



Detección de mastitis subclínica en vacas lecheras por modelos de regresión lineal y algoritmos de inteligencia artificial, San Carlos, Costa Rica

Subclinical mastitis detection in dairy cows using linear regression models and artificial intelligence algorithms in San Carlos, Costa Rica

Verny Estrada-Carvajal¹, Marlen Camacho Calvo², Rafael Molina Montero³, Wilfrido Paniagua Madrigal⁴

Palabras clave

Bovinos, células somáticas, mastitis, aprendizaje automatizado, modelos de regresión, sensibilidad, especificidad.

Resumen

Se compararon seis modelos de regresión lineal y algoritmos de inteligencia artificial para la predicción de mastitis subclínica en el hato de ordeño de la finca Ganadera La Uno S. A. Como variable de respuesta se utilizó una transformación de la variable "conteo de células somáticas" (CCS), denominada *score* de células somáticas (SCCS). Como variables predictoras de SCCS se utilizaron la conductividad eléctrica (CE), la producción de leche (PL) e indicadores técnicos de las vacas como número de partos (NP), días en lactación (DEL), edad (ED) y estado reproductivo (ER). Se definió como "vacas enfermas" aquella con un *score* de células somáticas (SCCS) de 4 o más, el cual representa un conteo de células somáticas (CCS) mayor a 200 000 cs/ml. Se desarrollaron varios modelos de regresión lineal por la técnica de modelos lineales y se usaron dos técnicas de regresión de *machine learning*: el algoritmo de clasificación de Naïve Bayes y un modelo de regresión desarrollado por aprendizaje automatizado. Los modelos de regresión lineal desarrollados por la técnica de modelos lineales mostraron valores de R^2 por debajo de 0,32, con un valor de predicción deficiente en términos de especificidad (ESP) y de sensibilidad (SEN), con ESP de 55% (a una SEN fijada en 80%), mientras que el área bajo la curva ROC de estos modelos se aproximó a 75%. El algoritmo de clasificación de Naïve Bayes no fue un buen estimador de mastitis subclínica, con valores de SEN y ESP similares a los modelos de regresión lineal mencionados anteriormente. El mejor predictor fue el modelo de regresión lineal que aplicó *machine learning*, con ESP (a 80% de SEN) y áreas bajo la curva ROC de 84,3% y 92,1% respectivamente. Se demostró que existe un gran potencial en los algoritmos de aprendizaje automático para la detección de mastitis subclínica en hatos lecheros.

Abstract

Six linear regression models and artificial intelligence algorithms for the prediction of subclinical mastitis in the milking herd of Ganadera La Uno S.A. farm were compared. As a response variable, a transformation of the "somatic cell count" (CCS) was used (somatic cell score, SCCS). As predictive variables of SCCS, there were used: 1) electrical conductivity (EC), 2) milk production (PL), and 3) technical indicators of cows, such as number of births (NP), days in lactation (DEL), age (ED) and reproductive status (ER). A "sick cow" was defined as one with a somatic cell score (SCCS) of 4 or more, which represents a somatic cell count (CCS) greater than 200,000 cs/ml. Four linear regression models were developed and two machine learning regression techniques were used: A) The Naïve Bayes classification algorithm and B) a regression model developed by automated learning. Linear regression models developed showed R^2 values below 0.32, with a poor prediction value in terms of specificity (SPE) and sensitivity (SEN), with SPE of 55% (SEN set at 80%); while the area under the ROC curve of these models approached 75%. The classification algorithm of Naïve Bayes was not a better estimator of subclinical mastitis either, with values of SEN and SPE similar to the linear regression models mentioned above. The best predictor was the linear regression model that machine learning applied, with SPE (at 80% SEN) and areas under the ROC curve of 84.3% and 92.1% respectively. It was shown that there is great potential in machine learning algorithms for the detection of subclinical mastitis in dairy herds.

Key words

Cattle, somatic cells, mastitis, machine learning, regression models, sensitivity, specificity.

1 Ing. Agr, especialista en Agricultura de precisión, ✉ag.bestrada@gmail.com
2 Docente, ITCR, mcamacho@tec.ac.cr
3 Médico veterinario, Cooperativa de Productores de Leche Dos Pinos R.L., ramolina@dospinos.com
4 Docente e investigador, ITCR, wpaniagua@tec.ac.cr

Recibido: 5 de abril del 2019
Aceptado: 10 de julio del 2019
Publicado: 15 de setiembre del 2019
DOI: 10.18860/rath.v211.4689

Introducción

La mastitis es una reacción inflamatoria clínica o subclínica del tejido mamario debido al daño producido por toxinas e irritantes liberados por bacterias que colonizan la ubre [1], la cual conlleva a un incremento en la producción de glóbulos blancos (células somáticas) que erradican los irritantes, reparan el tejido dañado y recuperan la ubre infectada. Incluso en granjas con rigurosos programas de control de infección por mastitis, este padecimiento se caracteriza por presentar niveles de incidencia alta [2]-[4], lo que repercute en una reducción en la producción de leche diaria, que alcanza hasta el 75% del total por vaca [5], [6].

La mastitis subclínica se cataloga como uno de los principales rubros de pérdidas en fincas lecheras [3], [5], [7]-[8] y compromete seriamente el bienestar animal; es también la enfermedad que más influye en los costos de producción de leche bovina [3], [9]-[11].

La detección temprana de enfermedades como la mastitis puede incidir en la inocuidad de la leche y en el bienestar animal [12]. Los métodos más comunes para realizar esta detección temprana son la prueba California para mastitis (CMT por sus siglas en inglés), el conteo de células somáticas (CCS) y el cultivo de bacterias [13]-[16]. Sin embargo, estos métodos son costosos y demandantes de tiempo por lo cual son evitados a menudo en granjas lecheras.

La incorporación de sistemas de ordeño automatizados reduce drásticamente la interacción del operario con el animal, lo que hace necesario contar con sistemas de detección en línea que utilicen parámetros de fácil medición *in situ* para detectar mastitis [17]. Entre las variables utilizadas destacan la medición de la conductividad eléctrica de la leche (CE), la producción de leche y otras características propias de cada animal como la cantidad de partos y la fase de lactancia [16]-[19], siendo la CE la más utilizada, puesto que un conteo alto coincide con una mayor concentración de iones (Na^+ , K^+ y Cl^-) en disolución [16].

Diversos estudios han intentado predecir la mastitis en función de la conductividad eléctrica

por medio de modelos clásicos de regresión [17], [20]-[22]; sin embargo, la mayoría de estos modelos muestran una utilidad cuestionable debido a que esta relación (mastitis – CE) es compleja y no siempre es lineal. Como contraparte, se han utilizado diversas técnicas de inteligencia artificial, tanto de clasificación como de predicción, para el desarrollo de modelos de detección de mastitis con resultados más prometedores [23]-[27]. Los métodos de predicción basados en algoritmos de inteligencia artificial son excelentes candidatos para el desarrollo de herramientas predictivas de mastitis principalmente porque pueden utilizar amplios y complejos bagajes de datos, son capaces de aprender por sí mismos y no dependen de los supuestos estadísticos paramétricos [28], [29].

