EPISTEMUS
CIENCIA, TECNOLOGÍA Y SALUD
ISSN: 2007-8196
Ciencia, Tecnología y Sociedad

EPISTEMUS 2022; Núm. 33, Vol. 16

DOI: https://doi.org/10.36790/epistemus.v16i33.220
www.epistemus.unison.mx

Árboles de decisión para clasificación de vacas lecheras usando información genética

EDELMIRA RODRÍGUEZ ALCÁNTAR[1]

RESUMEN

En este trabajo se presenta a los árboles de decisión como una técnica de aprendizaje automático para la clasificación de vacas como buenas productoras de leche a partir del uso de marcadores genéticos. La finalidad es realizar una selección de animales genéticamente superiores en menor tiempo y hacer más eficiente el proceso de reproducción asistida logrando con ello disminuir costos y aumentar ganancias en el sector lechero.

Los resultados de los experimentos realizados muestran hasta un 94.5% de precisión. Además, el algoritmo permitió la identificación del SNP más dominante para la clasificación, y el cromosoma que más influye en la predicción.

Palabras clave: *Clasificación, árboles de decisión, producción lechera*

[1] Dra. Rodríguez Alcántar E., Departamento de matemáticas, Universidad de Sonora, Hermosillo, Sonora, México, https://orcid.org/0000-0003-1825-102X

Autor de Correspondencia: Edelmira Rodríguez Alcántar, edelmira.rodriguez@unison.mx

Recibido: 17 / 01 / 2022
Aceptado: 20 / 07 / 2022
Publicado: 24 / 08 / 2022

Cómo citar este artículo:

RODRIGUEZ ALCANTAR, E. (2022). Árboles de decisión para clasificación de vacas lecheras usando información genética. EPISTEMUS, 16(33). https://doi.org/10.36790/epistemus.v16i33.220

Decision Tree to Classification of Dairy Cows from Genetic Information

ABSTRACT

This paper presents decision trees as a machine learning technique for classifying cows as good milk producers or not, based on the use of genetic markers. The purpose is to select genetically superior animals in less time and make the assisted reproduction process more efficient, thereby reducing costs and increasing profits in the dairy sector. Results are presented on the efficiency of decision trees for the classification of dairy cows, up to 94.5% accuracy was achieved. In addition, the algorithm allowed the identification of the most dominant SNP for classification, and the chromosome that most influences the prediction.

Key words: ***Classification, decision trees, dairy production***

INTRODUCCIÓN

Producción lechera en México

Según las estadísticas de la Cámara Nacional de Industriales de la Leche [1], actualizadas al mes de abril del 2021, México ocupa el 16º lugar como productor de leche con un 2% de la producción mundial. En la **Figura *1*** se ilustra esta información, para dar una mejor idea de la proporción que esto representa. La leche de bovino representa el 17% de la producción pecuaria nacional.

