UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
INSTITUTO DE INGENIERÍA
MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA

Aplicación de algoritmos de aprendizaje automático para la clasificación de ganado lechero utilizando SNPs de genoma completo

TESIS QUE PARA OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS

PRESENTA

Edelmira Rodríguez Alcántar

Director
Rafael Villa Angulo

Co-director
Julio Waissman Vilanova

Mexicali, B. C., a 5 deagosto de 2019

Agradecimientos

A **José Luis Gómez Michel**, por tu apoyo incondicional.

Al Dr. Rafael Villa Angulo, por la confianza brindada y la oportunidad para realizar este trabajo.

Al Dr. Julio Waissman Vilanova, por el apoyo brindado en el área de aprendizaje automático.

A los sinodales, por sus valiosas observaciones en la elaboración de este documento.

Al programa PRODEP y a la Universidad de Sonora por el apoyo brindado para la realización de los estudios de posgrado.

Al Área de Cómputo de Alto Rendimiento de la Universidad de Sonora (ACARUS), por el apoyo con el uso del hardware y software, así como el apoyo técnico brindado.

Resumen

A pesar de encontrarse entre los 15 principales productores de leche en el mundo, la balanza comercial de México en notablemente desfavorable, importando un monto casi 3 veces mayor de productos lácteos de lo que se exporta en el país. Vale la pena invertir para mejorar ese panorama pues ello puede llevar a un incremento en la fuente de empleos, entre otros beneficios. Si bien las técnicas de reproducción asistida han hecho una mejora dramática en la productividad de los animales en diversas áreas, para la producción lechera existe el problema que se requiere esperar varios años antes de medir el fenotipo lechero de una vaquilla. Cualquier mejora que pueda realizarse en ese sentido (disminuir el tiempo para conocer el fenotipo de interés y disminuir el intervalo generacional) será bien recibido. La selección genómica intenta predecir el fenotipo de un individuo partir de su información genética, así que puede usarse para lograr un hato de ganado que sea considerado genéticamente como buen productor de leche.

Este trabajo doctoral contribuye al estado del arte sobre el uso de técnicas de aprendizaje automático para la tipificación temprana del fenotipo lechero de vaquillas, a partir del uso de marcadores genéticos, con el fin de 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.

Se realizó un estudio sobre la eficiencia de dos algoritmos de aprendizaje automático (árboles de decisión y redes neuronales artificiales) para la clasificación de vacas lecheras. Se obtuvieron resultados alentadores, en particular con los árboles de decisión, logrando

hasta un 94.5% de precisión. Además, el algoritmo permitió la identificación del SNP más dominante para la clasificación, así como del cromosoma que más influye en la predicción.

Los algoritmos de aprendizaje automático estudiados son capaces de gestionar de manera efectiva la información del genoma completo bovino lo que los hace adecuados para la implementación de herramientas de predicción de rasgos económicos en la industria lechera.

Summary

Despite being among the 15 leading milk producers in the world, Mexico's trade balance is notably unfavorable, importing an almost 3 times higher amount of dairy products than it is exported in the country. It is worth investing to improve that panorama as this can lead to an increase in the source of jobs, among other benefits. Although the techniques of assisted reproduction have made a dramatic improvement in the productivity of animals in various areas, for milk production there is the problem that it is required to wait several years before measuring the dairy phenotype of a heifer. Any improvement that can be made in that sense (decrease the time to know the phenotype of interest and decrease the generational interval) will be well received. Genomic selection attempts to predict the phenotype of an individual from their genetic information, so it can be used to achieve a herd of cattle that are genetically considered a good milk producer.

This doctoral work contributes to the state of the art on the use of automatic learning techniques for the early typing of the dairy phenotype of heifers, from the use of genetic markers, in order to make a selection of genetically superior animals in less time and make more efficient the process of assisted reproduction, thus achieving lower costs and increasing profits in the dairy sector.

