UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2022, Vol. 15, No. 1, pp. 35–47

Predicción de la producción diaria de leche en bovinos Gyr a través de métodos de aprendizaje supervisado

Prediction of daily milk yield in Gyr cattle through supervised learning methods

Alberto Zea[a]
albertozea@usantotomas.edu.co

Dagoberto Bermúdez[b]
dagobertobermudez@usantotomas.edu.co

Ariel Jiménez[c]
ariel.jimenez@asocebu.com

Germán Gómez[d]
german.gomez@asocebu.com

Carlos Alberto Martínez[e]
camartinezn@unal.edu.co

Resumen

La Asociación Colombiana de Criadores de Ganado Cebú - ASOCEBU, tiene interés en desarrollar una máquina para predecir la producción total diaria de leche empleando mediciones de producción parciales en ganado Gyr y, en particular, responder dos preguntas: 1) ¿puede un método predictivo de referencia ser superado por métodos desarrollados a nivel local? 2) ¿cuál de los dos registros parciales (AM o PM) tiene un mejor desempeño predictivo? Por lo tanto, el objetivo de este artículo fue desarrollar una máquina predictiva para la producción diaria de leche en bovinos Gyr utilizando registros parciales, intervalo entre ordeños, días en lactancia, y número de partos (n=13806), mediante la implementación de métodos de aprendizaje supervisado. Además de la máquina predictiva de referencia, varias combinaciones de variables de entrada y modelo o método de aprendizaje fueron consideradas. Se emplearon redes neuronales artificiales, máquinas de soporte vectorial, bosques aleatorios y regresión lineal cuyos parámetros de localización se estimaron mediante mínimos cuadrados o los métodos de encogimiento Ridge y Lasso. El desempeño predictivo (DP) se evaluó mediante validación cruzada utilizando las siguientes funciones de error: la raíz del error cuadrático medio (RECM)

[a]Estudiante. Maestría en Estadística Aplicada, Universidad Santo Tomas
[b]Docente. Facultad de Estadística, Universidad Santo Tomas
[c]Asociación Colombiana de Criadores de Ganado Cebú-ASOCEBU
[d]Asociación Colombiana de Criadores de Ganado Cebú-ASOCEBU
[e]Departamento de Producción Animal, Facultad de Medicina Veterinaria y de Zootecnia, Universidad Nacional de Colombia, Sede Bogotá.

y el error absoluto medio (EAM). Se encontró que una red neuronal artificial con una capa oculta empleando el registro parcial AM, intervalo entre ordeños, número de partos y días en leche como variables de entrada presentó el mejor DP (RECM=1.5042, EAM=1.1389), pero en general, el desempeño de los diferentes métodos fue similar. Todas las máquinas cuyos parámetros se aprendieron empleando los datos locales fueron superiores al método de referencia y los registros parciales de la mañana presentaron mejor DP que los de la tarde. Estos resultados permiten direccionar el programa de control lechero de ASOCEBU y generan un método "a la medida" para predecir la producción total diaria de leche en ganado Gyr en Colombia, un componente importante de los programas de mejoramiento genético y modelamiento del nivel productivo en esta raza.

***Palabras clave*:** aprendizaje automático, control lechero, validación cruzada.

Abstract

The Asociación Colombiana de Criadores de Ganado Cebú - ASOCEBU, has interest in developing a machine to predict total daily milk yield using partial production measurements in Gyr cattle and, in particular, answering two questions: 1) can a reference predictive method be outperformed by locally developed methods? 2) which one of the two partial records (AM or PM) has a better predictive performance? Therefore, the objective of this paper was to develop a predictive machine for daily milk yield in Gyr cattle using partial records, milking interval, days in milk, and parity (n=13806), by implementing supervised learning methods. Besides the reference predictive machine, several combinations of input variables and model or learning method were considered. Artificial neural networks, support vector machines, random forests, and linear regression with location parameters estimated via least squares, or the shrinkage methods Ridge and Lasso were used. The predictive performance (PP) was assessed through crossvalidation using the following error functions: square root of mean square error (RMSE) and mean absolute error (MAE). It was found that an artificial neural network with a single hidden layer and the AM partial record, milking interval, parity and days in milk as input variables had the best PP (RMSE=1.5042, MAE=1.1389), but in general, the performance of the methods was similar. All machines whose parameters were learned using local data outperformed the reference method and the morning partial records showed a better PP than those from the afternoon. These results permit guiding ASOCEBU's milk control program and generate a "tailormade" method to predict total daily milk yield of Gyr cattle in Colombia, a relevant component of the genetic improvement and productivity modelling programs of this breed.

