



CLASIFICACIÓN DE LA PRODUCCIÓN LECHERA A PEQUEÑA ESCALA
EN LA ZONA FRONTERIZA ECUADOR-COLOMBIA.
UN ESTUDIO COMPARATIVO DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO
CLASSIFICATION OF SMALL-SCALE DAIRY PRODUCTION
IN THE ECUADOR-COLOMBIA BORDER AREA.
A COMPARATIVE STUDY OF AUTOMATIC LEARNING TECHNIQUES

✉L. CARVAJAL-PÉREZ^{1*}, ✉F. MONTENEGRO-ARELLANO¹, ✉G. TERÁN-ROSERO¹,
✉GLADYS URGILÉS-URGILÉS¹, ✉NAYELI CHULDE-CHULDE¹,
✉R. COBO-CUÑA², ✉MAGALY HERRERA-VILLAFRANCA²

¹Universidad Politécnica Estatal del Carchi, Antisana y Universitaria, Tulcán, Ecuador
²Instituto de Ciencia Animal, C. Central, km 47 ½, San José de las Lajas, Mayabeque, Cuba

*Email: luis.carvajal@upec.edu.ec

Se investigaron los factores socioeconómicos determinantes en la producción en granjas lecheras. Se involucró la clasificación de los productores a pequeña escala en la zona fronteriza entre Ecuador y Colombia. Un total de 532 agricultores participaron en la encuesta y los datos recopilados se analizaron mediante técnicas de aprendizaje automático. Los datos se sometieron a un preprocesamiento exhaustivo para eliminar errores y valores atípicos relacionados con los factores socioeconómicos en la producción de leche del Carchi, Ecuador. Entre las variables examinadas, el ingreso económico, el precio por litro de leche y la cantidad de litros utilizados para la producción de queso surgieron como los factores más influyentes. Los resultados mostraron que las técnicas de aprendizaje automático pueden clasificar eficazmente la producción láctea a pequeña escala, con precisión superior a 96 %. La presencia de un hijo que proporciona apoyo económico al hogar, la asignación de leche para la producción como para la venta de queso, junto con su utilización para el consumo familiar, influyeron significativamente en 90 % de los participantes encuestados.

Palabras clave: bienestar económico, modelos de clasificación, pequeños productores lecheros, productividad lechera

The socioeconomic factors determining production in dairy farms were researched. The classification of small-scale farmers in the border area between Ecuador and Colombia was involved. A total of 532 farmers participated in the survey and the data collected was analyzed using automatic learning techniques. The data were subjected to an exhaustive preprocessing to remove errors and outliers related to socioeconomic factors in milk production in Carchi, Ecuador. Among the variables examined, economic income, the price per liter of milk and the quantity of liters used for cheese production emerged as the most influential factors. The results showed that automatic learning techniques can effectively classify small-scale dairy production, with accuracy above 96 %. The presence of a child who provides economic support to the house, the allocation of milk for the production and sale of cheese, together with its use for family consumption, significantly influenced 90 % of the surveyed participants.

Key words: classification models, dairy productivity, economic well-being, small dairy farmers

Recibido: 10 de agosto de 2024

Aceptado: 21 de noviembre de 2024

Conflicto de intereses: Los autores declaran que no existe conflicto de intereses entre ellos.

Declaración de contribución de autoría CRediT: L. Carvajal-Pérez: **Conceptualización, Investigación, Análisis formal, Redacción-borrador original.** F. Montenegro- Arellano: **Conceptualización, Investigación.** G. Terán-Rosero: **Metodología, Análisis formal.** Gladys Urgilés-Urgilés: **Adquisición de fondos, Recursos.** Nayeli Chulde-Chulde: **Curación de datos.** R. Cobo-Cuña: **Validación.** Magaly Herrera-Villafraña: **Análisis formal, Redacción-borrador original**



Este artículo se encuentra bajo los términos de la licencia Creative Commons Attribution-NonCommercial (CC BY-NC 4.0). <https://creativecommons.org/licenses/by-nc/4.0/>



Introducción

La producción de leche es una importante actividad económica en el mundo. Para el 2023, la producción de leche superó los 950 millones de toneladas. En las economías emergentes, aproximadamente 80 % de la producción proviene de explotaciones familiares con uso limitado de insumos, lo que se traduce en menores rendimientos por animal. El 20 % de las explotaciones son medianas y grandes, de ellas 4 % invierte en tecnología para cumplir con los estándares de calidad (FAO 2023a).

En 2022, la Unión Europea (formada por 27 países) fue el mayor productor del mundo con 144 millones de toneladas. Le siguió los Estados Unidos con 103 millones de toneladas y la India con 97 millones de toneladas (Orús 2022). En Ecuador se producían, aproximadamente, 6.15 millones de litros de leche por día, que generaban ingresos para 1,3 millones de habitantes (Ionita 2022). La producción de leche aporta 4 % al producto interno bruto agroindustrial del país y muestra crecimiento de 10.92 % respecto al 2020. La región Sierra aporta 73 % de la producción, la Costa 19 % y la Amazónica 8 % (CIL Ecuador 2023).

La producción de leche usa factores de producción que incluyen la tierra, el capital, la mano de obra, la tecnología y, según algunos autores, la gestión empresarial para transformarlos y contribuir a mejorar las condiciones de vida de los productores.

Los factores sociales de mayor incidencia son el género, el nivel de educación, formación, experiencia o asociatividad (Zemarku et al. 2022). Asimismo, se identificaron factores económicos como ingresos, costos, tamaño del rebaño, volumen de producción (Vásquez et al. 2022); además, en el proceso productivo es fundamental la disponibilidad de tierra, alimentos y cuidados veterinarios (Peña et al. 2018), sin dejar de lado las innovaciones en el sistema de cría y el uso de equipos de automatización para una producción de calidad (Tangorra et al. 2022).

El sector lácteo permite a las poblaciones rurales producir y comercializar sus productos, lo que contribuye al desarrollo económico local, la seguridad alimentaria, el desarrollo económico y, por tanto, a una mejor calidad de vida de los agricultores (FAO 2022a). Es un sector que siempre está cambiando. Necesita invertir en nueva tecnología para ser eficiente. Esto perjudica a los pequeños productores, que no pueden permitirse invertir (Gily Hernández 2019). Además, la cadena de valor de los lácteos impulsa a los pequeños, micro y medianos productores, ya que les ayuda a procesar y vender productos lácteos (Gaudin y Padilla 2020).