A study on the efficiency of two automatic learning algorithms (decision trees and artificial neural networks) for the classification of dairy cows was conducted. Encouraging results were obtained, particularly with decision trees, achieving up to 94.5% accuracy. In addition, the algorithm allowed the identification of the most dominant SNP for classification, as well as the chromosome that most influences the prediction.

The machine learning algorithms studied are able to effectively manage the complete bovine genome information, which makes them suitable for the implementation of economic trait prediction tools in the dairy industry.

Tabla de contenido

Agradecimientos i
Resumen ii
Summary iv
Lista de figuras viii
Lista de tablas ix
Glosario x
1 Introducción 1
1.1 Genotipo y fenotipo 5
1.2 Equilibrio de Hardy-Weinberg 5
1.3 Frecuencia del alelo menor 6
1.4 Marcador genético 7
1.5 Estudio de asociación de genoma completo 8
1.6 Selección genómica 9
1.7 Aprendizaje automático 11
1.7.1 Árboles de decisión 12
1.7.2 Redes neuronales artificiales tipo *feedforward-backpropagation* 13
1.8 Planteamiento del problema y objetivos 17
2 Métodos y resultados 19
2.1 Métodos 19
2.1.1 Conjunto de datos utilizado 19
2.1.2 Filtros de control de calidad 22
2.1.3 Especificaciones de hardware y software 22
2.1.4 Análisis de componentes principales 22
2.1.5 Procesamiento de los datos 24
2.2 Resultados 31
2.2.1 Árboles de decisión 33
2.2.2 Redes neuronales 39
3 Discusión 43
4 Conclusiones y trabajo futuro 47
4.1 Conclusiones 47

4.2 Trabajo futuro 49
Referencias 50
Apéndices 55
A 56
B 57

Lista de figuras

Figura 1. Producción mundial anual de leche de bovino por países, 2017. _____ *1*

Figura 2. . Producción anual de leche de bovino en México (en millones de litros) [1]. _____ *2*

Figura 3. Producción de leche de bovino anual por entidad federativa 2018 (en millones de litros) [1]. _____ *3*

Figura 4. Un cambio en el genotipo puede representar un cambio en el fenotipo. _____ *7*

Figura 5. Una red neuronal tipo feedforward con tres capas. _____ *15*

Figura 6. Bosquejo del algoritmo de entrenamiento de una red neuronal. _____ *16*

Figura 7. Gráfica MCA con datos del cromosoma 14. Las marcas azules representan a los ejemplares con alta producción lechera, y las marcas rojas, a los bajos productores. _____ *24*

Figura 8. Bosquejo del entrenamiento usando GridSearchCV. _____ *28*

Figura 9. Variación del resultado al cambiar random_state en el árbol de decisión. _____ *36*

Figura 10. El SNP más influyente está en un QTL relacionado con la producción lechera a 305 días en ganado Holstein. _____ *37*

Figura 11. Listado de genes cercanos a la posición 1 455 997 en un rango de 110 000 bp. _____ *38*

Figura 12. Espacio de soluciones obtenidas (score) con redes neuronales. Los ejes representan la semilla, el número de neuronas por capa y la exactitud de la predicción obtenida por el algoritmo. _____ *42*

Lista de tablas

Tabla 1. Valor de la producción nacional de leche de bovino [1]. 3

Tabla 2. Muestra del conjunto de datos con información parcial del cromosoma 14. 20

Tabla 3. Número de QTLs relacionados con la producción lechera y número de SNPs, por cromosoma. 21

Tabla 4. Conjuntos de datos utilizados. 32

Tabla 5. Número de SNPs contenidos en cada conjunto, antes y después del filtrado de QC. 33

Tabla 6. Precisión de la clasificación usando árboles de decisión. 34

Tabla 7. Mejores parámetros para el árbol de decisión. 34

Tabla 8. Matriz de confusión arrojada por el clasificador DT para el conjunto_7. 35

Tabla 9. Resultados de la clasificación usando árboles de decisión, sin considerar al cromosoma 14. 39