***Keywords*:** machine learning, milk control, cross validation.

1. Introducción

La raza bovina Gyr es un recurso genético de gran relevancia en sistemas de producción de leche en el trópico bajo colombiano (Ferro et al., 2022). En estas

ganaderías es de suma importancia medir el rendimiento individual, siendo la producción diaria de leche una de las variables de mayor interés. En el país, lo más frecuente es realizar uno o dos ordeños diarios; en este último caso, uno se realiza en la mañana y otro en la tarde. La recolección de datos se logra a través de los denominados controles lecheros, los cuales siguen las pautas marcadas por el International Committee for Animal Recording–ICAR (ICAR, 2016). Este procedimiento busca garantizar la imparcialidad en la toma de datos y se basa en visitas de personal técnico a las fincas para registrar la producción de leche y tomar muestras para determinar su composición (por ejemplo, porcentaje de grasa y proteína). En ganaderías que realizan dos ordeños, hacer un solo control y emplear este registro para predecir la producción total del día brinda la oportunidad de reducir los costos del control lechero de manera sustancial.

Otras variables zootécnicas resultan importantes en este problema de predicción. Por razones fisiológicas, el lapso o intervalo entre ordeños (IEO) afecta la producción de leche, a mayor lapso se espera una mayor producción; también deben considerarse los días en leche (Berry et al., 2007). Incluir estas variables hace que las ecuaciones de predicción de producción sean más confiables (Cerón Muñoz et al., 2017).

Uno de los métodos de referencia en la actualidad (ICAR, 2016) se basa en el trabajo de Delorenzo and Wiggans (1986); en el que, además de los registros parciales, se usan los días en leche y el intervalo entre ordeños como variables de entrada y corresponde a un modelo lineal; en adelante, este será denotado como DW. La ecuación de predicción generada es ampliamente utilizada en diferentes países del mundo, por esto fue considerada como método de referencia en este artículo. Sin embargo, esta fue desarrollada empleando datos de una población de vacas Holstein en Estados Unidos, que es distinta a la población de ganado Gyr en Colombia, ya que se trata de otra raza y las condiciones medioambientales y de manejo zootécnico son muy diferentes. Existe otra propuesta similar desarrollada por Liu et al. (2000), que se basó en el método DW, realizando algunas extensiones al incorporar medias heterogéneas a través de intervalos entre ordeños y el número de partos.

Es así como nace la pregunta sobre el desempeño del método DW en rebaños con condiciones productivas distintas a aquellas bajo las cuales fue desarrollado. Por lo tanto, se requiere desarrollar un método de predicción a la medida para la raza Gyr en el trópico bajo colombiano.

Por otro lado, existe interés en el desempeño del registro de la mañana y el de la tarde, ya que el programa de control lechero debería enfocarse en aquel con mayor capacidad predictiva. Estudios previos como el conducido por Hargrove and Gilbert (1984) muestran las diferencias entre los ordeños de la mañana con respecto a los ordeños de la tarde y su relevancia en la estimación de la producción diaria. Varios autores han reportado una mejor habilidad predictiva del registro de la mañana (Liu et al., 2000; Lee and Min, 2013; Rodríguez Neira et al., 2013; Cerón Muñoz et al., 2017).

La Asociación Colombiana de Criadores de Ganado Cebú–ASOCEBU, lleva más

de una década ejecutando un programa de control lechero en la raza Gyr y tiene interés en desarrollar una ecuación para predecir la producción total diaria de leche empleando mediciones de producción parciales y, en particular, responder dos preguntas: 1) ¿puede el método DW ser superado por métodos desarrollados a nivel local?, 2) ¿cuál de los dos registros parciales (AM o PM) tiene un mejor desempeño?

En Colombia, Rodríguez Neira et al. (2013), realizaron una investigación para plantear un modelo de predicción para la producción de leche y el porcentaje de grasa a partir de producciones parciales (PP). Esta investigación se llevó a cabo en tres hatos lecheros de ganado Holstein del departamento de Antioquia. En este estudio se tuvo en cuenta el número de partos, el intervalo entre ordeños, y los análisis se basaron en modelos de regresión lineal múltiple. Similarmente, Cerón Muñoz et al. (2017) desarrollaron ecuaciones de predicción de producción total diaria de leche y de porcentaje de grasa y proteína para la raza Holstein en zonas de trópico alto del departamento de Antioquia. Con excepción de este trabajo, los demás basaron sus análisis enteramente en modelos lineales y en ninguno de los estudios consultados se validó la capacidad de generalización de las ecuaciones desarrolladas empleando técnicas apropiadas como la validación cruzada Hastie et al. (2009). En la actualidad se dispone de una gran cantidad de aproximaciones para desarrollar métodos predictivos (Bishop, 2006; James et al., 2013) que pueden aplicarse en este problema particular. Finalmente, cabe destacar que, hasta el momento, no se han realizado estudios de este tipo para la raza Gyr en el país.