El área de estudio comprende la provincia del Carchi. Está en el norte de Ecuador, en la frontera con Colombia. El 63 % del territorio se encuentra en la zona templada húmeda. Se encuentra entre 1.800 y 3.000 m s.n.m. y entre 12 y 18 °C. La temperatura depende de si es seco o lluvioso (Franco 2016).

El otro 37 % se encuentra en la región subtemplada, muy húmeda. Está en los páramos bajos, entre 3.000 y 4.000 m s.n.m. La temperatura es de 6 a 12 °C. Las precipitaciones son de 1000 a 1500 mm al año, sin ningún mes de máxima precipitación (Requelme y Bonifaz 2012).

La producción lechera del Carchi ocupa el tercer lugar en producción nacional. Se basa en las familias, tienen una fuerte presencia en el mercado informal (Morocho et al. 2021), emplea 36 % de la población (Terán y Cobo 2017). Existen 8957 fincas ganaderas (Prefectura del Carchi 2023).

El sistema principal es el extensivo, con prácticas tradicionales y presencia de mucho ganado criollo. Las vacas producen una media de 9.4 L al día. Esto es superior al promedio nacional de 5.9 L (Carvajal 2014). Las fincas con ganado Holstein alcanzan rendimientos de 15 a 18 L por vaca por día (Balarezo et al. 2016), pero son sólo el 6 % del total.

Las unidades de producción agrícola (UPA) cuentan con pequeñas instalaciones de ordeño o establos, lo que refleja su limitada capacidad económica (Velasteguí 2019). En términos de superficie terrestre, existe gran diferencia entre los grupos de productores. Los pequeños agricultores tienen un promedio de 3 ha. Los medianos productores cuentan con 7 ha. Los grandes productores cuentan con 120 ha (Requelme y Bonifaz 2012).

La edad promedio de los productores es de 50 años. Esto muestra pocos jóvenes y poco recambio generacional (Moreno 2018). En términos de educación, 60 % de los productores tiene educación primaria, 25 % tiene educación secundaria y 15 % educación universitaria. La cadena de producción no es competitiva, perjudica la producción y limita el sector agrícola de la región.

Se utilizan varias herramientas en el mundo para evaluar los factores socioeconómicos (FSE) y analizar estrategias para el desarrollo agrícola y alimentario sostenible (FAO 2018). Actualmente, la implementación de prácticas de inteligencia artificial (IA) inclusivas y sostenibles en la agricultura proporciona soluciones para lograr la seguridad alimentaria y nutricional. La IA se aplica en la robótica agrícola, el monitoreo de suelos y cultivos, así como para realizar análisis predictivos (FAO 2022b).

"Machine Learning" (ML) es el campo de estudio conocido como método o arte científico, donde las computadoras pueden aprender a partir de datos mediante programación (Valdez 2019 y Kassahun et al. 2022). Los datos utilizados para aprender se denominan muestras y forman parte del conjunto de entrenamiento. La parte del sistema ML que aprende y hace predicciones se denomina modelo, que comúnmente se prueba mediante el conjunto de pruebas (Gaurav y Patel 2020 y Slob et al. 2021). El aprendizaje automático es bueno, por ejemplo, en problemas que requieren muchas reglas, entornos fluctuantes y en problemas que requieren descubrir conocimientos en grandes cantidades de datos.

Géron (2019) propone tres sistemas ML principales: los supervisados durante el entrenamiento, los que pueden aprender incrementalmente sobre la marcha y los que permiten comparar nuevos puntos de datos con puntos de datos conocidos. Los sistemas de aprendizaje automático pueden clasificar datos según los datos de entrenamiento utilizados para aprender el modelo. Esto abre varias categorías, pero este estudio está impulsado por el aprendizaje supervisado, que requiere las soluciones en los datos de entrenamiento, comúnmente llamadas etiquetas. Un ejemplo de este aprendizaje es la clasificación de los correos electrónicos *spam* (Valdez 2019).

Para Alwadi *et al.* (2024), el clasificador de aumento de gradiente (GBC), usa amplios conjuntos de datos para desarrollar modelos que pronostican la producción y encuentran patrones relevantes. Este método usado en un estudio en Jordania, donde se emplearon sensores para rastrear a 4000 vacas, demostró gran potencial para aumentar la productividad. Asimismo, Bai *et al.* (2022) mostraron que el GBDT-AdaBoost alcanzó precisión promedio de reconocimiento de 98.0 %, superando a otros modelos como el bosque aleatorio y el árbol extremadamente aleatorio, que tuvieron precisiones de 79.9 % y 71.1 %, respectivamente.

Bovo *et al.* (2021) mostraron un clasificador de bosque aleatorio (RF) con error de predicción promedio de 18 % para la producción diaria de leche de cada vaca, y solo de 2 % para la producción total. Esto demuestra que el clasificador de bosque aleatorio es eficaz en la calibración de modelos que ayudan a mejorar la sostenibilidad y eficiencia en la ganadería lechera.

Piwczyński *et al.* (2020) utilizaron un clasificador de árbol de decisión (DT) para identificar los factores que influyen en la alta producción mensual de leche en vacas Holstein-Frisonas en 27 rebaños con robots de ordeño. Los resultados mostraron que la mayor producción mensual (47.24 kg) se registró en vacas múltiparas, ordeñadas más de tres veces al día, en establos con lecho profundo. En contraste, la menor producción (13.56 kg) se observó en vacas ordeñadas menos de dos veces al día, con promedio de menos de 3.97 cuartos

ordeñados. Este modelo permite a los criadores ajustar los factores para maximizar la producción de leche.