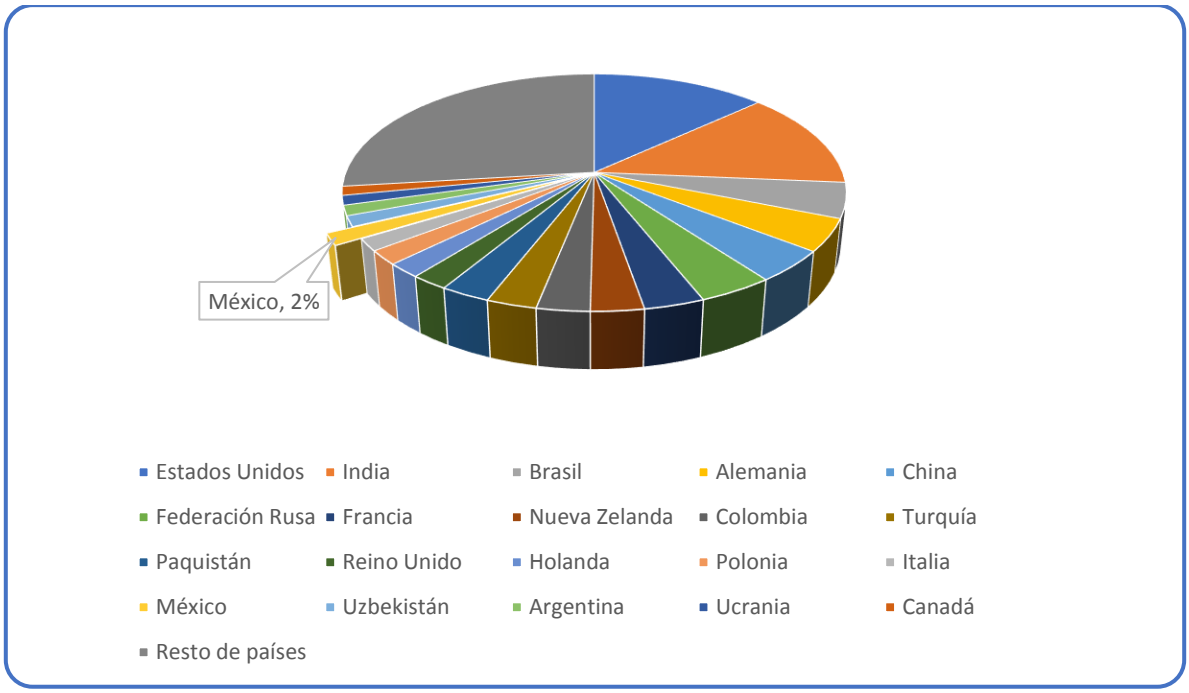


Figura 1. Principales países productores de leche de bovino, 2021.

Sin embargo, la producción nacional no es suficiente para satisfacer la demanda, pues, por ejemplo, en el 2020 se importaron productos lácteos por 1749 mdd desde EUA, de donde provino el 81% de las importaciones ese año. De hecho, la balanza comercial en notablemente desfavorable, importando productos lácteos en una

cantidad aprox. 3.4 veces mayor de lo que se exporta en el país (2020). Como puede notarse, la producción lechera en México tiene una oportunidad de crecimiento importante al pretender cubrir el déficit de producción nacional o, incluso, aumentar las exportaciones. Para ello, un punto clave puede ser optimizar las técnicas empleadas en la selección de las vacas que se destinarán a la producción lechera, debido a que se requiere esperar tres años para evaluar si una vaca es buena productora de leche. Al utilizar información genética la evaluación del fenotipo lechero puede ser más rápida y barata, pues no se estaría invirtiendo en la crianza de una vaca que no sea buena productora.

Genotipo y fenotipo

La genética está dedicada al estudio de las características hereditarias. Entre los temas básicos que se abordan en esta disciplina se encuentran el concepto de genotipo y de fenotipo. El genotipo es la información genética que posee un organismo en su ADN. Es decir, es información heredada de los padres. Los rasgos observables que caracterizan a un organismo constituyen su fenotipo (por ejemplo, la forma de los ojos, el largo de las orejas y el color del pelaje).

Marcadores genéticos

En Biotecnología, un “marcador molecular” es un fragmento de ADN en asociación con una cierta locación en el genoma (puede ser llamado también “marcador genético” o simplemente “marcador”); un marcador es usado para identificar una secuencia de ADN parcial en un grupo de ADN desconocido. Los Polimorfismos de Nucleótido Simple (*Single Nucleotide Polymorphisms,* SNP) son un tipo de

marcadores muy utilizados que, aunque por sí solos no proporcionan información sobre genes específicos, indican una localización cromosómica con probabilidades de estar asociada con un fenotipo dado. En diciembre de 2007, se liberó el chip BovineSNP50 BeadChip para análisis de DNA, con 54001 SNPs y en el 2009 se empezó a usar oficialmente para la selección genética en la industria lechera de Estados Unidos [2] .

Estudio de asociación de genoma completo

Un estudio de asociación de genoma completo (*Genome-wide Association Study*, GWAS), es un estudio de asociación del genoma con un fenotipo en particular. Los GWAS permiten utilizar un gran número de marcadores genéticos a lo largo de todo el genoma para detectar variaciones asociadas con una enfermedad o rasgo particular. Un GWAS permite a los investigadores analizar individuos sin conocer su pedigrí [3].

Aunque los SNPs pueden no ser ellos mismos responsables de la variación observada en un rasgo, debido a su proximidad a las variantes causales no genotípicas, han sido heredados conjuntamente y, por lo tanto, pueden actuar como representantes de las variantes causales desconocidas. De esta manera, los SNPs asociados significativamente con una enfermedad o rasgo pueden indicar una región del genoma que alberga variantes genéticas que influyen en la expresión de esa enfermedad o rasgo.