Por lo tanto, el objetivo de este estudio fue comparar el desempeño predictivo de diferentes máquinas correspondientes a combinaciones de métodos y variables de entrada, para construir una ecuación de predicción de la producción total diaria de leche en bovinos Gyr en el trópico bajo colombiano y establecer cuál de los dos ordeños parciales tiene mayor capacidad predictiva para orientar así los programas de control lechero.

2. Materiales y métodos

2.1. Datos

Los datos que se utilizaron en esta investigación fueron proporcionados por ASOCEBU (Asociación Colombiana de Criadores de Ganado Cebú). Estos provienen de 48 haciendas situadas en el trópico bajo colombiano, con un total de 13806 registros de producción de leche, pertenecientes a 3862 vacas.

Estos registros fueron recolectados por profesionales entrenados, que pertenecen al departamento técnico de ASOCEBU. Además, la base de datos cuenta con las siguientes variables: la identificación del animal; número de partos de la vaca (NP), días en leche (DEL) que corresponde a los días transcurridos entre el parto y la toma del registro de producción, producción de leche (lt) en el ordeño de la mañana, (PPAM), producción de leche en el ordeño de la tarde (PPPM), y la hora

de cada ordeño, que permite calcular el intervalo entre ordeños para el registro de la mañana (IEOAM) y en la tarde (IEOPM). Para la edición de estos datos se eliminaron los registros de producción de leche con DEL mayor a 450 días, con registros parciales menores a 0.5 lt y producciones diarias totales mayores a 60 lt. Similarmente, solo se consideraron animales de hasta 7 partos y se eliminaron categorías de número de parto con menos de tres individuos.

2.2. Análisis de datos

La variable de salida o respuesta fue la producción total del día (PT); que correspondió a la suma de los dos registros parciales, mientras que las variables de entrada o explicativas fueron los registros PPAM o PPPM, el IEOAM o el IEOPM, el NP y los DEL. Es importante considerar que, si bien se mencionaron seis variables de entrada, el número efectivo es cuatro porque solamente se considera uno de los registros parciales y su respectivo intervalo entre ordeños.

Ahora bien, en la literatura se encuentra una amplia gama de modelos y métodos para llevar a cabo análisis basados en aprendizaje supervisado. Así, hay que ser cuidadoso al elegir los modelos y métodos a emplear porque se puede llegar a tener un número muy grande de opciones que llevan a una situación en la que el análisis puede llegar a ser muy dispendioso; además, en este estudio el tamaño de muestra fue relativamente grande y esto aumentó el requerimiento computacional. Teniendo en cuenta estas características se consideraron los siguientes modelos: el propuesto por Delorenzo and Wiggans (1986), que se basa en un modelo lineal, redes neuronales artificiales (RNA), bosques aleatorios (BA), máquinas de soporte vectorial (MSV) y regresión lineal con tres métodos para estimar los parámetros de localización: mínimos cuadrados ordinarios (MCO), Ridge y Lasso. Es importante considerar que algunas de estas aproximaciones corresponden a diferentes modelos, mientras que otras corresponden a diferentes métodos para estimar los parámetros del mismo modelo. Por ejemplo, Ridge, Lasso y MCO son tres métodos para estimar los parámetros de localización del modelo de regresión lineal múltiple, simplemente Lasso y Ridge añaden penalizaciones a la función objetivo de MCO, una de tipo l2 para Ridge y l1 para Lasso, (Hastie et al., 2009).

Cabe notar que, en este trabajo, el término máquina predictiva hace referencia a la combinación de un modelo/método y un conjunto particular de variables de entrada; no es lo mismo la máquina definida por una RNA con dos variables de entrada que aquella definida por una RNA con cuatro variables de entrada.

