple scales. Ecosystem Services. 17:1-4. http://dx.doi.org/10.1016/j.ecoser.2015.10.022

34. WAWERU WANGAI, P.; BURKHARD, B.; MULLER, F. 2016. A review of studies on ecosystem services in Africa. International Journal of Sustainable Built Environment. 5:225-245. http://dx.doi.org/10.1016/j.ijsbe.2016.08.005

35. WILLCOCK, S.; MARTÍNEZ-LÓPEZ, J.; HOOFTMAN, D.A.P.; BAGSTAD, K.J.; BALBI, S.; MARZO, A.; PRATO, C.; SCIANDRELLO, S.; SIGNORELLO, G.; VOIGT, B.; VILLA, F.; BULLOCK, J.M.; ATHANASIADIS, I.N. 2018. Machine learning for ecosystem services. Ecosystem Services. 33:165-174. https://doi.org/10.1016/j.ecoser.2018.04.004

36. ZIMMERMAN, D.L. 2020. Linear model theory. With examples and exercises. Springer. Switzerland. 525p. https://doi.org/10.1007/978-3-030-52063-2