Se han desarrollado nuevos valores de crianza estimada basados en los marcadores SNPs densos (*Genomic Estimated Breeding Values,* GEBV) cubriendo

el genoma bovino completo, capturando así todos los locus de un carácter cuantitativo (*Quantitative Trait Loci,* QTL) que contribuyen a la variación de un rasgo, dando lugar al nuevo campo llamado Selección Genética (*Genetic Selection*, GS) [4]. La principal limitación para la implementación de selección genómica ha sido el gran número de marcadores requeridos y el costo de genotipado de estos marcadores [5].

En este trabajo, se utilizan árboles de decisión, una estrategia de Aprendizaje Automático (*Machine Learning*, ML), para predecir si una vaca será alta-productora de leche a partir de su información genética.

Aprendizaje automático

El análisis de grandes datos genómicos se ve obstaculizado por problemas como un pequeño número de observaciones y un gran número de variables predictivas, alta dimensionalidad o estructuras de datos altamente correlacionadas. Los métodos de aprendizaje automático son famosos por tratar estos problemas [6].

ML es un área de Inteligencia Artificial basada en la idea de que los sistemas informáticos pueden aprender mediante el análisis de datos en la búsqueda de patrones para generar un modelo capaz de hacer predicciones. Un problema de aprendizaje puede definirse como el problema de mejorar alguna medida de rendimiento, a través del entrenamiento, al realizar una tarea [7]. ML tiene dos categorías principales: métodos de aprendizaje supervisados [8] y métodos de aprendizaje no supervisados [9]. Los métodos supervisados se entrenan con ejemplos etiquetados y luego se usan el modelo entrenado para hacer predicciones

sobre ejemplos no etiquetados, mientras que los métodos no supervisados encuentran la estructura en un conjunto de datos sin usar etiquetas. El aprendizaje puede usarse para predecir datos categóricos (lo que se denomina predicción categórica o clasificación) o para predecir datos de valores reales, lo que se denomina regresión.

El objetivo general en este trabajo es demostrar el potencial del aprendizaje automático, en particular de los árboles de decisión, como un marco poderoso para el análisis genético que puede permitir realizar predicciones sobre la producción lechera de ganado bovino raza Holstein, partiendo de marcadores SNP a lo largo del genoma completo.

DESARROLLO

Conjunto de datos utilizado

El conjunto de datos usado en este trabajo fue obtenido de Chen et al. [10]. Los datos consisten en muestras genotípicas de 1092 vacas Holstein, en un panel de 164312 SNPs con 29 cromosomas autosomales. Los valores de los genotipos son 0, 1 y 2 para representan homocigotos de alelo menor, heterocigotos y homocigotos de alelo mayor, respectivamente. Se cuenta con medidas de fenotipos para diferentes rasgos y se seleccionó la producción lechera promedio a 305 días ('milk_le_ave_305').

Los cromosomas en el dataset están etiquetados como chr1, chr2 ,… , chr29; para realizar el análisis se seleccionan aquellos cromosomas que contienen el número más significativo de QTLs relacionados con la producción lechera, según la base de

datos Cattle QTLdb. En la ***Tabla 1*** se presenta, para cada cromosoma, el número de QTLs relacionado con la producción lechera y el número de SNPs muestreados en el *dataset*.

Nótese que el cromosoma 14 resalta del resto debido a su número de QTLs asociados con la producción lechera. Los cromosomas con un mayor número de QTLs relacionados con la producción lechera son: chr1, chr5, chr6, chr14, chr17, chr20 y chr26. Estos son los cromosomas seleccionados para formar los subconjuntos con los que se harán las pruebas.

Tabla 1. Número de QTLs relacionados con la producción lechera y número de SNPs, por cromosoma.