Se consideraron RNA regularizadas con una sola capa oculta y se trató el número de unidades en dicha capa como un parámetro de sintonización junto con el parámetro de penalización de la función objetivo denotado por WD. La función de activación fue la logística y la función de salida fue la función identidad. Por otro lado, las máquinas de soporte vectorial consideraron kernel de dos tipos: lineal y radial, el parámetro de sintonización fue el costo. Los bosques aleatorios tuvieron como parámetros de sintonización el número de árboles y el número de variables que se muestrean al azar para aumentar el número de ramas del árbol

(mtry). Finalmente, respecto a los modelos de regresión lineal, cuando se emplean los mínimos cuadrados ordinarios (MCO) como método de estimación, no se tienen parámetros de sintonización, en tanto que los métodos de encogimiento presentan un único parámetro de este tipo llamado el parámetro de penalización. Cabe resaltar que Lasso y Ridge se emplean con modelos de dos o más variables de entrada. La Tabla 1 presenta los valores que se usaron para los parámetros de sintonización de cada modelo/método; cuando se tuvieron dos parámetros (RNA y BA), el proceso de sintonización se llevó a cabo con todas las combinaciones de los valores de las cuadrículas individuales que aparecen en la Tabla 1. La sintonización se llevó a cabo mediante validación cruzada interna en el conjunto de entrenamiento de cada partición como se describe en Hastie et al. (2009).

Modelo/Método	Parámetro y valores considerados.
RNA	Número de unidades en la capa oculta:{1,2,...,6} WD:{0,0.2,0.4,...,1}
BA	Número de árboles:{50,150,...,450,500} mtry:{1,2,3}
LASSO	Parámetro de Penalización:Secuencia de tamaño 500 desde 0.0001 hasta 100
RIDGE	Parámetro de Penalización:Secuencia de tamaño 500 desde 0.0001 hasta 100
RLMCO	NA
MSV	Costo:{0.001,0.005,...,20}
DW	NA

Tabla 1: Conjuntos de valores considerados para cada uno de los parámetros de sintonización.

Debido a que en este problema se tienen solamente cuatro variables de entrada, fue posible considerar modelos basados en todos los subconjuntos de tamaño 1, 2, 3 y 4. Además, los conjuntos que contenían registros parciales o intervalos entre ordeños tenían dos versiones, una con el registro de la mañana y otra con el registro de la tarde.

Asimismo, considerando que para los intervalos entre ordeños y registros parciales se tenían valores tomados en la mañana y en la tarde y que nunca se combinan en el mismo grupo PPAM con IEOPM ni PPPM con IEOAM, se tuvieron en total 27 conjuntos de variables de entrada distribuidas así: seis grupos conformados por una sola variable, 11 de dos variables, ocho de tres y dos de cuatro.

La evaluación del desempeño predictivo se llevó a cabo mediante validación cruzada en diez partes utilizando la raíz cuadrada del error cuadrático medio (RECM) y el error absoluto medio (EAM) como funciones de error. Además, también se calculó el estimador de validación cruzada de la correlación entre el valor predicho y el observado, esto es, se obtuvo esta correlación en el conjunto de validación en cada una de las particiones y luego se obtuvo la media aritmética de los diez valores resultantes. Para un conjunto de validación de tamaño k las funciones de error RECM y EAM se calculan como sigue:

$$RECM = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}$$

$$EAM = \frac{1}{k}\sum_{i=1}^{k}|y_i - \hat{y}_i|.$$

El primer criterio que se tuvo en cuenta para seleccionar un modelo fue RECM, mientras que el EAM se utilizó para decidir entre modelos con valores de RECM muy cercanos. Además, para el caso de modelos con funciones de error muy similares, CP se empleó como criterio para decidir. La razón para dar prioridad a RECM es que combina exactitud y precisión, y un buen modelo predictivo debe balancear estos dos atributos (Hastie et al., 2009). También se calculó la razón entre RECM y la media general de PT (error cuadrático medio relativo ECMR) tanto para los modelos con registros AM (ECMRAM) como aquellos basados en registros PM (ECMRPM) y la razón entre RECMAM y RECMPM denominada como desempeño predictivo relativo (RDP). Siguiendo la escala empleada por Lee and Min (2013), la capacidad predictiva se categorizó según ECMR así: buena cuando el ECMR ≤ 0.2, regular si ECM se encuentra en el intervalo (0.2, 0.3) y pobre si ECMR ≥ 0.3.

La edición y análisis de datos se llevó a cabo utilizando las librerias: ggplot2, lattice, caret, randomforest, gtools, glmnet, matrix, del programa R (R Core Team, 2021).

3. Resultados

La Tabla 2 muestra algunas estadísticas descriptivas para las variables de entrada y la variable de salida. Como puede verse, la mayor variabilidad, cuantificada mediante el coeficiente de variación, la exhibió DEL, mientras que IEOPM presentó la menor.