Tabla 10. Precisión de la clasificación usando NN. 40

Tabla 11. Mejores parámetros para la red neuronal. 41

Tabla 12. Concentrado de los resultados obtenidos con los métodos utilizados. 44

Glosario

GWAS Genome-wide Association Study

HWE Hardy-Weinberg Equilibrium

MAF Minor Allele Frequency

ML Machine Learning

DT Decision Tree

ANN Artificial Neural Network

SNP Single Nucleotide Polymorphisms

EBV Estimated Breeding Value

GEBV Genetic Estimated Breeding Value

QTL Quantitative Trait Loci

GS Genetic Selection

CV Cross-Validation

mdd millones de dólares

bp base par

1 Introducción

En este capítulo se da un bosquejo del estado de la producción lechera en México y en Baja California, según estadísticas oficiales y actualizadas a la fecha.

Con base en las estadísticas obtenidas de la Cámara Nacional de Industriales de la Leche [1], se sabe que en el 2017, México ocupaba el 14º lugar en la producción mundial de leche de bovino, con el 2% de la producción total. En la Figura 1 se ilustra esta información, para dar una mejor idea de la proporción que esto representa.

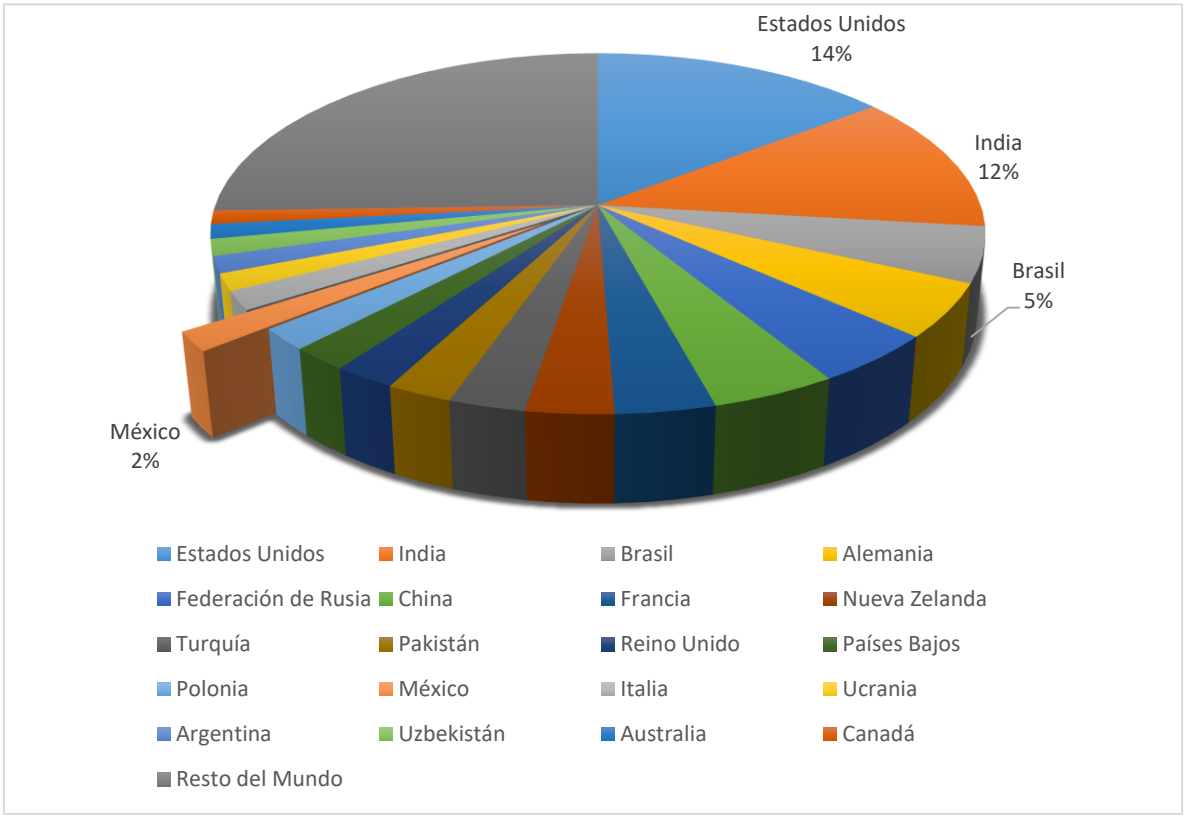


Figura 1. Producción mundial anual de leche de bovino por países, 2017.