Con base en lo anterior, este estudio se enfocó en desarrollar y comparar modelos de predicción de mastitis subclínica en un sistema de ordeño automatizado, mediante la técnica clásica de regresión lineal y dos algoritmos de inteligencia artificial.

Materiales y métodos

Descripción del sitio de estudio

El estudio se desarrolló en la finca Ganadera La Uno S.A, ubicada en La Marina de San Carlos, Costa Rica, específicamente entre las coordenadas CRTM05 458742.53 E y 1147645.93 N. Esta zona se caracteriza por la presencia de explotaciones ganaderas mayoritariamente para la producción de leche.

El periodo de estudio se extendió de setiembre a noviembre del 2017, comprendiendo once mediciones semanales efectuadas los días lunes. Durante ese periodo, la finca contó con un hato de 111 vacas en ordeño y cerca de 70 reemplazos, principalmente de la raza Holstein, Jersey, Pardo Suizo y sus cruces, con una producción promedio de 14 kg de leche/vaca/día. El hato se maneja de forma intensiva, utilizando un establo de alrededor de 10 000 m² donde se ofrece una dieta balanceada que es revisada constantemente con base en criterios productivos y reproductivos del hato con el propósito de optimizar la producción. El suministro de alimento obedece

a un ciclo mecanizado desde el establecimiento, mantenimiento, cosecha, acarreo, mezcla y distribución de forraje, junto con alimentos balanceados para completar una ración total mezclada (RTM). Además, la finca cuenta con 20 ha para el pastoreo de las vacas secas, zonas de producción de forrajes rastreros (como estrella africana) y erectos (sorgo, maíz), zonas de sesteo e infraestructura para la alimentación diaria de los reemplazos y protección del hato ante condiciones climáticas adversas. Para la limpieza y desinfección de las instalaciones, tanto de los galrones de alimentación como de la sala de ordeño, se utiliza una solución de ácido peracético (LIPERAX) que se aplica hasta dos veces por semana.

La sala de ordeño de la lechería es de tipo espina de pescado con capacidad para diez vacas por fila. Cada estación de ordeño posee una consola individual que permite controlar el ordeño y monitorear diversos parámetros relacionados con cada animal. El equipo de ordeño cuenta con un sistema automatizado (AMS por sus siglas en inglés), que permite monitorear y controlar de manera eficiente el ordeño de cada vaca. El sistema monitorea la conductividad eléctrica y la producción de leche (entre otros), emitiendo una serie de alertas que detectan alteraciones en estas variables ($\pm 12\%$)

que permiten identificar animales con posibles eventos de mastitis subclínica.

Durante el estudio se siguió el protocolo de ordeño de la finca, de manera que se realizaron dos ordeños diarios (3:00 am y 1:00 pm) con una duración de una hora aproximadamente. Una vez concluida esta actividad, la ubre de la vaca fue sellada con yodo al 0,25% y al finalizar la operación de ordeño del hato se realizó el protocolo de lavado del equipo.

Población y variables evaluadas

La población evaluada fue la totalidad de vacas en producción, las cuales presentaron edades que iban desde 17 meses hasta 154 meses, distribuidas en todas las etapas de lactación desde el día 10 postparto hasta el final de la cosecha de leche. La cantidad de partos en el hato varió de uno a ocho. En el Cuadro 1 se presentan las variables cuantificadas en este estudio.

La variable estado reproductivo (ER) fue discretizada mediante la siguiente escala: 1= vaca gestante, 2= vaca abierta (no inseminada), 3= vaca inseminada, 4= vaca no gestante y 5= vaca estéril. Además, tres de las variables cuantitativas (NP, ED y DEL) fueron multi-clasificadas con el propósito de simplificar su uso en el análisis de

Cuadro 1. Variables evaluadas en el estudio de detección de mastitis subclínica. Finca Ganadera La Uno S. A., San Carlos, Costa Rica, 2017.

Table 1. Variables evaluated. Ganadera La Uno S.A., San Carlos, Costa Rica.

Variable	Abreviatura	Unidad	Equipo de medición
Conductividad eléctrica (máx.)	CE	mS/cm	SMARTFLO™
Producción de leche	PL	kg/vaca/día	SMARTCONTROL™ METER – US
Conteo de células somáticas	CCS	cs/ml	Fossmatic Minor
Edad	ED	meses	
Días en lactación	DEL	días	
Número de partos	NP	n	Registros (SMARTDAIRY - US)
Estado reproductivo	ER	clases	

Nota: CCS fue expresado en células somáticas por mililitro (cs/ml)

los datos. La edad del animal (ED) se clasificó en cinco clases: 1= menor o igual a 24 meses, 2= de 24,1 a 48 meses, 3= de 48,1 a 72 meses, 4= de 72,1 a 96 meses y 5= 96,1 o más meses. Los días en lactación (DEL) se agruparon en cinco clases: 1= de 0 a 60 días, 2= de 61 a 120 días, 3= de 121 a 200 días, 4= de 201 a 300 días y 5= mayor a 300 días de lactación. El número de partos (NP) se agrupó en cuatro clases: 1= un parto; 2= dos partos; 3= tres y cuatro partos; 4= cinco o más partos. Se utilizó la letra "c" como subíndice para diferenciar una variable en escala original y la misma variable en escala ordinal (ER_c, NP_c, ED_c, DEL_c).

Dado que la variable de respuesta en este estudio (CCS) no presentó una distribución normal, se siguió la recomendación de diversos autores [4], [30]-[31] de normalizar su distribución por medio de una transformación logarítmica, mediante la ecuación desarrollada por [32], denominada "score de células somáticas", de manera que:

$$SCCS = \text{Log}_2(CCS / 100000) + 3$$

Metodología de toma de datos

Sin interrumpir el proceso normal de ordeño, una vez que las vacas estuvieron correctamente ubicadas en cada estación y debidamente identificadas se preparó una bolsa para muestras de leche. En cada bolsa se recolectó aproximadamente 70 ml de leche/vaca y se identificó con la fecha de muestreo, número de vaca y número de finca. Una vez que las bolsas quedaron debidamente selladas, se introdujeron en una hielera para bajar la temperatura y evitar la proliferación de microorganismos. Inmediatamente después de completar el ordeño, las muestras se trasladaron al Laboratorio de Calidad de Leche Dos Pinos, Ciudad Quesada, donde se realizó el conteo celular somático. Los resultados fueron brindados pasados tres días de cada muestreo.

Por otra parte, las mediciones de CE y PL fueron tomadas automáticamente por el equipo instalado, en tiempo real, mientras que el resto de variables (edad, fase de lactancia, número de partos y estado de la vaca) se extrajo del registro histórico de la finca.