Variable	Media	Mediana	Min.	Max.	DE	CV	n
NP	3.0236	3.0000	1.0000	9.0000	1.7906	59.2220	13806
DEL	116.6251	104.0000	5.0000	441.0000	77.0900	66.1007	13806
PPAM	7.4405	7.2000	1.0000	25.2000	2.7220	36.5835	13806
PPPM	4.7326	4.3000	0.5000	17.8000	2.0846	44.0474	13806
IEOAM	9.9055	10.0000	6.5000	13.0000	0.8855	8.9391	13806
IEOPM	14.0945	14.0000	11.0000	17.5000	0.8855	6.2824	13806
PT	12.1731	11.6000	1.5000	33.3000	4.3965	36.1161	13806

Tabla 2: Estadísticas descriptivas de las variables de entrada y la variable respuesta.

Por otro lado, la Tabla 3 resume el desempeño predictivo de diferentes máquinas. Para cada modelo o método, se presenta la combinación de variables de entrada con el mejor desempeño (menor RECM, menor EAM, y mayor CP) junto con

los resultados de aquella que consideró el registro parcial (PPAM o PPPM) y su respectivo intervalo entre ordeños. La regresión Lasso es un método que hace selección de variables de manera automática y por ello, este método se corrió una sola vez considerando las cuatro variables de entrada; por esta razón, en la tabla 3 se presenta un único valor. Como se puede apreciar, en todos los casos, el mejor desempeño se obtuvo con el conjunto completo de cuatro variables.

Modelo/ método	Variables de entrada	CPAM	RECMAM	MAEAM	CPPM	RECMPM	MAEPM
RIDGE	PP, IEO	0.9358	1.5560	1.1823	0.8876	2.0292	1.5470
RIDGE	NP, DEL, PP, IEO	0.9369	1.5438	1.1753	0.8924	1.9879	1.5035
RNA	PP, IEO	0.9372	1.5401	1.1702	0.8941	1.9739	1.4991
RNA	NP, DEL, PP, IEO	0.9402	1.5042	1.1389	0.9010	1.9116	1.4377
BA	PP, IEO	0.9189	1.7668	1.3177	0.8735	2.1668	1.6602
BA	NP, DEL, PP, IEO	0.9367	1.5480	1.1710	0.8964	1.9547	1.4766
MSV	PP, IEO	0.9358	1.5593	1.1791	0.8876	2.0412	1.5412
MSV	NP, DEL, PP, IEO	0.9368	1.5487	1.1722	0.8922	1.9999	1.4993
MSV RA	PP, IEO	0.9374	1.5408	1.1628	0.8944	1.9821	1.4784
MSV RA	NP, DEL, PP, IEO	0.9398	1.5116	1.1295	0.9001	1.9258	1.4297
RLMCO	PP, IEO	0.9358	1.5560	1.1823	0.8876	2.0292	1.5470
RLMCO	NP, DEL, PP, IEO	0.9369	1.5438	1.1753	0.8924	1.9879	1.5035
LASSO	NP, DEL, PP, IEO	0.9369	1.5438	1.1753	0.8924	1.9879	1.5035
DW	PP, IEO, DEL	0.9153	2.1070	1.5314	0.8752	2.6576	1.9979

CPAM: Correlación Predictiva AM; RECMAM: Raíz del error cuadrático medio AM
MAEAM: Error absoluto medio AM; CPPM: Correlación Predictiva PM
RECMPM: Raíz del error cuadrático medio PM; MAEPM: Error absoluto medio PM
1. En cada caso se presenta la combinación de variables de entrada con el mejor desempeño y la máquina basada solamente en el registro parcial y su respectivo intervalo entre ordeños.

Tabla 3: Desempeño predictivo de algunas de las máquinas consideradas[1].

El siguiente comportamiento se presentó tanto para las maquinas que consideraron registros parciales de la mañana como las que consideraron los de la tarde. Teniendo en cuenta el RECM, la red neuronal basada en las cuatro variables de entrada presentó el mejor desempeño predictivo, seguida muy de cerca por la máquina de soporte vectorial con kernel radial. No obstante, según el EAM, esta última presentó el mejor desempeño, pero nuevamente, las diferencias no fueron grandes. Finalmente, el criterio CP favoreció a la red neuronal con las cuatro variables sobe la máquina de soporte vectorial con kernel radial, también basada en las cuatro variables explicativas. Por lo tanto, el mejor desempeño bajo dos de los tres criterios (RECM y CP) fue alcanzado por la red neuronal artificial que incluyó PPAM, IEOAM, NP y DEL como variables de entrada. Los resultados para todas las máquinas consideradas se presentan en el material suplementario.

Además, la Tabla 3 también muestra que el desempeño predictivo de los diferentes métodos fue similar, observándose que todas las máquinas desarrolladas con los datos obtenidos de la población de interés mostraron un desempeño superior a la ecuación de referencia (DW) según RECM y EAM. Bajo el criterio CP se presentó una excepción en el caso del bosque aleatorio que empleó PPPM e IEOPM como variables de entrada, pues este presentó un valor de CP levemente menor al obtenido con el método DW.