Finalmente, Fadillah *et al.* (2023) en un estudio con productores lecheros indonesios sobre la calidad de la leche y los factores asociados con el recuento total en placa (TPC) y el recuento de células somáticas (SCC). Se utilizaron modelos de regresión multinomial y regresión logística de tipo Firth para identificar factores relacionados con el conocimiento de TPC y SCC. Permitieron identificar como variables significativas la pertenencia a cooperativas, la distancia a productores vecinos y la adopción de tecnología para aumentar la conciencia sobre la calidad de la leche entre pequeños productores. En general, tales resultados muestran evidencia de que son modelos aplicables a cualquier región y facilitan la toma de decisiones, basados en resultados con mediciones efectivas.

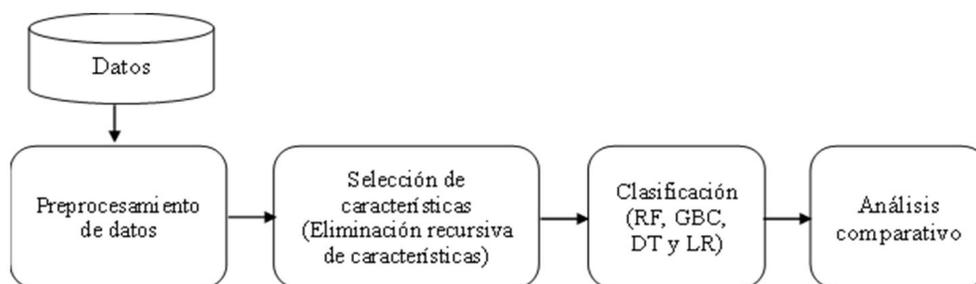
Esta investigación comparó cuatro técnicas diferentes de aprendizaje automático: clasificador de aumento de gradiente (GBC), clasificador de bosque aleatorio (RF), clasificador de árbol de decisión (DT) y regresión logística (LR). Los resultados mostraron que GBC y RF fueron las técnicas de aprendizaje automático más efectivas para clasificar la producción de leche.

Metodología

El presente estudio implica un análisis experimental que consta de cuatro fases: preprocesamiento de datos, selección de características, clasificación y análisis comparativo de los clasificadores.

El flujo de trabajo de la metodología propuesta se muestra en la figura 1, que ilustra las relaciones entre las diferentes fases y la aplicación de algoritmos específicos en cada etapa.

Recopilación de datos: Se encuestó la población de pequeños y medianos productores lecheros de la provincia del Carchi, totalizando 532 individuos. Se empleó un enfoque de investigación aplicada con una metodología exploratoria y correlacional (Hernández-Sampieri y Mendoza 2018). El cuestionario abordó una variedad de factores, brindando información sobre aspectos relevantes para la comunidad de productores lecheros:



Fuente: Elaboración propia

Figura 1. Flujo de trabajo para predecir la producción láctea a pequeña escala

- **Social:** edad, género, nivel educativo, estructura familiar, formación, acceso a la tecnología, condiciones de vivienda, servicios básicos, empleo, asociatividad, gobernanza y participación, apoyo técnico gubernamental
- **Económico:** ingresos ganaderos, otros ingresos, costos de producción, distribución del ingreso, financiamiento, comercialización, tamaño de la finca.
- **Productivo:** uso de la tierra, tamaño y estructura del rebaño, número de cabezas de ganado, pastos, producción de leche por hectárea (L ha⁻¹), adopción de tecnología y diversificación productiva. número de cabezas de ganado.

Se incorporaron 17 preguntas con información cuantitativa, 23 preguntas de intervalo y 10 preguntas dicotómicas. El cuestionario fue desarrollado rigurosamente y se validó su contenido y estructura. La recolección de datos en campo se llevó a cabo en colaboración con estudiantes de Administración de Empresas de la Universidad Politécnica Estatal del Carchi (UPEC), Ecuador, durante el segundo semestre de 2022. Se aplicó un muestreo aleatorio simple.

Preprocesamiento de datos: Los datos recopilados se sometieron a un riguroso proceso de preprocesamiento, que incluyó la eliminación de errores y valores atípicos, así como el tratamiento de valores faltantes. Se aplicó la normalización Min-Max para garantizar que todas las características tuvieran un rango común y fueran comparables entre sí (Treviño Cantú 2022). Esto permitió eliminar cualquier sesgo, debido a la escala de los datos, asegurando un análisis más preciso y justo.

Selección de características: La selección de funciones desempeña una función fundamental en la fase de preprocesamiento de datos antes de aplicar las técnicas de aprendizaje automático (Siddiqui y Amer 2024). Implica seleccionar las características más relevantes e informativas del conjunto de datos, descartando al mismo tiempo las características irrelevantes o redundantes. En este estudio, se utilizó la selección de características para mejorar el rendimiento y la interpretabilidad de los modelos de aprendizaje automático para clasificar los productores lecheros a pequeña escala en la región fronteriza entre Ecuador y Colombia.

El conjunto de datos utilizado en esta investigación contiene varias variables socioeconómicas y relacionadas con la producción que potencialmente podrían influir en la producción de leche. Sin embargo, no todas estas variables son igualmente importantes para la tarea de predicción. Algunas características pueden introducir ruido, aumentar la carga computacional o provocar un sobreajuste, lo que dificulta la capacidad del modelo para generalizar bien datos invisibles.

Para abordar estos desafíos e identificar las características más influyentes, se empleó la técnica de eliminación de características recursivas (RFE). Es un método de selección de funciones popular y potente que funciona ajustando recursivamente el modelo de aprendizaje automático,

eliminando las funciones menos significativas en cada iteración. El proceso continúa hasta obtener el número deseado de características. La importancia de RFE radica en su capacidad para clasificar características en función de su contribución al rendimiento del modelo, lo que permite centrarse en los atributos más relevantes y descartar los menos informativos (Mannepalli et al. 2024).

La base de datos inicial constaba de 134 ítems, incluidas variables numéricas, dicotómicas y categóricas. Con el objetivo de reducir la dimensionalidad de los datos y el costo computacional durante el entrenamiento del modelo, se aplicó la selección de características y finalmente se redujo el conjunto a 10 variables. Se incluyeron el tipo de vivienda, el acceso a agua potable y electricidad, la comercialización de leche cruda, las ventas de queso pasteurizado, el uso de leche para la producción de queso, las relaciones con los clientes, el ingreso total anual de la actividad primaria, los litros utilizados para la producción de queso y el precio por litro.