Análisis de datos

Análisis exploratorio

La base de datos original (n=966), se sometió a un proceso de depuración en el cual se eliminaron los registros incompletos y aquellos que mostraron extremos o valores atípicos (más de 3 veces el valor del rango intercuartil para cada variable). De la base de datos depurada (n=942), se extrajo al azar el 16,6% de los registros, sobre los cuales se validaron los modelos elaborados. Con los registros restantes (83,4% de la base de datos original = 786 registros), se realizó una exploración previa de la información recabada, con el uso de la técnica de análisis de componentes principales (ACP).

Modelos de regresión clásicos

Posteriormente, se procedió a desarrollar modelos de predicción para SCCS, en función de las variables predictivas con la técnica clásica de regresión lineal. Se seleccionaron los modelos que cumplieran con los mejores criterios de ajuste (mayor R², adecuada distribución de los residuos, menores valores de los criterios de información de Akaike (AIC), Bayesiano (BIC) y error cuadrático medio de predicción (ECMP)). Se utilizó el programa estadístico InfoStat/P [33] para el desarrollo de los modelos de regresión y su evaluación.

Algoritmos de aprendizaje automático

Los algoritmos de aprendizaje automático (*machine learning* o ML) se entrenaron con el mismo conjunto de datos. En este análisis se utilizó el método de aprendizaje automático supervisado. Para este caso las "características" fueron las variables regresoras, utilizadas de forma simple o en multi-clases. Como "etiqueta" se utilizó la variable de respuesta (SCCS).

Los archivos fueron manipulados con el programa Pandas [34] y el lenguaje de programación utilizado para entrenar y probar los algoritmos de *machine learning* fue Python mediante la librería Scikit-learn [35]. Se probaron dos tipos de algoritmos de ML: i) de clasificación (MLC), denominado Naïve Bayes y ii) de regresión (MLRG), con la función "*LinearRegresion*" [36].

En el caso de MLC, el algoritmo detectó patrones que intentaron clasificar los animales en dos grupos: 0) sanas y 1) enfermas. En el marco de los MLC se entrenaron dos modelos, el primero (MLCD) utilizó como características las cuatro variables discretas (ED, DEL, NP y ER) y las dos variables continuas (PL y CE) para predecir el estado de salud de la vaca. El otro modelo que se entrenó (MLCC) contempló las mismas variables (excepto ER) en sus valores originales (no en multi-clases). Ambos modelos (MLCC y MLCD) reportaron los resultados en forma binaria (0,1, es decir: sanas o enfermas).

El modelo *machine learning* de tipo regresión (MLRG) generó una salida de datos continuos de SCCS, que después fueron clasificados en sanas o enfermas. Este tipo de modelo se definió con la siguiente ecuación.

$$SCCS_{MLRG} = b_0 + m_0 * ED + m_1 * DEL + m_2 * NP + m_3 * PL + m_4 * CE$$

Por medio del entrenamiento del algoritmo con las reiteradas corridas de datos, se obtuvieron los valores de $b_0, m_0, m_1, \dots, m_4$. Posteriormente, estos se sustituyeron en la ecuación para calcular la predicción.

Todos los modelos de ML se pusieron a prueba con la base de datos de validación (n=157). De estos, solo para el MLRG se ejecutó la prueba de ajuste anteriormente mencionada.

Sensibilidad y especificidad

Con el 16,6% de la base de datos original se realizó la evaluación de los modelos de predicción de mastitis subclínica desarrollados, considerando criterios de sensibilidad (SEN), especificidad (ESP), tasa de error y precisión. Utilizando estos parámetros, fue posible hacer comparaciones entre los modelos obtenidos y otros modelos de detección de mastitis reportados en la literatura.

Para utilizar estos parámetros de comparación se determinó que el punto crítico (Pc) entre una vaca sana y una enferma correspondió a un SCCS de 4. Cuando el resultado fue menor o igual a este Pc, la vaca se consideró sana. Así, cuando el modelo indicó que una vaca estaba

sana y el resultado de análisis en laboratorio mostró valores iguales o menores a 200 000 cs/ml, se consideró que fue un “verdadero negativo” (VN); si el modelo indicó que la vaca estaba sana y el resultado del laboratorio fue positivo (> 200 000 cs/ml), se le llamó “falso negativo” (FN). Por el contrario, cuando el modelo predijo que la vaca estaba enferma (positiva a mastitis), se consideró “verdadero positivo” (VP) si el resultado de laboratorio fue positivo y si el resultado de laboratorio fue negativo se denominó “falso positivo” (FP).

Los valores de SEN, ESP, precisión y tasa de error se calcularon de acuerdo con las siguientes ecuaciones.

$$SEN(\%) = \frac{VP}{(VP + FN)} \times 100$$

$$ESP(\%) = \frac{VN}{(VN + FP)} \times 100$$

$$Precisión(\%) = \frac{(VN + VP)}{(VN + FP + FN + VP)} \times 100$$

$$Tasa\ de\ error(\%) = \frac{FP}{(FP + VP)} \times 100$$

Las curvas de SEN y ESP fueron utilizadas para determinar gráficamente el punto de corte entre ambas curvas, conocido como *threshold*. También se calculó, a partir de la gráfica, el valor de ESP a una SEN fijada en 80%, como parámetro de comparación entre modelos. Asimismo, se crearon curvas ROC con el fin de determinar el potencial del modelo para el diagnóstico de la enfermedad, medido como el área bajo la curva ROC.

Resultados y discusión

Descripción de la población

En términos generales se observó que la población de vacas en producción de la finca Ganadera La Uno S.A., mantuvo niveles promedio entre 391 000 cs/ml y 644 000 cs/ml durante el periodo en estudio, mostrando los conteos más

altos en los muestreos uno y dos (511 000 y 644 000, respectivamente); para luego ubicarse entre 391 000 y 488 000 cs/ml. Aproximadamente el 50% de las vacas se encontraban sanas (<200 000 cs/ml) y el otro 50% estaban enfermas (>200 000 cs/ml), con variaciones entre los muestreos. Esta característica de la población evaluada permitió tener una cantidad de datos similar por encima y por debajo del punto crítico de diagnóstico de mastitis (200 000 cs/ml), lo que favoreció el posterior análisis de sensibilidad y especificidad.

Resulta importante destacar que la finca maneja una carga de antibióticos para el tratamiento de mastitis relativamente baja (se utiliza solamente en casos muy severos), pues el manejo persigue optimizar el sistema inmunológico mediante la aplicación de productos que estimulen el sistema linfático para la producción de células de defensa.

Se observó una tendencia en el descenso de la producción láctea conforme aumentó el conteo celular somático del hato, tendencia que coincide con la encontrada por otros autores [5], [37].