En la Figura 2, se ilustra la producción anual de leche de bovino en México en el período de 1990-2018, la cual va de 6,141,545,000 litros en 1990 a 12,008,286,000 litros en 2018. Como puede observarse, se ha duplicado la producción en ese período.

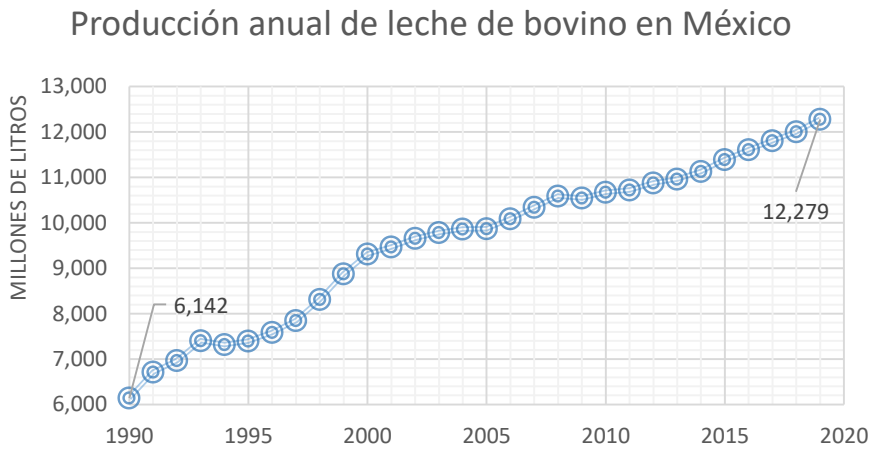


Figura 2. . Producción anual de leche de bovino en México (en millones de litros) [1].

En la Tabla 1, se muestra el valor económico de la producción de leche de bovino en México durante el período 2010-2017; por ejemplo en el 2017, esta actividad representó un valor de $70 660 026 000 pesos.

En lo que respecta a la producción local, en el estado de Baja California se produjeron 186 139 000 litros de leche durante el 2018, esta cantidad representa <2% de la producción nacional, tal como se ilustra en la Figura 3.

Tabla 1. Valor de la producción nacional de leche de bovino [1].

Año	Valor de la producción nacional de leche de bovino (en miles de pesos)
2010	$50 801 773
2011	$53 029 602
2012	$56 445 380
2013	$60 678 408
2014	$65 000 180
2015	$66 969 586
2016	$67 751 168
2017	$70 660 026

México importa anualmente 2 011 mdd de productos lácteos, exporta 722 mdd y tiene un déficit comercial de 1 289 mdd en este sector. Como puede verse, la producción lechera en México tiene una oportunidad de crecimiento importante al pretender cubrir el déficit de producción nacional o incluso aumentar las exportaciones. Al incrementar la inversión en esta actividad comercial puede incrementarse también el número de empleos.