Algoritmo de clasificación

Gradient Boosting Classifier (GBC): Es un clasificador que destaca por su precisión y velocidad de predicción sobre conjuntos de datos grandes y complejos. También minimiza el error de sesgo del modelo (Bentéjac et al. 2020). Este método se utiliza cuando sólo hay dos clases en las características objetivo, es decir, clases binarias (positivas y negativas). La función de pérdida como log-verosimilitud se utiliza en la creación (entrenamiento) del modelo (Natekin y Knoll 2013). Esta pérdida se muestra en la ecuación (1):

$$L(\theta) = - \sum y_i \log(p(y_i | x_i; \theta)) \quad (1)$$

donde y_i es el objetivo de clasificación, p es la probabilidad prevista de clase 1 y θ es la entrada.

La función de pérdida encuentra los residuos después de crear el árbol de decisión con todas las variables independientes y objetivo. Cuando se construye el primer árbol, la salida final se encuentra por las hojas (Saini 2021). La fórmula directa para calcular el resultado final se muestra en la ecuación (2):

$$Y = \frac{\sum_{i=1}^n \text{Residual}_i}{\sum_{i=1}^n [\text{Previous probability}_i \times (1 - \text{Previous probability}_i)]} \quad (2)$$

donde Y es la función objetivo para la decisión de clasificación.

Clasificador de bosque aleatorio (RF): Se denomina bosque de árbol de decisión. Este método se basa en el principio de ensacado con selección aleatoria de características y el modelo utiliza votación para combinar predicciones de árboles. RF funciona bien para la mayoría de los problemas, puede manejar el ruido y selecciona solo las funciones más importantes. Sin embargo, la interpretabilidad

del modelo es limitada y su ajuste requiere cierto esfuerzo en el manejo de los datos (Gaurav y Patel 2020).

Clasificador de árbol de decisión (DT): Es un algoritmo de aprendizaje automático supervisado que se puede utilizar para categorización o predicción. Los DT están diseñados para imitar el pensamiento humano, lo que hace que los resultados sean fáciles de entender e interpretar. Los seis componentes clave de un DT son el nodo raíz, la división, el nodo de decisión, el nodo hoja, la poda y la rama (Suthaharan 2016).

Los DT se utilizan en problemas que involucran datos y variables, numéricos como categóricos. Son eficaces para modelar problemas con múltiples resultados y para probar la confiabilidad de los árboles. Otra ventaja de los DT es que requieren menos limpieza de datos en comparación con otras técnicas de modelado de datos. Sin embargo, es importante reconocer que los DT se pueden ver afectados por el ruido y pueden no ser ideales para conjuntos de datos más grandes (Kliś et al. 2021).

Regresión logística (LR): También llamada regresión logit, se utiliza para estimar la probabilidad de que una instancia pertenezca a una clase determinada. Normalmente, se utiliza para tareas de clasificación binaria donde las clases se etiquetan como 0 y 1, según un umbral de probabilidad (Géron 2019). La probabilidad estimada de LR se presenta en la ecuación (3):

$$\hat{p} = h_{\theta}(\mathbf{x}) = \sigma(\theta^t \cdot \mathbf{x}) \quad (3)$$

donde $\sigma(t)$ es una función sigmoidea que produce un número entre 0 y 1, dado por la función logística que se muestra en la ecuación (4):

$$\sigma(t) = \frac{1}{1 + e^{-t}} \quad (4)$$

donde t es el tiempo

La evaluación de los modelos de aprendizaje automático se describe a continuación:

- *Precisión o Proximidad de los resultados*: Utiliza los parámetros verdadero positivo (TP), verdadero negativo (TN), falso positivo (FP), falso negativo (FN).
- *El área bajo la curva (AUC)*: Mide la capacidad del modelo para discriminar entre dos clases.
- *Recuerdo o probabilidad de clasificar verdaderos positivos*: Utiliza los parámetros verdadero positivo (TP), falso negativo (FN).
- *Precisión o dispersión del conjunto de valores obtenidos*: Utiliza los parámetros verdadero positivo (TP), falso positivo (FP).
- *F1 (F-Score)*: Combina medidas de precisión y recuperación en un solo valor.
- *Kappa* cuantifica la concordancia entre las predicciones realizadas por un modelo y las clases verdaderas. Se utiliza para evaluar el rendimiento predictivo de diferentes clases.

- El tiempo de entrenamiento (TE Seg) mide el tiempo que tarda un modelo en aprender del conjunto de datos de entrenamiento y ajustar sus parámetros para obtener predicciones precisas.

Resultados y Discusión

La preparación de algoritmos de aprendizaje automático, incluida la selección de funciones y el entrenamiento de modelos, se realizó utilizando una combinación de herramientas de ciencia de datos de última generación. El código utilizado para este propósito, basado en las bibliotecas 'pycaret' y 'scikit-learn' en Python, formó la piedra angular del enfoque metodológico.

La implementación del modelo utilizando funciones estándar de 'scikit-learn' proporcionó una base sólida para el proceso de capacitación. En este estudio, se omitió intencionalmente el ajuste de hiperparámetros, basándose en su lugar en los parámetros predeterminados inherentes a cada modelo. Esta elección estratégica se hizo para mantener la coherencia metodológica y facilitar las comparaciones directas entre modelos. La adopción de configuraciones predeterminadas inherentes a cada algoritmo tenía como objetivo mantener un marco estandarizado en todos los análisis, garantizando la transparencia y reproducibilidad de los experimentos.

El mejor modelo entrenado con el conjunto de datos discutido anteriormente fue GBC, que logró 96.77 % de predicciones correctas en la fase de prueba. Adicionalmente, el porcentaje de la habilidad de evaluación predictiva del modelo entrenado fue de 96.9 %, y en la evaluación del desempeño alcanzó 93.50 %. También se midieron otras métricas importantes como AUC, recuperación y precisión, que obtuvieron 99.4, 97.90 y 96.10 %, respectivamente. Asimismo, las métricas para modelos como RF, DT y LR se presentan en la tabla 1.