Adriaens *et al.* [38] respaldan esta información y añaden que la disminución de producción se da tanto en el/los cuarto/s afectados de manera directa, como en los cuartos adyacentes de forma indirecta. Por su parte, Green *et al.* [39] mencionan que la relación entre PL y CCS se da, en gran medida, por un efecto de dilución y que debe ser estudiada con mayor profundidad, ya que la merma de la producción láctea puede estar sobreestimada.

Exploración de las variables evaluadas

El análisis de componentes principales (ACP) mostró que con la totalidad de las variables evaluadas fue posible explicar 55,4% de la variabilidad total, siendo ER una variable de poco peso y redundante con PL. Al extraer esta variable del análisis, la variación total explicada aumentó a 67,7% (*biplot* en la Figura 1a). Sin embargo, la conductividad eléctrica (CE) resultó de mediana importancia, además de estar positivamente correlacionada con el número de partos (NP). Al excluir dicha variable (Figura 1b), se obtuvo una ganancia en la explicación de la variación total (76,7%). No obstante, se

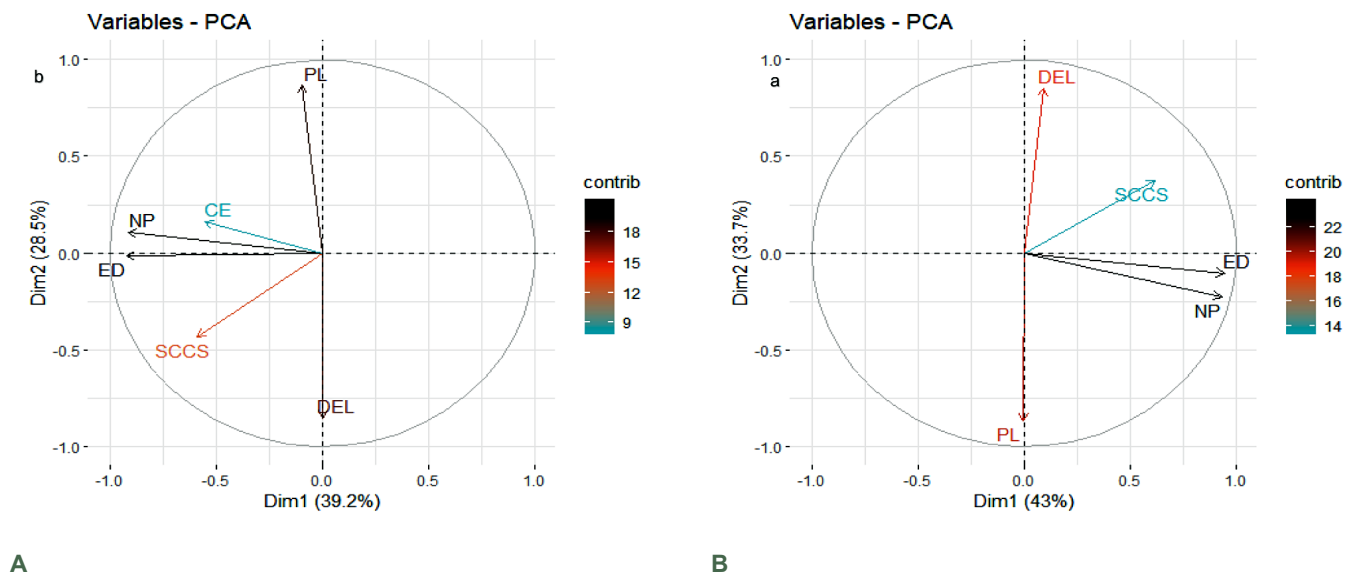


Figura 1. Análisis de componentes principales (ACP) para las variables regresoras y dependientes. a) considera la variable CE, b) excluye la variable CE. Finca Ganadera La Uno S. A., San Carlos, Costa Rica, 2017.

Figure 1. Principal component analysis (PCA) for the set of variables. a) CE variable include, b) CE variable exclude. Ganadera La Uno S.A., San Carlos, Costa Rica.

consideró pertinente conservar esta variable en el análisis, puesto que la misma se ha utilizado en otros estudios para predecir el *score* de células somáticas, o en conjunto con este para diagnosticar la mastitis [16]-[20], [40].

Selección del mejor modelo de regresión lineal clásico (MR)

Entre todos los modelos de regresión lineal desarrollados con la técnica clásica, se seleccionaron cuatro modelos para la predicción del SCCS, los cuales se presentan en el Cuadro 2. Todos los modelos se caracterizaron por utilizar variables continuas y discretas, aunque con un aumento en la complejidad del modelo, debido a la utilización de un mayor número de variables y a la combinación entre ellas, lo que pudo propiciar mayor absorción de variabilidad cuando los individuos se agruparon en conjuntos. De estos cuatro modelos de regresión lineal múltiple, el MR4 presentó los mejores ajustes: mayor coeficiente de determinación ajustado (R^2_{aj}) y valores menores de AIC, BIC y ECMP.

Selección del modelo para algoritmos de machine learning

Los resultados de los modelos que utilizaron el algoritmo de Naïve Bayes se obtuvieron en notación binaria, en los cuales las vacas sanas mostraron el valor de “0”, mientras que las vacas enfermas se identificaron con “1”. El algoritmo de clasificación de Naïve Bayes también fue estudiado por [28] para la predicción de resultados de inseminación, con una tasa de éxito baja. Esta deficiencia del modelo para predecir en determinadas condiciones podría estar explicado, ya que este asume que todas las características son independientes [41] y como se observó en el ACP de este estudio, algunas de las variables utilizadas están correlacionadas.

Durante el entrenamiento de MLRG se realizaron corridas simultáneas de datos para entrenar el algoritmo, de manera que se obtuvieron los valores de b_0 , m_0 , m_1 , m_2 , m_3 y m_4 . Estos se incluyeron en la ecuación para obtener el resultado de SCCS. El modelo a prueba se muestra en la siguiente ecuación.

$$SCCS_{MLRG} = 1,1863 + 0,0053*ED + 0,0022*DEL + 0,2616*NP - 0,0810*PL + 0,4102*CE$$

Cuadro 2. Modelos de regresión lineal para la detección de mastitis subclínica y criterios de diagnóstico. Finca Ganadera La Uno S.A, San Carlos, Costa Rica, 2017.

Table 2. Linear regression models for the detection of subclinical mastitis and diagnostic criteria. Ganadera La Uno S.A, San Carlos, Costa Rica.