Por otro lado, según la escala descrita previamente, la capacidad predictiva de

las máquinas relacionadas en la Tabla 4 es buena excepto por el método DW empleando registros de la tarde, la cual tiene una capacidad predictiva regular. También cabe notar que el registro parcial de la mañana y su respectivo intervalo entre ordeños presentaron un mejor desempeño que los registros de la tarde con un RDP que vario entre 0.7668 y 0.7928 (Tabla 4) indicando menores valores de RECM.

Modelo/método	Variables de entrada	RDP	ECMRAM	ECMRPM
RIDGE	PP, IEO	0.7668	0.1280	0.1669
RIDGE	NP, DEL, PP, IEO	0.7766	0.1270	0.1635
RNA	PP, IEO	0.7802	0.1267	0.1624
RNA	NP, DEL, PP, IEO	0.7869	0.1237	0.1573
BA	PP, IEO	0.8154	0.1453	0.1783
BA	NP, DEL, PP, IEO	0.7919	0.1273	0.1608
MSV	PP, IEO	0.7639	0.1283	0.1679
MSV	NP, DEL, PP, IEO	0.7744	0.1274	0.1645
MSV RA	PP, IEO	0.7773	0.1268	0.1631
MSV RA	NP, DEL, PP, IEO	0.7849	0.1244	0.1584
RLMCO	PP, IEO	0.7668	0.1280	0.1669
RLMCO	NP, DEL, PP, IEO	0.7766	0.1270	0.1635
LASSO	NP, DEL, PP, IEO	0.7766	0.1270	0.1635
DW	PP, IEO, DEL	0.7928	0.1733	0.2186

RDP: Razón de desempeño predictivo

ECMRAM: Error cuadrático medio relativo AM

ECMRPM: Error cuadrático medio relativo PM

1. En cada caso se presenta la combinación de variables de entrada con el mejor desempeño y la máquina basada solamente en el registro parcial y su respectivo intervalo entre ordeños.

Tabla 4: Razón de desempeño predictivo y raíz del error cuadrático medio relativo para cada una de las técnicas consideradas[1].

4. Discusión

Los resultados indican que las producciones parciales fueron las variables de entrada que más aportaron a la capacidad predictiva, ya que, bajo todos los modelos o métodos considerados, las máquinas que incluían los registros parciales mostraron superioridad; de hecho, para aquellas máquinas que consideraron una sola variable de entrada, cuando se empleó PPAM o PPPM, los valores de RECM, MAE y CP fueron cercanos a aquellos obtenidos con el conjunto completo de variables explicativas (ver material suplementario). En todos los casos, agregar variables diferentes a la producción parcial de leche (AM o PM) no generó aumentos marcados en capacidad predictiva. Sin embargo, el conjunto de las cuatro variables de entrada siempre mostró el mejor desempeño predictivo; además, los resultados obtenidos con el método Lasso soportan esta idea ya que todas las variables fueron selec-

cionadas. Sumado a esto, si bien el método Ridge no hace selección de variables, este induce un encogimiento en los estimadores que es gobernado por el parámetro de penalización (James et al., 2013); durante el proceso de sintonización siempre eligieron valores muy cercanos a cero, fenómeno también observado para el Lasso y, en consecuencia, los resultados para MCO, Ridge (con las 4 variables) y Lasso fueron casi iguales (Tabla 3).

Por lo tanto, los resultados del estudio soportan la idea de que la predicción parcial de leche es la principal variable para predecir la producción total del día tal y como lo han reportado otros autores (Cassandro et al., 1995; Liu et al., 2000); además, en el caso de contenidos de grasa y proteína, Cerón Muñoz et al. (2017) encontraron que los modelos con mejor desempeño incluían únicamente el registro parcial como variable predictora. Por otro lado, Liu et al. (2000) consideraron una máquina muy sencilla en la cual la producción total del día se predice como el doble de la producción parcial. Si bien esta regla desconoce que generalmente la producción de la mañana es mayor, los resultados encontrados por estos autores mostraron que, aunque su desempeño fue inferior al de los demás modelos lineales considerados, las diferencias no fueron grandes e incluso, en algunos casos esta ecuación predictiva ingenua equiparó al método DW.