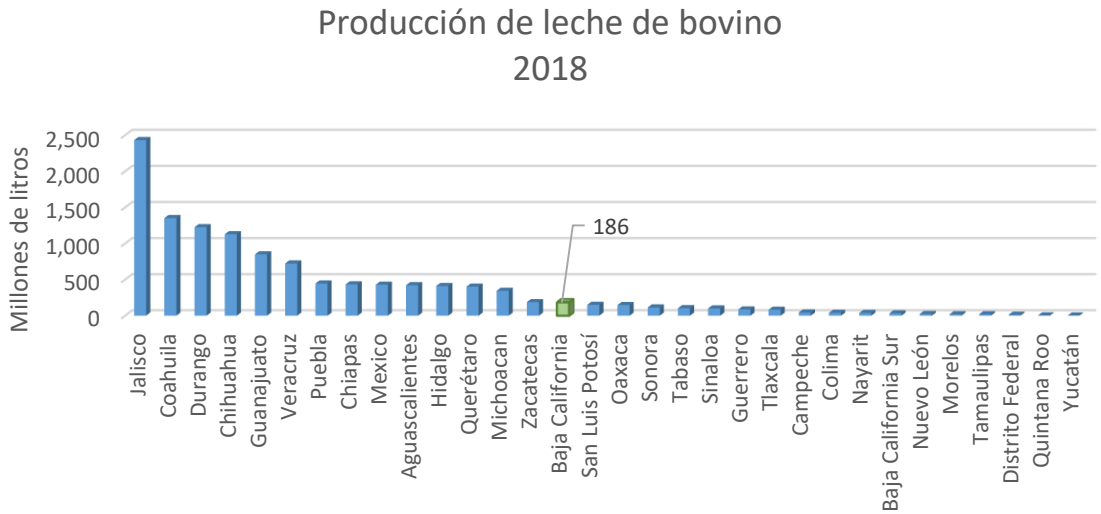


Figura 3. Producción de leche de bovino anual por entidad federativa 2018 (en millones de litros) [1].

La actividad lechera es compleja y hay diversas áreas en las que podrían enfocarse los esfuerzos con vistas a mejorar la producción. Estas áreas comprenden mejorar la alimentación del ganado, la construcción de hábitats con ambiente controlado, el cuidado de la salud animal y la crianza selectiva de los animales. Todo esto con miras a obtener un hato en las mejores condiciones para la producción lechera.

La crianza selectiva puede llevarse a cabo de la manera tradicional (usando el pedigrí y el fenotipo) o apoyándose en la información genética del animal. El método tradicional de selección genómica incluye la evaluación de la progenie del animal (ascendencia y descendencia). Aunque las técnicas de reproducción asistida (tales como inseminación artificial, múltiple ovulación y transferencia de embriones) han hecho una mejora dramática en la productividad de los animales [2], aún no es suficiente, sobre todo en lo que respecta a la producción lechera ya que es un fenotipo que requiere esperar varios años antes de medirse. Para evaluar el fenotipo lechero de una vaquilla debe esperarse aproximadamente 3 años desde su nacimiento (hay que esperar a que alcance la madurez sexual después del primer año de vida, aproximadamente un año de preñez y un año más para evaluar su producción lechera –aproximadamente 305 días). Para hacer una evaluación estimada del fenotipo a partir de los padres, debe esperarse al menos 5 años para evaluar al padre, y aproximadamente 3 años para evaluar a la madre. Esta valoración debería de ser más precisa utilizando información sobre la variación en la secuencia de ADN entre los animales [3], y podría ayudar a seleccionar ejemplares genéticamente superiores y en menor tiempo.

1.1 Genotipo y fenotipo

La genética es la ciencia que se encarga del estudio de la variación de la herencia de los organismos vivos [4]. 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 se caracterizan a un organismo constituyen su fenotipo (por ejemplo, la forma de los ojos, el largo de las orejas y el color del pelaje). El genotipo y el fenotipo están fuertemente relacionados, y el medio ambiente de desarrollo del individuo también es relevante en la expresión del fenotipo. Esta relación se representa mediante la siguiente expresión

fenotipo = genotipo + medioambiente

donde el signo '+' representa una interacción en lugar de una operación de suma.