Chr	Número de QTLs	Número de SNPs	Chr	Número de QTLs	Número de SNPs
1	23	7338	16	10	5269
2	16	7049	17	22	4750
3	21	8064	18	10	7579
4	12	7572	19	16	6108
5	31	7733	20	25	3395
6	36	5312	21	20	5871
7	18	7465	22	5	3765
8	8	6088	23	18	4548
9	10	5273	24	1	3622
10	14	5952	25	6	5773
11	7	7120	26	22	3536
12	8	6640	27	8	3492
13	13	6736	28	7	3680
14	51	4004	29	9	5004
15	4	5574			

Como estrategia de control de calidad (*Quality Control*, *QC*), se aplicaron ciertos filtros al *dataset* para asegurar la calidad general de las muestras y que se trata de un conjunto consistente de genotipos. El filtrado incluye remover:

1. Todas aquellas muestras que tengan >20% de genotipos faltantes.
2. Todos aquellos SNPs que violen la distribución de frecuencias de Hardy-Weinberg, como se aplica en [11].
3. Todos aquellos SNPs que tienen una frecuencia del alelo menor (MAF) <5%, por considerarse de variación no significativa.

Especificaciones de Hardware y software

Se utilizó un *cluster* de computadoras que usa el sistema operativo Linux CentOS 6.7, con un i5-2500 Quad-Core, procesador a 3.30 GHz, 4 GB RAM DDR3 1333 MHz.

El lenguaje de programación utilizado es Python (v.3.6.8) y la librería *scikit-learn*, con herramientas simples y eficientes para el aprendizaje automático.

Evitar sobreajuste

Suponer que se tiene un conjunto de datos compuesto por (X, y) donde X es la matriz de los genotipos (atributos) de tamaño (*nMuestras*,*nAtributos*) y y es el vector donde se tiene el fenotipo esperado de cada muestra, de tamaño *nMuestras*.

Un error común al entrenar una función de aprendizaje es usar los mismos datos para el entrenamiento y para las pruebas pues se tendrá un desempeño perfecto, ya que se conocen todas las muestras. Esto se conoce como *sobreajuste*. Para evitarlo, se reserva una parte de los datos como un conjunto de pruebas, al que llamamos (X_test, y_test). El resto de los datos se utiliza para el entrenamiento, lo llamamos (X_train, y_train). Python tiene una función para separar al conjunto de datos en entrenamiento/prueba, llamada train_test_split(). Se utilizó esta función

designando un 10% de las muestras para la etapa de pruebas y 90% para el entrenamiento.

RESULTADOS

El fenotipo utilizado es la producción promedio de leche a 305 días ('*milk_le_ave_305*'), con un rango de valores $[-8.576,16.838]$. Para propósitos de clasificación, se transforman los valores del fenotipo a valores categóricos, asignándole la clase “no-lechera”, si el fenotipo fue ≤ 0, y la clase “lechera”, si el fenotipo fue >0. Como resultado, se tienen 670 vacas catalogadas como “no-lechera” y 422, como “lechera”.

Para evaluar el desempeño de cada algoritmo, se generaron diferentes subconjuntos conteniendo SNPs de grupos de cromosomas que contienen un alto número de QTLs relacionados a la producción lechera. Los *datasets* considerados se describen a continuación:

a) El conjunto completo de SNPs con los 29 cromosomas.
b) El conjunto formado con los SNPs correspondientes al cromosoma 14.
c) El conjunto formado por los SNPs de los cromosomas 6 y 14.
d) El conjunto formado por los SNPs de los cromosomas 5, 6 y 14.
e) El conjunto formado por los SNPs de los cromosomas 1, 5, 6 y 14.
f) El conjunto formado por los SNPs de los cromosomas 1, 5, 6, 14 y 20,
g) El conjunto formado por los SNPs de los cromosomas 1, 5, 6, 14, 17 y 20.
h) El conjunto formado por los SNPs de los cromosomas 1, 5, 6, 14, 17, 20 y 26.

i) El conjunto formado por los SNPs de los cromosomas 1 y 14.

La **Tabla *2*** presenta los resultados para el árbol de decisión, donde la primera columna indica el *dataset* utilizado, la segunda columna muestra la precisión de la clasificación y la tercera columna muestra el número de nodos en el árbol resultante.

Tabla 2. Precisión de la clasificación usando árboles de decisión.