Modelos de regresión						
MR1. $SCCS_{MR1} = 1,6445 + 0,3811CE - 0,0961PL + 0,0203ED$						
MR2. $SCCS_{MR2} = 0,9629 + 0,5930NPc - 0,0878PL + 0,2376DELc + 0,3560CE$						
MR3. $SCCS_{MR3} = 1,2954 + 0,4949NPc + 0,1203EDc + 0,2932CE - 0,0785PL + 0,0003CE*DEL$						
MR4. $SCCS_{MR4} = 1,1580 + 0,5442NPc + 0,0529ED + 0,0018DEL - 0,0153NPc*ED - 0,0885PL + 0,1179CE*NPc$						
	R^2_{aj}	AIC	BIC	ECMP	p-valor modelo	p-valor Wilcoxon
1. MR1	0,2318	2803,18	2826,51	2,0890	<0,0001	0,7018
2. MR2	0,2744	2759,42	2787,41	1,9922	<0,0001	0,6811
3. MR3	0,2765	2758,07	2790,73	1,9764	<0,00 01	0,5287
4. MR4	0,3115	2720,15	2757,47	1,8838	<0,0001	0,5696

Aunque el análisis de componentes principales indicó que la CE fue de menor peso que el resto de las variables (Figura 1), para el modelo con mayor precisión resultó ser un parámetro de mucho peso, con el valor del coeficiente más alto que asignó el algoritmo. Nielen *et al.* [42] demostraron el potencial de esta variable para clasificar vacas sanas y enfermas de mastitis mediante el uso del algoritmo de regresión lineal supervisado.

No se han reportado estudios similares, en el ámbito pecuario de Costa Rica y Centroamérica, que mencionen la aplicación de modelos basados en algoritmos de inteligencia artificial para clasificar vacas de acuerdo con el estado sanitario de la ubre. Por esta razón el presente estudio abre una ventana para la investigación en este campo que permitirá perfeccionar la implementación del *machine learning* en los sistemas de producción ganaderos para su optimización, con la ventaja de que los algoritmos de ML son capaces de aprender las relaciones por sí mismos.

Prueba de sensibilidad (SEN) y especificidad (ESP)

La comparación de sensibilidad y de especificidad entre los modelos, utilizando la

base de datos de validación (n= 157 registros) (Cuadro 3), reveló que los modelos MR1, MR2 y MR3 presentaron una ESP más alta (de 70,0 a 75,0%) en comparación al modelo MR4 (61,3%). Además, MR4 obtuvo el valor más alto (71,4%) en SEN entre todos los modelos de regresión, lo que mostró su buen desempeño para predecir vacas enfermas en el punto crítico fijado.

Por otra parte, MLCC obtuvo la SEN más baja (54,5%), mientras que MLRG la ESP más alta (88,8%), lo cual le da al primero un bajo poder de detección de vacas enfermas y al segundo el mayor potencial para discriminar vacas sanas. MR1, MR3 y MLCD obtuvieron los mismos resultados en este análisis para la detección de mastitis (vacas enfermas).

Los resultados sugieren que tanto los modelos de regresión lineal clásicos, como aquellos de *machine learning* que utilizaron el algoritmo de clasificación de Naïve Bayes (MLCC y MLCD), no resultaron apropiados para la predicción de mastitis a partir del punto crítico (Pc) determinado, ya que la precisión de estos modelos varió entre 66,2 y 67,5%, mientras tanto la tasa de error fue mayor al 30%. MR4 fue el modelo menos preciso y con mayor tasa de error, a pesar de los parámetros obtenidos entre los modelos de regresión lineal. Por otro lado, se encontró que

Cuadro 3. Sensibilidad (SEN), especificidad (ESP), precisión y tasa de error para los modelos de regresión lineal clásico y de *machine learning*. Ganadera La Uno S.A, San Carlos, Costa Rica, 2017.

Table 3. Sensitivity (SEN), specificity (ESP), accuracy and error rate, for classical linear regression and machine learning models. Ganadera La Uno S.A, San Carlos, Costa Rica.

Modelo	Variables	SEN (%)	ESP (%)	Precisión (%)	Tasa de error (%)
Modelos de Regresión lineal					
1. MR1	3	59,7	75,0	67,5	30,3
2. MR2	4	64,9	70,0	67,5	32,4
3. MR3	5	59,7	75,0	67,5	30,3
4. MR4	5	71,4	61,3	66,2	36,0
Modelos de <i>machine learning</i>					
5. MLCD	6	59,7	75,0	67,5	30,3
6. MLCC	5	54,5	77,5	66,2	30,0
7. MLRG	5	68,8	88,8	79,0	14,5

el modelo con mayor precisión (79,0%) y menor tasa de error (14,5%) fue MLRG, resultando con valores de SEN y ESP de 68,8 y 88,8% respectivamente, lo cual sugiere una relación lineal entre los parámetros utilizados y el SCCS.

En la Figura 2 se muestran las curvas de SEN y ESP únicamente para MLRG. Los gráficos incluyen el *threshold* obtenido para cada uno con el objetivo determinar el valor de la variable de pronóstico, que es el punto donde se igualan SEN y ESP. El valor de ESP al 80% de SEN fija fue de 84,3% para MLRG, mientras que para los modelos MR (1, 2, 3 y 4) alcanzaron valores de 52,0, 56,7, 54,2 y 53,0% respectivamente.

Cavero *et al.* [24] utilizaron una red neuronal de tipo clasificación, con cinco variables regresoras (rendimiento de leche, tasa promedio de flujo de leche, tiempo entre ordeños, tasa de producción de leche y CE de cada cuarto), alcanzando ESP de 51,1 y 74,9% a una SEN fija de 80%, para dos Pc (100 000 cs/ml y 400 000 cs/ml, respectivamente). Ninguno de estos resultados superó la ESP a una SEN fija de 80% del MLRG

de este estudio, con un punto crítico de 3,67 SCCS, equivalente a 159 000 cs/ml.

Ferrero [43] construyó una red neuronal considerando un conjunto mayor de variables, tanto productivas (número de lactancias, días en lactación, época de parto, incidencias previas de la enfermedad), como ambientales (dos variables de precipitación y dos de estrés calórico). La precisión del modelo obtenido fue de 90,74%, y la SEN y ESP fueron de 86,9% y 91,36% respectivamente. En total se utilizaron 12 variables y no todas de fácil medición. En el presente estudio se obtuvo una precisión aceptable con un set de cinco variables (dos de fácil medición en la línea de leche y tres del registro histórico de los animales).