En este estudio, se midió el tiempo de entrenamiento de los modelos. En GBC, la capacitación duró aproximadamente 0.9 segundos. RF, DT y LR en su entrenamiento lograron 1, 0.63 y 0.77 segundos respectivamente. Estos resultados y la precisión de cada modelo se muestran en la figura 2.

Una fase esencial en la formación del mejor modelo fue la importancia de las características. El modelo GBC, que es el mejor, encontró que la característica correspondiente a "ingreso principal" tuvo una métrica del 80 %. La importancia de las características se presenta en la figura 3.

La figura 4 muestra la matriz de predicción y los cuadros superior izquierdo e inferior derecho corresponden a las predicciones correctas, mientras que los cuadros superior derecho e inferior izquierdo contienen predicciones incorrectas o falsos positivos.

Nyambo et al. (2023) aplicaron técnicas de aprendizaje automático (ML) en la industria láctea de Tanzania.

Tabla 1. Resultados de los algoritmos de clasificación

Algoritmo	Exactitud, %	AUC, %	Recordar, %	Prec, %	F1, %	Kappa, %	TT, seg.
GBC	0.9677	0.994	0.979	0.961	0.969	0.935	0.90
RF	0.9518	0.984	0.964	0.946	0.954	0.903	1.00
DT	0.9489	0.956	0.943	0.96	0.95	0.898	0.63
LR	0.9141	0.977	0.948	0.894	0.919	0.828	0.77

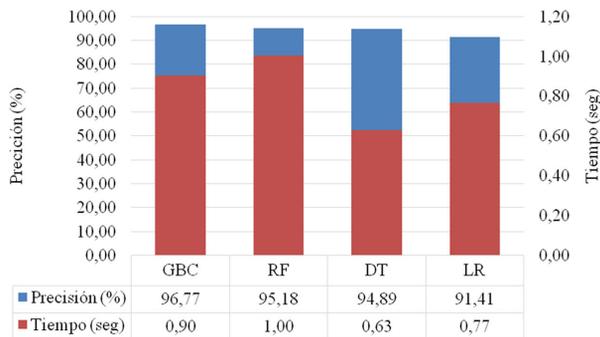


Figura 2. Precisión y tiempo de ejecución de los algoritmos de aprendizaje automático mejor valorados

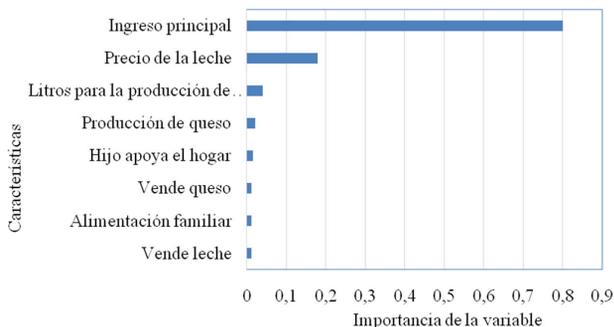


Figura 3. Características importantes del modelo GBC

Clase verdadera	0	71	6
	1	5	78
		0	1
		Clase predicha	

Figura 4. Matriz de confusión del mejor modelo de clasificación.

Su estudio se centró en tres cuestiones principales: infraestructura inadecuada, tecnología obsoleta y baja productividad. Analizaron los datos y encontraron grupos de producción homogéneos. Luego, hicieron recomendaciones

para aumentar la producción de leche. De manera similar, *Mwanga et al. (2020)* utilizaron ML para identificar grupos de agricultores. En su caso, la clasificación se basó en la ubicación de la finca. También se basó en el sistema de alimentación y cuidado de los animales. Esta información favoreció mejor planificación y gestión de recursos. Permitió intervenciones más precisas en cada grupo para mejorar los servicios.

Autores como *Abdukarimova et al. (2016)* mencionan que estimar la producción de leche ayuda a evaluar el desempeño de los recursos. Sin embargo, existen varios desafíos asociados con la previsión de la producción de leche, especialmente en la clasificación eficaz.

Ji et al. (2022) ejecutaron un marco de aprendizaje automático con la utilización de cinco años de datos de productividad y salud conductual de 80 vacas. Lograron una precisión de más de 80 %. Otros autores como *Radwan et al. (2020)* han propuesto un modelo lineal dinámico (DLM) y una red neuronal artificial (RNA) en la predicción de la producción de leche. El DLM logró precisión de 95 %, utilizando un conjunto de datos que constaba de 1094780 observaciones de información de sensores proporcionados por Lely Industries (Masslui, Países Bajos). La RNA logró precisión del 79.5 %, superando las expectativas de producción de leche.

A pesar de los desafíos involucrados, este trabajo comparó diferentes modelos de aprendizaje automático (GBC, RF, DT, LR) en un conjunto de datos de producción de leche de Carchi, provincia de Ecuador. Los resultados mostraron precisión de clasificación significativa: GBC logró una precisión del 96.77 % y recuperación del 97.9 %. RF logró precisión de 95.18 % y puntuación F1 de 95.4 %.

La abundancia de datos en el sector ganadero requiere enfoques de análisis innovadores. Este estudio investigó el potencial de los modelos de aprendizaje profundo, específicamente seis algoritmos de redes neuronales, como alternativa a los métodos estadísticos tradicionales. En comparación con estos métodos tradicionales, los modelos de aprendizaje profundo pueden lograr una mayor precisión, lo que los convierte en herramientas valiosas para identificar variables agrícolas y desarrollar productos lácteos seguros, y prácticas de gestión de riesgos (*Suseendran y Duraisamy 2021*).

Los investigadores utilizaron métodos de clasificación para identificar variables relevantes, y luego emplearon estas variables para entrenar varios modelos predictivos. Estos modelos incluían no solo algoritmos de aprendizaje profundo, sino también establecidas como regresión logística, k vecinos más cercanos, árboles de decisión y bosques aleatorios. Si bien la mayoría de los modelos lograron un alto rendimiento predictivo de 93 %, las redes neuronales y los modelos de mezcla gaussiana demostraron ser más sensibles a las variaciones en el conjunto de datos. En respuesta, los investigadores combinaron algoritmos de bosque aleatorio y de árbol de decisión para mejorar la selección de factores (Mwanga *et al.* 2020).