Conjunto de datos	Precisión del conjunto de pruebas (%)	Núm. de nodos en el árbol resultante
a	93.6	67
b	91.8	117
c	91.8	107
d	90.9	93
e	90.9	95
f	93.6	93
g	90.9	85
h	94.5	85
i	94.5	97

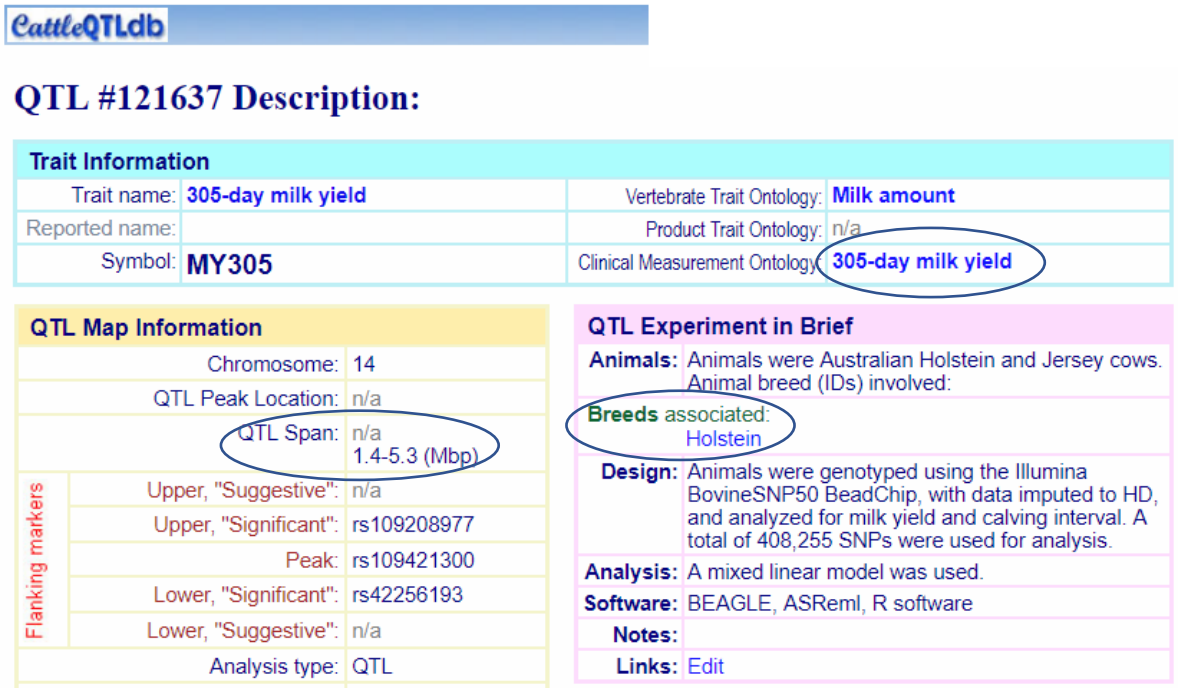
Como puede verse, en todos los casos se obtuvo una precisión superior al 90%. Los mejores resultados de la clasificación se obtuvieron para los *datasets* (h) e (i), de los cuales el *dataset* más pequeño es el (i) que incluye los cromosomas 1 y 14, alcanzando una precisión de 94.5%.

En cuanto al tiempo de procesamiento, con el conjunto de datos que incluye todos los cromosomas, se requirieron, aproximadamente, 30 horas de procesamiento. Cuando se utilizó el *dataset* que incluye solo al cromosoma 14, el tiempo de ejecución fue de, aproximadamente, 1.4 horas.

En Python, la clase *DecisionTreeClassifier* cuenta con un método llamado *feature_importances_* que devuelve la importancia de las características empleadas

en la clasificación (los valores de entrada). Con todos los conjuntos de datos, el algoritmo selecciona el SNP con posición 1455997 perteneciente al cromosoma 14 como el SNP más influyente. Su valor de influencia fue de aproximadamente 0.46, pudiendo variar ligeramente según el *dataset* considerado. Como una verificación adicional se calculó el coeficiente de correlación de Pearson entre todos los SNPs y el fenotipo de interés: el SNP con posición 1455997 tiene una correlación de 0.73 con el fenotipo, mientras que ningún otro SNP obtuvo una correlación más significativa de 0.30.

Con la intención de investigar si el SNP más influyente está asociado con un QTL asociado con la producción de leche, se realizó una búsqueda genómica en los QTL usando la base de datos *Cattle QTL database* (Cattle QTL database, https://www.animalgenome.org/cgi-bin/QTLdb/BT/index). El SNP en cuestión está dentro del QTL #121637 relacionado con la producción de leche a 305 días, en ganado de raza Holstein, considerando una expansión de 1.4 a 5.3 Mbp en el cromosoma 14 según puede verse en la **Figura *2***.