Utilizando modelos de inteligencia artificial y modelos aditivos generalizados, Ankinakatte *et al.* [40] lograron diferenciar entre vacas sanas y enfermas (por cultivo bacteriológico), con una ESP a un 80% de SEN fija cercana a 75%, aunque la ausencia del conteo de células somáticas en el modelo disminuyó la habilidad

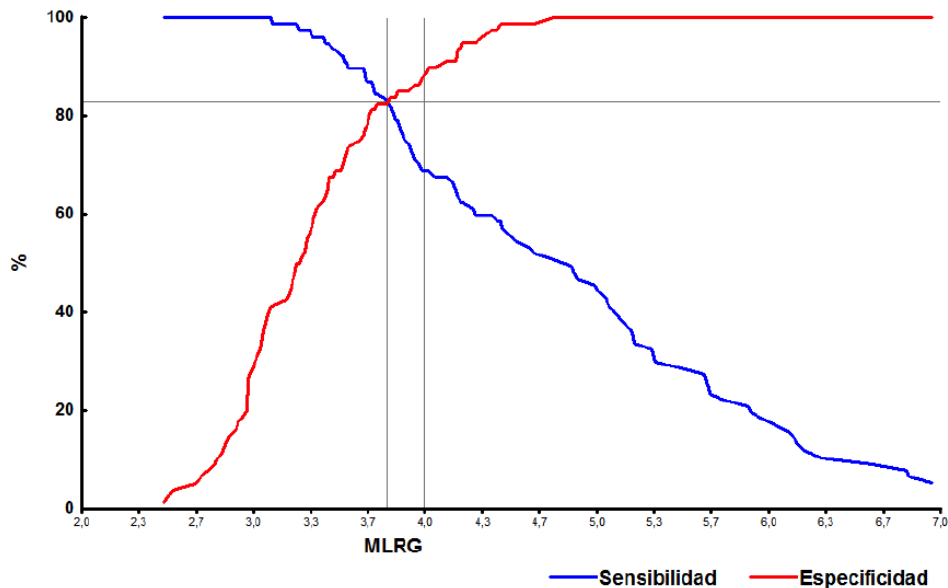


Figura 2. Curvas de sensibilidad (SEN) y especificidad (ESP) del modelo MLRG. Se incluye el *threshold* y se indica la ESP a un 80% de ESP fija. Ganadera La Uno S.A, San Carlos, Costa Rica, 2017.

Figure 2. Sensitivity (SEN) and specificity (ESP) curves of the MLRG model (include threshold and ESP at 80% fixed ESP). Ganadera La Uno S.A, San Carlos, Costa Rica.

de predicción en al menos 5%. Estos autores utilizaron ambos algoritmos para clasificar los datos y concluyeron que los algoritmos de redes neuronales fueron ligeramente mejores para especificidades altas, tendencia que coincide con los resultados obtenidos por *machine learning* para los modelos de clasificación (MLC) y regresión (MLRG) (Cuadro 3).

Las curvas ROC expresan la relación entre la sensibilidad y 1 - especificidad, siendo el área bajo esta curva el criterio más importante para comparar entre métodos de predicción de una enfermedad: cuando este valor se aproxima a 1 aumenta el valor predictivo, y cuando se acerca a 0,5 pierde su efectividad, indicando que el modelo en este estudio no tendrá ningún poder de discriminación entre vacas sanas y enfermas. El mejor resultado en función del área bajo la curva ROC lo mostró el modelo MLRG que se presenta en la Figura 3, con el 0,921 del área cubierta; mientras que los modelos restantes se mantuvieron entre 0,735 y 0,754. En un estudio realizado en vacas Holstein en pastoreo, Elizalde *et al.* [18] desarrollaron un modelo de predicción basado en conductividad eléctrica diferencial (entre cuartos), con un área bajo la curva ROC de 0,82 y niveles estimados SEN y ESP de 78,0 y 79,1% respectivamente, lo

cual concuerda con los resultados obtenidos en esta investigación, donde se encontró evidencia que demuestra que la conductividad eléctrica puede ser una variable predictora de mastitis. La utilización de la conductividad eléctrica en los modelos de predicción de mastitis para sistemas de detección en línea se ha vuelto muy común desde los inicios de la aplicación de inteligencia artificial en este campo [42]. Por otra parte, Hassan *et al.* [44] reportan variaciones a este parámetro donde se prueba la resistencia eléctrica en lugar de la conductividad eléctrica, obteniendo resultados satisfactorios en la identificación de agentes causales de hasta 99% de especificidad en modelos de redes neuronales artificiales no supervisados y de 97%-80% en modelos supervisados.

Utilizar algoritmos de inteligencia artificial para obtener buenos resultados en la predicción de enfermedades a partir de variables sencillas es todavía una interrogante. La literatura reporta el uso de modelos de predicción de mastitis con estos algoritmos con variables de fácil medición como la conductividad eléctrica, la producción láctea o patógenos en leche [42], [23], [24], [44], [25], [40], [26], adicionando variables ambientales [43] u otras más complejas que integran incluso parámetros de la genealogía del

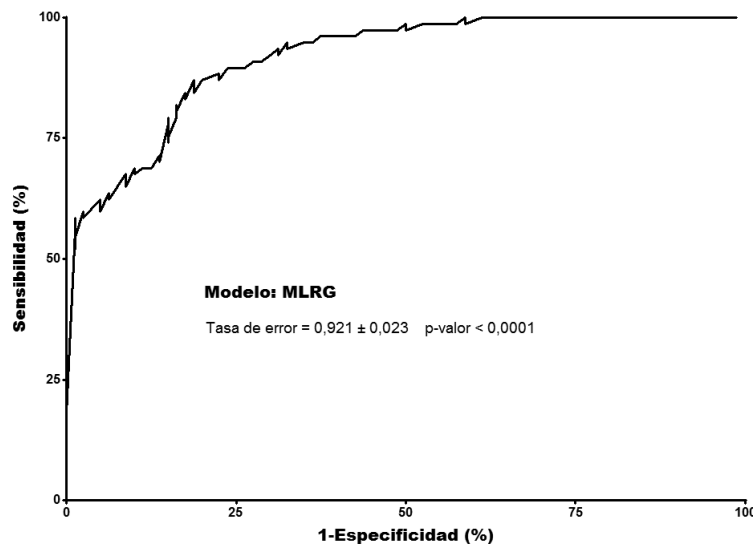


Figura 3. Curva ROC para el modelo MLRG, error y p-valor de la curva. Ganadera La Uno S.A, San Carlos, Costa Rica, 2017.

Figure 3. ROC curve for MLRG model, error and p-value of the curve. Ganadera La Uno S.A, San Carlos, Costa Rica.

animal [45]. La ventaja que estos tienen es que son capaces de aprender de la información que se les provee (datos), buscando constantemente la minimización del error. Tanto las ecuaciones obtenidas en los modelos de regresión lineal (MR) como en el MLRG, son ecuaciones lineales; la diferencia radica en que el algoritmo de aprendizaje automático (ML) realiza reiteradas corridas de datos que le permiten encontrar la intersección con el eje “y” (b_0) y los pesos ($m_0 \dots m_n$) de “n” variables predictoras que se ajusten al mejor resultado de clasificación; contrario a los modelos de regresión lineal clásicos en los que se busca, por lo general, maximizar el coeficiente de determinación.