1.2 Equilibrio de Hardy-Weinberg

En un solo locus autosomal con dos alelos, un individuo diploide puede tener uno de tres genotipos posibles: AA, Aa o aa [5]. Donde A corresponde al alelo mayor (más frecuente) y a al alelo menor (menos frecuente). El equilibrio de Hardy-Weinberg (HWE) hace referencia a la independencia de alelos en un locus entre los dos cromosomas homólogos. En concreto establece la relación entre las frecuencias alélicas y genotípicas en una población lo suficientemente grande, sin selección, mutación o migración. Si consideramos un locus con dos alelos A y a la probabilidad de que aparezca el alelo A es $Pr(A) = p$ y la probabilidad del alelo a es $Pr(a) = q = 1 - p$, si los dos alelos que porta un individuo se heredan independientemente, las probabilidades esperadas de los genotipos serán $Pr(AA) = p^2$,

$Pr(Aa) = 2pq$ y $Pr(aa) = q^2$, respectivamente. La suma de las probabilidades de cada combinación posible es $p^2 + 2pq + q^2 = 1$.

En un estudio de asociación es necesario verificar que los SNPs genotipificados se encuentran en equilibrio de Hardy-Weinberg, ya que desviaciones de este equilibrio pueden ser indicativas de errores de genotipificación o bien de la existencia de factores que podrían afectar a las frecuencias alélicas (por ejemplo, migración reciente) diferentes del rasgo en estudio. Si no se examina puede dar lugar a asociaciones espurias. Se considera que una población está en equilibrio de Hardy-Weinberg si la estadística de una prueba de igualdad de proporciones Chi-Cuadrada es menor o igual a 3.84 ($\chi^2 \leq 3.84$). Por convención, todos los SNPs que no están en equilibrio de Hardy-Weinberg son eliminados de un estudio de asociación.

1.3 Frecuencia del alelo menor

La frecuencia del alelo menor (*Minor Allele Frequency*, MAF), se define como la frecuencia del alelo menos común en un determinado loci, en una población. Para considerar una variación de nucleótido simple como un SNP, el alelo menor debe estar presente en, al menos, el 5% de la población. Aquellos alelos con MAF ≤ 0.05 son considerados raros, y son candidatos a ser mutaciones. Por convención, todos aquellos SNPs con un MAF ≤ 0.05 son considerados como monomórficos y son eliminados, y, por ende, no son utilizados en los estudios de asociación [6][7].

1.4 Marcador genético

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. El mayor desafío que enfrentan los genetistas es identificar los marcadores de genes que controlan la variación fenotípica en rasgos de interés [2]. Los marcadores de polimorfismo de nucleótido simple (*Single Nucleotide Polymorphisms*, SNP), son atractivos pues son abundantes, genéticamente estables y apropiados para análisis automático de alto rendimiento [8]. Aunque los SNPs por sí solos no proporcionan información sobre genes específicos, indican una localización cromosómica con probabilidades de estar asociada con un fenotipo dado. Esto se ilustra en la Figura 4.

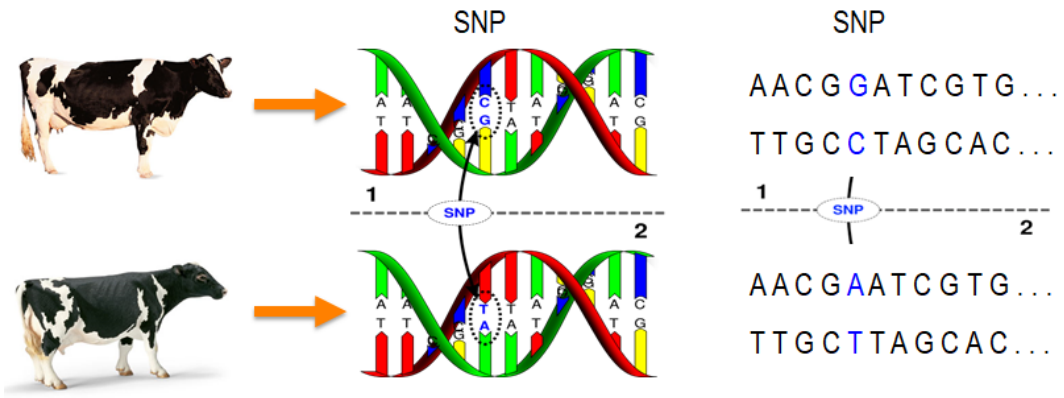


Figura 4. Un cambio en el genotipo puede representar un cambio en el fenotipo.