CattleQTLdb

QTL #121637 Description:

Trait Information

Trait name:	305-day milk yield	Vertebrate Trait Ontology:	Milk amount
Reported name:		Product Trait Ontology:	n/a
Symbol:	MY305	Clinical Measurement Ontology:	305-day milk yield

QTL Map Information

	Chromosome:	14
	QTL Peak Location:	n/a
	QTL Span:	n/a 1.4-5.3 (Mbp)
Flanking markers	Upper, "Suggestive":	n/a
	Upper, "Significant":	rs109208977
	Peak:	rs109421300
	Lower, "Significant":	rs42256193
	Lower, "Suggestive":	n/a
	Analysis type:	QTL

QTL Experiment in Brief

Animals:	Animals were Australian Holstein and Jersey cows. Animal breed (IDs) involved:
Breeds associated:	Holstein
Design:	Animals were genotyped using the Illumina BovineSNP50 BeadChip, with data imputed to HD, and analyzed for milk yield and calving interval. A total of 408,255 SNPs were used for analysis.
Analysis:	A mixed linear model was used.
Software:	BEAGLE, ASReml, R software
Notes:	
Links:	Edit

Figura 2. El SNP más influyente está en un QTL relacionado con la producción lechera a 305 días en ganado Holstein.

CONCLUSIONES

En este estudio se muestra el resultado de utilizar una técnica de ML para para la clasificación de vacas lecheras, específicamente se utilizan árboles de decisión.

Los árboles de decisión además de identificar con muy alta precisión a las muestras dadas, identifica con éxito el SNP más importante al hacer la clasificación. Esto puede conducir a ahorros económicos pues solo se requiere genotipificar al cromosoma 14 para obtener muy buenos resultados. Aunque ya se sabía que éste

es el cromosoma más relacionado con la producción lechera, no se sabía que era suficiente para determinar la clasificación.

El algoritmo de árboles de decisión estudiado es capaz de gestionar de manera efectiva la información del genoma completo bovino lo que lo hace adecuado para la implementación de herramientas de predicción de rasgos económicos en la industria lechera.

REFERENCIAS

[1] Cámara Nacional de Industriales de la Leche (CANILEC), “Estadísticas del sector lácteo 2010-2020,” 2021.

[2] G. R. Wiggans, J. B. Cole, S. M. Hubbard, and T. S. Sonstegard, “Genomic Selection in Dairy Cattle: The USDA Experience∗,” *Annu. Rev. Anim. Biosci.*, vol. 5, pp. 309–327, 2017.

[3] C. E. Rabier, P. Barre, T. Asp, G. Charmet, and B. Mangin, “On the accuracy of genomic selection,” *PLoS One*, vol. 11, no. 6, pp. 1–23, 2016.

[4] B. Hayes and M. Goddard, “Genome-wide association and genomic selection in animal breeding,” *Genome*, vol. 53, no. 11, pp. 876–883, 2010.

[5] M. E. Goddard and B. J. Hayes, “Genomic selection,” *J. Anim. Breed. Genet.*, no. 124, pp. 323–330, 2007.

[6] B. Li, N. Zhang, Y. G. Wang, A. W. George, A. Reverter, and Y. Li, “Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods,” *Front. Genet.*, vol. 9, no. JUL, pp. 1–20, 2018.

[7] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” vol. 349, no. 6245, 2015.

[8] S. B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques,” *Informatica*, vol. 31, pp. 249–268, 2007.

[9] T. Hastie, R. Tibshirani, and J. Friedman, “Unsupervised Learning,” in *The Elements of Statistical Learning*, 2009, pp. 486–585.

[10] Z. Chen, Y. Yao, P. Ma, Q. Wang, and Y. Pan, “Haplotype-based genome-wide association study identifies loci and candidate genes for milk yield in Holsteins,” *PLoS One*, vol. 13, no. 2, pp. 1–13, 2018.

[11] M. A. Cleveland, J. M. Hickey, and S. Forni, “A common dataset for genomic analysis of livestock populations,” *G3 Genes, Genomes, Genet.*, vol. 2, no. 4, pp. 429–435, 2012.

Cómo citar este artículo:
RODRIGUEZ ALCANTAR, E. (2022). Árboles de decisión para clasificación de vacas lecheras usando información genética. EPISTEMUS, 16(33). https://doi.org/10.36790/epistemus.v16i33.220