Los resultados de la encuesta mostraron que el principal ingreso económico derivado de la producción de leche (89 %), el precio por litro de leche (46 %) y la cantidad de litros de leche utilizados para la producción de queso (18 %) fueron los factores más importantes en la producción. También tuvieron un impacto significativo la presencia de un hijo como sustento económico del hogar (5 %), el destino de la leche para la producción y venta de queso (21 %) y el uso de la producción de leche y queso para el consumo doméstico (53 %), pero en menor medida.

El estudio describe los FSE clave que dan forma a la dinámica familiar y la producción agrícola en la comunidad estudiada. Se destaca que el 90 % de los productores que mantienen condiciones adecuadas del hogar, el nivel educativo no muestra influencias en las decisiones de bienestar familiar. Sin embargo, el nivel educativo universitario con que cuentan algunos productores señala la presencia de mayores ingresos y mejores tasas de producción. Además, prevalece un modelo patriarcal de sostén de familia, en el que los maridos asumen esta función en 75 % de los hogares. La edad también emerge como un factor. Se observó aumento de la convivencia entre los 50 y 55 años. Además, la experiencia se entrelaza con la educación, ya que ambas tienen un impacto significativo en los niveles de producción. Estos hallazgos subrayan la compleja interacción entre educación, ingresos, estructura del hogar y productividad agrícola y brindan información valiosa para la elaboración de modelos socioeconómicos y estrategias de desarrollo.

El estudio sugiere mayor exploración a través de un análisis de eficiencia técnica productiva, que incluiría variables como infraestructura, mano de obra, manejo de productos, procesos de ordeño, manejo, prácticas ambientales y control de calidad. Un análisis de este tipo permitiría optimizar las capacidades de producción en una unidad de producción. De esto se pueden derivar intervenciones específicas para mejorar la eficiencia de la producción, facilitar el acceso justo a los mercados y racionalizar las actividades de procesamiento de lácteos con valor agregado.

Conclusiones

Este estudio ha identificado los factores que influyen en la producción en pequeñas explotaciones lecheras en la región fronteriza entre Ecuador y Colombia. Los resultados de este estudio se pueden utilizar para informar futuras investigaciones y decisiones destinadas a apoyar la sostenibilidad y el desarrollo del sector lácteo en la región. Al arrojar luz sobre los determinantes clave de la producción de leche y su impacto en el bienestar económico de las familias rurales, esta investigación proporciona una guía valiosa para las partes interesadas y los responsables para formulación de intervenciones e iniciativas específicas.

Este estudio, en el contexto único de la zona fronteriza ecuatoriana, destaca el potencial de las técnicas de aprendizaje automático para clasificar con precisión la producción de leche de los pequeños agricultores. La aplicación exitosa de algoritmos de aprendizaje automático, incluidos Gradient Boosting Classifier y Random Forest, ha demostrado su eficacia para clasificar la producción de leche con una precisión notable.

Los resultados de este estudio tienen implicaciones significativas para la industria láctea en la región fronteriza entre Ecuador y Colombia, y más allá. Los factores identificados que influyen en la producción de leche proporcionan una hoja de ruta para mejorar la productividad y los medios de vida en las comunidades de productores lecheros de pequeña escala.

A medida que el sector lácteo continúa desempeñando un papel fundamental en la economía de la región, aprovechar el poder del aprendizaje automático para identificar variables relevantes será fundamental para dar forma a los modelos predictivos, promover el crecimiento sostenible y fortalecer el bienestar económico general del sector.

Referencias

- Abdukarimova, M., Abdukarimov, A. & Abdukarimov, N. 2016. *Handbook of Industrial and Innovation Economics*, editado por Munisa, 466p. Uzbekistan: Independently. ISBN: 979-8412353852. Available at: https://www.researchgate.net/profile/Munisa-Abdukarimova/publication/344279960_Handbook_of_Industrial_and_innovation_economics/links/62493f3621077329f2ed6414/Handbook-of-Industrial-and-innovation-economics.pdf.
- Alwadi, M., Alwadi, A., Chetty, G. & Alnaimi, J. 2024. Smart dairy farming for predicting milk production yield based on deep machine learning. *International Journal of Information Technology*, 16: 4181-4190, ISSN: 2511-2112. <https://doi.org/10.1007/s41870-024-01998-5>.
- Bai, J., Xue, H., Jiang, X. & Zhou, Y. 2022. Recognition of bovine milk somatic cells based on multi-feature extraction and a GBDT-AdaBoost fusion model. *Mathematical Biosciences and Engineering: MBE*, 19(6): 5850-5866, ISSN: 1551-0018. <https://doi.org/10.3934/mbe.2022274>.