También se encontró que el registro parcial de la mañana predice mejor la producción total del día que el de la tarde (Tablas 3 y 4), resultado que ha sido reportado en estudios que consideraron diferentes razas y diferentes poblaciones (Liu et al., 2000; Lee and Min, 2013; Rodríguez Neira et al., 2013; Cerón Muñoz et al., 2017). Este fenómeno puede ser explicado, al menos parcialmente, por el hecho de que generalmente la producción de la mañana es mayor que la producción de la tarde (Delorenzo and Wiggans, 1986; Lee and Min, 2013); en consecuencia, este registro es más cercano a la variable que se quiere predecir; Liu et al. (2000) reportaron correlaciones entre registro parcial y producción total del día levemente superiores para la producción de la mañana. En este estudio, la media de los registros de la mañana fue 57.21 % mayor a la media de los registros de la tarde (Tabla 2). En contraste, Lee and Min (2013) encontraron que, para ciertos intervalos entre ordeño y números de parto, el registro PM generó mayores exactitudes.

Estudios preliminares que han investigado el mismo problema (Delorenzo and Wiggans, 1986; Cassandro et al., 1995; Liu et al., 2000; Lee and Min, 2013; Rodríguez Neira et al., 2013; Cerón Muñoz et al., 2017), lo han abordado desde un punto de vista diferente debido a que no se reconoció completamente su naturaleza. Se trata de un problema de predicción y no de un problema de pruebas de hipótesis o estimación de parámetros; una distinción que hasta hace relativamente poco tiempo no se hacía. La diferencia entre la búsqueda de variables significativamente asociadas con una variable respuesta y variables con alta capacidad para predecirla se discute en Lo et al. (2015), quienes motivados por problemas de genética-estadística, estudiaron un fenómeno relevante en diferentes ramas de la ciencia: el hecho de que las variables estadísticamente significativas (existiendo muchas formas para declararlas como tal), no necesariamente son buenos predictores. Esto se debe a que, desde el punto de vista estadístico, no se trata del mismo

problema y así, las variables explicativas que se declaran como estadísticamente significativas no siempre conforman el modelo con el mejor desempeño predictivo. Por consiguiente, el problema que se estudia en este trabajo implica evaluar la capacidad para predecir la producción total del día empleando valores de las variables de entrada que no han sido "vistos" por la máquina; por lo tanto, lo relevante no es juzgar la bondad de ajuste a los datos de entrenamiento sino el error de generalización, siendo la validación cruzada una forma relativamente sencilla de estimarlo (Bishop, 2006; Hastie et al., 2009).

Es por esto que, la búsqueda de métodos para predecir la producción diaria de leche empleando registros parciales y otras variables zootécnicas no debe enfocarse en los análisis que juzgan bondad de ajuste a la base de datos con la que se entrena el modelo, sino en aquellos empleados para optimizar el error fuera de la muestra. La mayoría de los estudios encontrados en la literatura emplearon pruebas de hipótesis, y criterios como el coeficiente de determinación y confiabilidad para comparar modelos y evaluar su capacidad predictiva. El problema es que al utilizar los mismos registros que se emplearon para entrenar las máquinas, su capacidad predictiva parecerá mayor de lo que en realidad es, fenómeno denominado optimismo (Hastie et al., 2009); además, se debe evitar el sobreajuste en el conjunto de datos de entrenamiento; por ejemplo, cuando se tiene un coeficiente de determinación demasiado alto, porque en estos casos el error de generalización tiende a ser muy alto (James et al., 2013). De los estudios consultados, solamente el conducido por Liu et al. (2000) consideró una validación del modelo predictivo en una base de datos independiente, que corresponde a una partición sencilla de la base de datos en conjuntos de entrenamiento y validación.

Este es el primer estudio de predicción de la producción diaria de leche total para bovinos Gyr en el país. Además de su importancia para los criadores de la raza, existen otros aportes relevantes para el sector que vale la pena mencionar. Aquí se evidencia la importancia de desarrollar ecuaciones de predicción de producción total diaria de leche a la medida, esto es, empleando registros de individuos pertenecientes a la población de interés. Por otro lado, en aspectos metodológicos, se muestra la utilidad y necesidad de abordar este problema de interés zootécnico desde el punto de vista predictivo mediante métodos de aprendizaje supervisado que son más apropiados al reconocer la naturaleza del problema. Sumado a esto, con excepciones como Lee and Min (2013) y Cerón Muñoz et al. (2017), los estudios preliminares consideraron solamente modelos lineales; los resultados de este trabajo ilustran el buen desempeño predictivo de los modelos no lineales. Además, con la excepción de los tradicionales modelos lineales, ninguno de los estudios encontrados en la literatura usó las metodologías empleadas en aprendizaje automático, las cuales mostraron superioridad en este trabajo. Sin embargo, algunos autores han implementado estas aproximaciones en otros problemas de predicción de producción de leche. Por ejemplo, Grzesiak et al. (2003) estudiaron problemas de predicción que tienen que ver con la proyección de curvas de lactancia. Estos autores compararon el desempeño predictivo de redes neuronales artificiales y regresión lineal múltiple para predecir la producción total de leche hasta los 305 DEL encontrando desempeños similares. Así, se espera que este trabajo aporte al sector

lechero al mostrar la utilidad de la implementación de los métodos de aprendizaje de máquina para construir ecuaciones de predicción.