Conclusiones

- Aproximadamente el 50% de las vacas del ható de la finca Ganadera La Uno S.A. se encontraban enfermas, lo que favoreció en gran medida el estudio ya que se contaba con cantidades similares de vacas sanas y enfermas.
- La conductividad eléctrica fue una variable importante en el algoritmo de *machine learning* para la predicción de vacas sanas y enfermas, a pesar de que demostró poca relación con el *score* de células somáticas en el análisis previo (ACP).
- Los modelos de regresión lineal desarrollados para la predicción del *score* de mastitis fueron muy similares entre ellos, con valores de especificidad a un 80% de sensibilidad fija, cercana a 55% y áreas de curvas ROC de 75% aproximadamente.
- El mejor modelo para la predicción de mastitis a partir de dos variables de rápida medición *in situ* e indicadores técnicos del animal fue el algoritmo elaborado con la función de *machine learning* “Linear Regression”, caracterizado por un área bajo la curva ROC de 0,921 y una especificidad de 84,3 a una sensibilidad fijada de 80%.

Recomendaciones

Identificar cambios repentinos en las dos variables de medición en línea de leche (CE y PL), podría

generar mejores predictores de mastitis, así como la inclusión de más variables de rápida medición en los sistemas, teniendo en cuenta parámetros de clima y estrés de los animales.

Estudiar el comportamiento de los modelos obtenidos con otros hatos que mantengan niveles de células somáticas menores a los del ható analizado.

Crear una arquitectura de red neuronal que permita explicar mejor el comportamiento de la mastitis en función de estas variables.

Bibliografía

- [1] C. Viguier, S. Arora, N. Gilmartin, K. Welbeck, R. O’Kennedy, “Mastitis detection: current trends and future perspectives”, *Trends in Biotechnology*, vol. 27, no. 8, pp. 486-493, 2009.
- [2] M. L. Fernández, J. P. Ramírez, C. Chaves, M. L. Arias, “Disminución en la incidencia de mastitis en ganado vacuno con la aplicación de un sellador de barrera experimental”, *Agronomía Costarricense*, vol. 32, no. 1, pp. 107-112, 2008.
- [3] M. Gráff, E. Mikó, “Analysis of mastitis in Holstein-Friesian cows and economic effects of mastitis”, *Agricultural Management / Lucrari Stiintifice*, vol. 17, no. 1, pp. 294-297, 2015.
- [4] M.G. Mora, B. Vargas, J. J. Romero, J. Camacho, “Efecto de factores genéticos y ambientales sobre el recuento de células somáticas en ganado lechero de Costa Rica”, *Agronomía Costarricense*, vol. 40, no. 2, pp. 7-18, 2016.
- [5] J. L. Guimarães, M. A. Brito, C. C. Lange, M. R. Silva, J. B. Ribeiro, L. C. Mendonça, G. N. Souza, “Estimate of the economic impact of mastitis : a case study in a Holstein dairy herd under tropical conditions”, *Preventive Veterinary Medicine*, vol. 142, pp. 46-50, 2017.
- [6] J. L. Gonçalves, C. Kamphuis, C. M. Martins, J. R. Barreiro, T. Tomazi, A. H. Gameiro, H. Hogeveen, M. V. dos Santos, “Bovine subclinical mastitis reduces milk yield and economic return”, *Livestock Science*, vol. 210, pp. 25-32, 2018.
- [7] K. Huijps, T. Lam, H. Hogeveen, “Costs of mastitis: facts and perception”, *Journal of Dairy Research*, vol. 75, no. 1, pp. 113-120, 2008.
- [8] C. Nielsen, “Economic impact of mastitis in dairy cows”, Ph.D. Tesis, Swedish University of Agricultural Sciences, Uppsala, 2009.
- [9] R. A. Batavani, S. Asri, H. Naebzadeh, “The effect of subclinical mastitis on milk composition in dairy cows”, *Iranian Journal of Veterinary Research*, vol. 8, no. 3, pp. 205-211.

- [10] N. Sharma, N. K. Singh, M. S. Bhadwal, "Relationship of somatic cell count and mastitis: an overview", *Asian-Australasian Journal of Animal Science*, vol. 24, no. 3, pp. 429-438, 2011.
- [11] X. Zhao, P. Lacasse, "Mammary tissue damage during bovine mastitis: Causes and control", *Journal of Animal Science*, vol. 86, no. 13, pp. 57-65, 2014.
- [12] R. M., de Mol, W. Ouweltjes, "Detection model for mastitis in cows milked in an automatic milking system", *Preventive Veterinary Medicine*, vol. 49, no. 1-2, pp. 71-82, 2001.
- [13] W. J. Zambrano, Jr. A. de Pinho, "Evaluación de la glándula mamaria y composición química de la leche en vacas primíparas mestizas lecheras en el parto, hasta el quinto mes de la lactación", *Revista Científica*, vol. 18, no. 5, pp. 562-569, 2008.
- [14] B. Polat, A. Colak, M. Cengiz, L.E. Yanmaz, H. Oral, A. Bastan, S. Kaya, A. Hayirli, "Sensitivity and specificity of infrared thermography in detection of subclinical mastitis in dairy cows", *Journal of Dairy Science*, vol. 93, no. 8, pp. 3525-3532, 2010.
- [15] N. Sharma, V. Pandey, N. A. Sudhan, "Comparison of some indirect screening tests for detection of sub-clinical mastitis in dairy cows", *Bulgarian Journal of Veterinary Medicine*, vol. 13, no. 2, pp. 98-103, 2010.
- [16] A. Ribeiro, J. dos Santos, D. Zanol, L. Leme, S. Bruzaroski, A. Ludovico, E. de Santana, "Evaluation of an electrical conductivity portable device as an alternative for subclinical mastitis detection", *Revista de Salud Animal*, vol. 38, no. 2, pp. 31-135, 2016.
- [17] M. Khantun, C. Clark, N. Lyons, P. Thomson, K. Kerrisk, S. García, "Early detection of clinical mastitis from electrical conductivity data in an automatic milking system", *Animal Production Science*, vol. 57, no. 7, pp. 1226-1232, 2017.
- [18] E. F. Elizalde, M. L. Signorini, V. R. Canavesio, A. Cuatrin, H. D. Tarabla, L. F. Calvino, "Medición de la conductividad eléctrica en leche como método diagnóstico de mastitis subclínica bovina", *Revista FAVE - Ciencias Veterinarias*, vol. 8, no. 1, pp. 5-28, 2009.
- [19] G. Kaçıkçı, Ö Çetin, E. Bingöl, M. Gündüz, "Relations between electrical conductivity, somatic cell count, California mastitis test and some quality parameters in the diagnosis of subclinical mastitis in dairy cows", *Turkish Journal of Veterinary and Animal Sciences*, vol. 36, no. 1, pp. 49-55, 2012.
- [20] E. Norberg, G. W. Rogers, R. C. Goodling, J. B. Cooper, P. Madsen, "Genetic Parameters for Test-Day Electrical Conductivity of Milk for First-Lactation Cows from Random Regression Models", *Journal of Dairy Science*, vol. 87, pp. 1917-1924, 2004.
- [21] D. Cavero, K. Tölle, C. Buxadé, J. Krieter, "Mastitis detection in dairy cows by application of fuzzy logic", *Livestock Science*, vol. 105, no. 1-3, pp. 207-213, 2006.
- [22] D. Cavero, K. Tölle, G. Rave, C. Buxadé, J. Krieter, "Analysing serial data for mastitis detection by means of local regression", *Livestock Science*, vol. 110, no. 1-2, pp. 101-110, 2007.
- [23] C. W. Heald, T. Kim, W. M. Sischo, J. B. Cooper, D. R. Wolfgang, "A Computerized Mastitis Decision Aid Using Farm-Based Records: An Artificial Neural Network Approach", *Journal of Dairy Science*, vol. 83, no. 4, pp. 711-720, 2000.
- [24] D. Cavero, K. Tölle, C. Henze, C. Buxadé, J. Krieter, "Mastitis detection in dairy cows by application of neural networks", *Livestock Science*, vol. 114, no. 2-3, pp. 280-286, 2008.
- [25] C. Kamphuis, H. Mollenhorst, A. Feelders, D. Pietersma, H. Hogeveen, "Decision-tree induction to detect clinical mastitis with automatic milking", *Computers and Electronics in Agriculture*, vol. 70, no. 1, pp. 60-68, 2010.
- [26] H. Küçükönder, F. Üçkardeş, A. Ceyhan, M. Cinar, "Determination of the effect of somatic cell count on udder measurements and subclinical mastitis with data mining method", *Pakistan Veterinary Journal*, vol. 35, no. 4, pp. 441-445, 2015.
- [27] B. Sitkowska, D. Piwczynski, J. Aerts, M. Kolenda, S. Özkaya, "Detection of high levels of somatic cells in milk on farms equipped with an automatic milking system by decision trees technique", *Turkish Journal of Veterinary and Animal Sciences*, vol. 41, pp. 532-540, 2017).
- [28] S. Shahinfar, D. Page, J. Guenther, V. Cabrera, P. Fricke, K. Weigel, K. "Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms". *Journal of Dairy Science*, 97 (2), 731-742, 2014.
- [29] P. Shine, M. D. Murphy, J. Upton, T. Scully, "Machine-learning algorithms for predicting on-farm direct water and electricity consumption on pasture based dairy farms", *Computers and Electronics in Agriculture*, vol. 150, pp. 74-87, 2018.
- [30] J. Hai-Guo, Z. Guo-li, C. Yang, Y. Ming-Zing, "Study on genetic variation of 4 microsatellite DNA markers and their relationship with somatic cell counts in cow milk", *Asian-Australasian Journal of Animal Science*, vol. 16, no. 10, pp. 1535-1539, 2003.
- [31] N. Sharma, T. Y. Kang, S. J. Lee, J. N. Kim, C. H. Hur, J. C. Ha, V. Vohra, D. K. Jeong, "Status of bovine mastitis and associated risk factors in subtropical Jeju Island, South Korea", *Tropical Animal Health and Production*, vol. 45, no. 8, pp. 1829-1832, 2013.
- [32] G. R. Wiggans, G. E. Shook, "A lactation Measure of Somatic Cell Count", *Journal of Dairy Science*, vol. 70, pp. 2666-2672, 1987.
- [33] JA. DiRienzo, F. Casanoves, M. Balzarini, L. González, M. Tablada, C. Robledo, InfoStat versión 17-11-2016. Córdoba: Universidad Nacional de Córdoba, 2017.