- Balarezo, L., García-D, J., Hernández, M. & García-L, R. 2016. Metabolic and reproductive state of Holstein cattle in the Carchi region, Ecuador. *Cuban Journal of Agricultural Science*, 50(3): 381-392, ISSN: 2079-3480. <https://cjasience.com/index.php/CJAS/article/view/632/699>.
- Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. 2020. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3): 1937-1967, ISSN: 1573-7462. <https://doi.org/10.1007/s10462-020-09896-5>.
- Bovo, M., Agrusti, M., Benni, S., Torreggiani, D., & Tassinari P. 2021. Random Forest Modelling of Milk Yield of Dairy Cows under Heat Stress Conditions. *Animals*, 11(5): 1305, ISSN: 2076-2615. <https://doi.org/10.3390/ani11051305>.
- Carvajal, L.A. 2014. La asociatividad en el sector agropecuario del Carchi y su potencial de producir y comercializar semielaborados de papa y leche. *SATHIRI*, 7(7): 153-163, ISSN: 2631-2905. <https://doi.org/10.32645/13906925.348>.
- CIL Ecuador. 2023. *La industria láctea fomenta la economía circular, a través de una producción sostenible*, Comprometidos con el Desarrollo de la Cadena Láctea. Available at: <https://www.cil-ecuador.org/post/la-industria-láctea-fomenta-la-economía-circular-a-través-de-una-producción-sostenible>. [Consulted: March 10, 2024].
- Fadillah, A., van den Borne, B.H.P., Poetri, O.N., Hogeveen, H., Umberger, W., Hetherington, J., & Schukken, Y.H. 2023. Smallholder milk-quality awareness in Indonesian dairy farms. *Journal of Dairy Science*, 106(11): 7965-7973, ISSN: 0022-0302. <https://doi.org/10.3168/JDS.2023-23267>.
- FAO. 2018. *Panorama de la pobreza rural en América Latina y el Caribe*. Roma, 114p. ISBN: 978-92-5-131085-4 Available at: <https://openknowledge.fao.org/handle/20.500.14283/ca2275es>. [Consulted: February 03, 2024].
- FAO. 2022a. *The State of Food and Agriculture 2022*. Roma, 182p. ISBN: 978-92-5-136043-9. <https://doi.org/10.4060/cb9479en>.
- FAO. 2022b. *La aplicación de las mejores prácticas de la inteligencia artificial en el contexto de la agricultura*, editado por Bishan Dong, 136. Roma: FAO Publications Catalogue 2022. ISBN: 78-92-5-136969-2.
- FAO. 2023a. *FAO analiza fortalezas y brechas de la producción láctea en América Latina y el Caribe*, Más Allá de La Finca Lechera. Available at: <https://www.fao.org/americas/noticias/ver/es/c/1617544/>. [Consulted: July 18, 2024].
- Franco, W. 2016. Propuestas para la innovación en los sistemas agroproductivos y el desarrollo sostenible del Valle Interandino en Carchi, Ecuador. *Tierra Infinita*, 2(1): 49-87, ISSN: 2631-2921. <https://doi.org/10.32645/26028131.104>.
- Gaudin, Y. & Padilla, R. 2020. Los intermediarios en cadenas de valor agropecuarias: un análisis de la apropiación y generación de valor agregado (N° 186 (LC/TS.2020/77; LC/MEX/TS.2020/15)). Serie Estudios y Perspectivas-Sede Subregional de La CEPAL en México. Available at: <https://www.cepal.org/es/publicaciones/45796-intermediarios-cadenas-valor-agropecuarias-un-analisis-la-apropiacion-generacion>. [Consulted: August 20, 2024].
- Gaurav, K.A. & Patel, L. 2020. Machine Learning With R. In S. Khalid (Ed.), *Applications of Artificial Intelligence in Electrical Engineering* (pp. 291-331), ISBN: 9781799827184. IGI Global. <https://doi.org/10.4018/978-1-7998-2718-4.ch015>.
- Géron, A. 2019. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media. ISBN: 978-1-492-03264-9. Available at: https://books.google.com.ec/books?id=HnetDwAAQBAJ&printsec=frontcover&hl=es&source=gbs_book_other_versions#v=onepage&q&f=false. [Consulted: August 10, 2024].
- Gil Montelongo, M. & Hernández Villa, X. 2019. Risk management as a tool in the internal control on organizations of the dairy sector. *Ekotemas*, 5(2): 51-66, ISSN: 2414-4681. <https://www.ekotemas.cu/index.php/ekotemas/article/view/63/54>.
- Hernández-Sampieri, R., & Mendoza, C. 2018. Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta. In Interamericana (Ed.), McGRAW-HILL Interamericana Editores S.A. de C.V. Mc Graw Hill. ISBN: 978-1-4562-6096-5.
- Ionita, E. 2022. La producción de leche en Ecuador, Veterinaria Digital. Available at: <https://www.veterinariadigital.com/articulos/la-produccion-de-leche-en-ecuador/>. [Consulted: January 20, 2024].
- Ji, B., Banhazi, T., Phillips, C.J.C., Wang, C. & Li, B. 2022. A machine learning framework to predict the next month's daily milk yield, milk composition and milking frequency for cows in a robotic dairy farm. *Biosystems Engineering*, 216(9): 186-197, ISSN: 1537-5110. <https://doi.org/10.1016/j.biosystemseng.2022.02.013>.
- Kassahun, A., Bloo, R., Catal, C. & Mishra, A. 2022. Dairy Farm Management Information Systems. *Electronics*, 11(2): 1-18, ISSN: 2079-9292. <https://doi.org/10.3390/electronics11020239>.
- Kliś, P., Piwczyński, D., Sawa, A. & Sitkowska, B. 2021. Prediction of Lactational Milk Yield of Cows Based on Data Recorded by AMS during the Periparturient Period. *Animals*, 11(383): 1-11, ISSN: 2076-2615. <https://doi.org/10.3390/ANI11020383>.
- Mannepalli, P.K., Kulurkar, P., Jangade, V., Khan, A., & Singh, P. 2024. An Enhanced Classification Model for Depression Detection Based on Machine Learning