Los resultados de este estudio indican que se puede predecir la producción diaria de leche con alta confiabilidad empleando registros parciales, IEO, DEL y NP; además, el registro de producción de la mañana tiene un poder predictivo mayor que el de la tarde. En consecuencia, cuando se requiera disminuir costos, el programa de control lechero de ASOCEBU puede diseñarse para obtener solamente la producción de leche de la mañana y predecir la producción total del día utilizando la red neuronal artificial que se entrenó en este estudio.

5. Conclusiones

Es importante que los investigadores del área noten que este es un problema de predicción y no un problema de inferencia; por lo tanto, el abordaje estadístico que tiene relevancia cuando se van a llevar a cabo pruebas de hipótesis o estimación de parámetros, no es el más apropiado para evaluar la capacidad predictiva de un modelo o método dado; no prima el ajuste a los datos de entrenamiento sino el error de generalización.

Los resultados de este trabajo permiten asistir la toma de decisiones para el direccionamiento del programa de control lechero de ASOCEBU y generan el primer modelo predictivo de la producción total de leche del día para la raza Gyr en el país.

Otra contribución de este trabajo es mostrar la utilidad de los métodos de aprendizaje supervisado en el estudio de este problema de relevancia zootécnica.

Recibido: mayo 2022
Aceptado: junio 2022

Referencias

D. P. Berry, F. Buckley, and P. Dillon. Body condition score and live-weight effects on milk production in irish holstein-friesian dairy cows. *Animal*, 1(9):1351–1359, 2007.

C. M. Bishop. Linear models for classification. *Pattern recognition and machine learning*, pages 179–224, 2006.

M. Cassandro, P. Carnier, L. Gallo, R. Mantovani, B. Contiero, G. Bittante, and G. Jansen. Bias and accuracy of single milking testing schemes to estimate daily and lactation milk yield. *Journal of dairy science*, 78(12):2884–2893, 1995.

M. F. Cerón Muñoz, J. D. Corrales Álvarez, and J. P. Ramírez Arias. Predicción de la producción de leche, porcentaje de grasa y proteína diaria a partir de

registros del ordeño de la mañana o de la tarde en vacas holstein en pastoreo. *Livestock Research for Rural Development*, 29(9):166, 2017.

M. A. Delorenzo and G. R. Wiggans. Factors for estimating daily yield of milk, fat, and protein from a single milking for herds milked twice a day. *Journal of Dairy Science*, 69(9):2386–2394, 1986.

D. Ferro, J. Gil, A. Jiménez, C. Manrique, and C. A. Martínez. Estimation of lactation curves of gyr cattle and some associated production parameters in the colombian low tropic. *Revista Colombiana de Ciencias Pecuarias*, 35(1), 2022.

W. Grzesiak, R. Lacroix, J. Wójcik, and P. Blaszczyk. A comparison of neural network and multiple regression predictions for 305-day lactation yield using partial lactation records. *Canadian Journal of Animal Science*, 83(2):307–310, 2003.

G. L. Hargrove and G. R. Gilbert. Differences in morning and evening sample milkings and adjustment to daily weights and percents. *Journal of Dairy Science*, 67(1):194–200, 1984.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

ICAR. Knowledge Management in Agriculture. *THE GLOBAL STANDARD FOR LIVESTOCK DATA*, 2016. URL `http://www.icar.org.in/en/information-resources.htm`.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

D. Lee and H. Min. Estimation of daily milk yields from am/pm milking records. *Journal of Animal Science and Technology*, 55(6):489–500, 2013.

Z. Liu, R. Reents, F. Reinhardt, and K. Kuwan. Approaches to estimating daily yield from single milk testing schemes and use of am-pm records in test-day model genetic evaluation in dairy cattle. *Journal of Dairy Science*, 83(11): 2672–2682, 2000.

A. Lo, H. Chernoff, T. Zheng, and S.-H. Lo. Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences*, 112(45):13892–13897, 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

J. D. Rodríguez Neira, G. A. Correa Londoño, and J. J. Echeverri Zuluaga. Prediction models for total milk yield and fat percentage using partial samples. *Revista Facultad Nacional de Agronomía Medellín*, 66(1):6909–6917, 2013.