- [34] T. Augspurger, C. Bartak, P. Cloud, A. Hayden, S. Hoyer, W. McKinney, J. Reback, C. She, M. Horikoshi, J. Bossche, "Pandas". [Online]. Disponible en: <https://www.py pandas.cn/about/#>. [Accesado el 3 ene., 2019]
- [35] J. Cabral, N. Luczywo, J. Zanazzi, "Scikit-Criteria: Colección de métodos de análisis multi-criterio integrado al stack científico de Python", Simposio Argentino de Investigación Operativa (SIO), Buenos Aires, Argentina, 2016.
- [36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
- [37] J. Gonçalves, R. Cue, B. Botaro, J. Horst, A. Valloto, M. Santos, "Milk losses associated with somatic cell counts by parity and stage of lactation", *Journal of Dairy Science*, vol. 101, no. 5, pp. 4357-4366, 2018.
- [38] I. Adriaens, T. Huybrechts, B. Aernouts, K. Geerinckx, S. Piepers, B. de Ketelaere, W. Saeys, "Method for short-term prediction of milk yield at the quarter level to improve udder health monitoring", *Journal of Dairy Science*, vol. 101, no. 11, pp. 1-10, 2018.
- [39] L. E. Green, Y. H. Schukken, M. J. Green, "On distinguishing cause and consequence: Do high somatic cell counts lead to lower milk yield or does high milk yield lead to lower somatic cell count?", *Preventive Veterinary Medicine*, vol. 76, no. 1-2, pp. 74-89.
- [40] S. Ankinakatte, E. Norberg, P. Løvendahl, D. Edwards, S. Højsgaard, "Predicting mastitis in dairy cows using neural networks and generalized additive models: A comparison", *Computers and Electronics in Agriculture*, vol. 99, pp. 1-6, 2013.
- [41] T. Patil, S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", *International Journal of Computer Science and Applications*, vol. 6, no. 2, pp. 256-261, 2013.
- [42] M. Nielen, M. H. Spigt, Y. H. Schukken, H. A. Deluyker, K. Maatje, A. Brand, "Application of a neural network to analyse on-line milking parlour data for the detection of clinical mastitis in dairy cows", *Preventive Veterinary Medicine*, vol. 22, no. 1-2, pp. 15-28, 1995.
- [43] M. Ferrero, "Generación de un modelo mediante el uso de redes neuronales artificiales para la detección de mastitis en vacas lecheras del INTA estación experimental agropecuaria Rafaela", XLIII Jornada Argentina de Informática e Investigación Operativa (43JAIO), Buenos Aires, Argentina, 2014.
- [44] K. J. Hassan, S. Samarasinghe, M. G. Lopez-Benavides, "Use of neural networks to detect minor and major pathogens that cause bovine mastitis", *Journal of Dairy Science*, vol. 92, no. 4, pp. 1493-1499, 2009.
- [45] X. Z. Yang, R. Lacroix, K. M. Wade, "Investigation into the production and conformation traits associated with clinical mastitis using artificial neural networks", *Canadian Journal of Animal Science*, vol. 80, no. 3, pp. 415-426, 2000.

De acuerdo con la norma IEEE, este documento debe citarse:

V. Estrada-Carvajal, M. Camacho-Calvo, R. Molina-Montero, W. Paniagua-Madrigal, "Detección de mastitis subclínica en vacas lecheras por modelos de regresión lineal y algoritmos de inteligencia artificial, San Carlos, Costa Rica", *Revista AgroInnovación en el Trópico Húmedo*, vol. 2, no. 1, pp.18-30, 2019, DOI: 10.18860/rath.v2i1.4689