- with Feature Selection Technique. En P. K. Jha, B. Tripathi, E. Natarajan, & H. Sharma (Eds.), *Proceedings of Congress on Control, Robotics, and Mechatronics* (Vol. 364, pp. 589-601). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-5180-2_46
- Moreno, F. 2018. Caracterización socioeconómica y productiva de la cadena de valor agroalimentaria de la leche en la provincia de Tungurahua. Tesis presentada en opción al Título de carrera de Ingeniería de los alimentos, Universidad Técnica de Ambato, Ecuador.
- Morocho, B., Carvajal, H. & Vite, H. 2021. Análisis socioeconómico del agronegocio ganadero: Caso productores de la Aso Ganaderos del Altiplano Orense 5 de noviembre del cantón Atahualpa. *Revista Metropolitana de Ciencias Aplicadas*, 4(1): 26-32, ISSN: 2631-2662.
- Mwanga, G., Lockwood, S., Mujibi, D., Yonah, Z. & Chagunda, M. 2020. Machine learning models for predicting the use of different animal breeding services in smallholder dairy farms in Sub-Saharan Africa. *Tropical Animal Health and Production*, 52(3): 1081-1091, ISSN: 1573-7438. <https://doi.org/10.1007/s11250-019-02097-5>.
- Natekin, A. & Knoll, A. 2013. Gradient boosting machines, a tutorial. *Frontiers in Neurobotics*, 7(21): 1-21, ISSN: 1662-5218. <https://doi.org/10.3389/fnbot.2013.00021>.
- Nyambo, D.G., Malamsha, G.C. & Mavura, F. 2023. Leveraging Machine Learning Techniques to Improve Learning and Recommendations Within Dairy Farms: Towards High Milk Yields for Small-Scale Farmers. In F. Mtenzi, G. Oreku, & D. Lupiana (Eds.), *Impact of Disruptive Technologies on the Socio-Economic Development of Emerging Countries* (pp. 172-188), ISBN: 9781668468739. IGI Global. <https://doi.org/10.4018/978-1-6684-6873-9.ch011>.
- Orús, A. 2022. Leche de vaca: principales productores a nivel mundial en 2022. Estadista. Available at: <https://es.statista.com/estadisticas/600241/principales-productores-de-leche-de-vaca-en-el-mundo-en/>. [Consulted: April 30, 2024].
- Piwczyński, D., Sitkowska, B., Kolenda, M., Brzozowski, M., Aerts, J. & Schork, P.M. 2020. Forecasting the milk yield of cows on farms equipped with automatic milking system with the use of decision trees. *Animal Science Journal*, 91(1): e13414, ISSN: 1740-0929. <https://doi.org/10.1111/asj.13414>.
- Peña, Y., Benitez, D., Ray, J. & Fernández, Y. 2018. Factores determinantes de la producción ganadera en una comunidad campesina del suroeste de Holguín, Cuba. *Cuban Journal of Agricultural Science*, 52(2): 155-163, ISSN: 2079-3480. http://scielo.sld.cu/scielo.php?pid=S2079-34802018000200155&script=sci_arttext&tlng=es
- Prefectura del Carchi. 2023. Datos informativos de la provincia. Available at: <https://carchi.gob.ec/2016f/index.php/informacion-provincial.html>. [Consulted: April 25, 2024].
- Radwan, H., Qaliouby, H. & Elfadl, E. 2020. Classification and prediction of milk yield level for Holstein Friesian cattle using parametric and non-parametric statistical classification models. *Journal of Advanced Veterinary and Animal Research*, 7(3): 429-435, ISSN: 2311-7710. <https://doi.org/10.5455/javar.2020.g438>.
- Requelme, N. & Bonifaz, N. 2012. Caracterización de sistemas de producción lechera de Ecuador. *La Granja*, 15(1): 56-69, ISSN: 1390-3799.
- Saini, A. 2021. Gradient Boosting Algorithm: A Complete Guide for Beginners. Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/>. [Consulted: March 21, 2024].
- Siddiqui, T. & Amer, A.Y.A. 2024. A comprehensive review on text classification and text mining techniques using spam dataset detection. In *Mathematics and Computer Science*, vol. 2, editado por Ghosh, S., Niranjanamurthy, M., Deyasi, K., Mallik, B. & Das, S., 1-17. Editorial Wiley, ISBN: 978-111989671-5. <https://doi.org/10.1002/9781119896715.ch1>.
- Slob, N., Catal, C. & Kassahun, A. 2021. Application of machine learning to improve dairy farm management: A systematic literature review. *Preventive Veterinary Medicine*, 187: 105237, ISSN: 1873-1716. <https://doi.org/10.1016/j.prevetmed.2020.105237>.
- Suthaharan, S. 2016. Decision Tree Learning, In *Machine Learning Models and Algorithms for Big Data Classification*, Integrated Series in Information Systems, vol 36. Springer, Boston, MA., 237-269, ISBN: 9781489976413. https://doi.org/10.1007/978-1-4899-7641-3_10.
- Suseendran, G. & Duraisamy, B. 2021. Predication of Dairy Milk Production Using Machine Learning Techniques. In: Peng, S.L., Hsieh, S.Y., Gopalakrishnan, S., Duraisamy, B. (eds) *Intelligent Computing and Innovation on Data Science. Lecture Notes in Networks and Systems*, 248: Springer, Singapore, ISBN: 978-981-16-3153-5. https://doi.org/10.1007/978-981-16-3153-5_60.
- Tangorra, F. M., Calcante, A., Vigone, G., Assirelli, A. & Bisaglia, C. 2022. Assessment of technical-productive aspects in Italian dairy farms equipped with automatic milking systems: A multivariate statistical analysis approach. *Journal of Dairy Science*, 105(9): 7539-7549, ISSN: 0022-0302. <https://doi.org/10.3168/jds.2021-20859>.
- Terán, G. & Cobo, R. 2017. Determining management factors in dairy farms in Carchi, Ecuador. *Cuban Journal of Agricultural Science*, 51(2): 175-182, ISSN: 2079-3480. <http://cjas.science.com/index.php/CJAS/article/view/724>.

- Treviño Cantú, J.A. 2022. Alternativas de estandarización para índices compuestos espacio-temporales. El caso del rezago educativo en los estados de México, 2000 a 2020. *Investigaciones Geográficas*, 109: 1-14, ISSN: 2448-7279. <https://doi.org/10.14350/ig.60615>.
- Valdez, A. 2019. Machine Learning para todos. En *IV Congreso Nacional de Profesionales de Computación, Informática y Tecnologías*. pp. 60. Perú: Ministerio de Educación. <https://doi.org/10.13140/RG.2.2.13786.70086>.
- Vásquez, H., Barrantes, C., Vigo, C. & Maicelo, J. 2022. Factores socioeconómicos que influyen en la adopción de tecnologías para mejoramiento genético de ganado vacuno en Perú. *Agricultura, Sociedad y Desarrollo*, 19(3): 312-330, ISSN: 2594-0244. <https://doi.org/10.22231/asyd.v19i3.1358>.
- Velasteguí, N. 2019. Cadena productiva del sector lechero en la provincia de Tungurahua, cantón Pillaro: Un estudio socio-económico de la producción de la leche cruda. Tesis presentada en opción al Título de carrera de Economía, Universidad Técnica de Ambato, Ecuador.
- Zemarku, Z., Senapathy, M. & Bojago, E. 2022. Determinants of Adoption of Improved Dairy Technologies: The Case of Offa Woreda, Wolaita Zone, Southern Ethiopia. *Advances in Agriculture*, 2022: 1-19, ISSN: 2314-7539. <https://doi.org/10.1155/2022/3947794>.