En [9] se propone un método para la predicción de valores de crianza individuales sin fenotipos pero que han sido genotipados con un panel de marcadores de alta densidad. En

diciembre de 2007, se liberó el chip BovineSNP50 BeadChip[1] 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 [10]. Actualmente, existen kits de análisis de ADN bovino con más de 770 000 SNPs. La secuenciación del ganado bovino ha hecho que gran cantidad de marcadores genéticos estén disponibles en la forma de SNPs. La disponibilidad de esta información y el desarrollo de la tecnología han hecho posible la selección asistida por marcadores y los estudios de asociación de genoma completo sean llevados a cabo en poblaciones de bovinos.

1.5 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. La mayoría de los GWAS se centran en la asociación entre los SNPs y algún rasgo de interés [11]. Un GWAS permite a los investigadores analizar individuos sin conocer su pedigrí [12].

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.

[1] Illumina, San Diego, CA

En general, los estudios GWAS actúan como una herramienta de selección inicial, desde la cual las regiones significativamente asociadas pueden refinarse aún más utilizando una mayor densidad de marcadores (potencialmente por re-secuenciación), con el objetivo final de identificar genes candidatos que se cree subyacen en el rasgo de interés. Los genes candidatos se pueden caracterizar aún más en un intento de identificar los mecanismos funcionales subyacentes a un rasgo [13]. Esto conducirá a una mayor comprensión del rasgo en cuestión.

No obstante, este método ha demostrado tener debilidades en los siguientes casos:

- El número de participantes no es lo suficientemente elevado.
- El grupo de casos no es representativo.
- No se siguen las metodologías establecidas con estricto control.

Cuando alguna de estas situaciones ocurre, se producen sesgos que comprometen el estudio.

1.6 Selección genómica

La selección genómica es una forma de selección asistida por marcadores en la que se utilizan marcadores genéticos que cubren todo el genoma, de modo que todos los loci de rasgos cuantitativos (QTL) están en desequilibrio de enlace con al menos un marcador. Este enfoque se ha vuelto factible gracias a la gran cantidad de SNPs descubiertos por la secuenciación del genoma y los nuevos métodos para obtener, de manera eficiente, el genotipo de gran cantidad de SNPs.

La domesticación del ganado ocurrió hace aproximadamente 10 000 años, y la selección se ha practicado desde entonces con el propósito de mejorar la salud de los animales, su apariencia física y su producción [6]. Tradicionalmente, la selección de los

animales se ha basado en su historial familiar y en el fenotipo del animal [14]. Esta información es usada para calcular el valor de crianza estimada (*Estimated Breeding Value*, EBV) de cada animal, alcanzando una confiabilidad del 15-28% en la predicción de rasgos asociados con la producción lechera [15].

Después de la secuenciación de ganado bovino [16], [17], se ha desarrollado una revolución de las tecnologías de genotipado de alto rendimiento posibilitando la inspección de miles de SNPs en el genoma completo [18], [19]. 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) [20], [21].

Comparados a los esquemas de crianza tradicionales, la selección genética permite una selección más eficiente de los animales, guiados por rasgos de baja heredabilidad. Además, la GS permite la evaluación de un número más alto de candidatos para la crianza, así como la intensificación de la crianza reduciendo el intervalo generacional [22] pues el genotipado puede hacerse desde una edad muy temprana, al momento del nacimiento o incluso durante la gestación [23]. En [15], Meawissen y colaboradores reportaron que la confiabilidad de los GEBVs varía de 44 a 49% para rasgos asociados con la producción lechera. Más aún, en [24] se reporta una confiabilidad de ~60% usando selección genómica en ganado lechero.

Una cuestión importante de GS es que los efectos QTL y los GEBV deben calcularse primero para una gran población de referencia con información genotípica y fenotípica, mientras que en las generaciones posteriores solo se requiere información de marcadores para

estimar el GEBV [25]. 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 [3]. Aunque con el tiempo, la tecnología para el genotipado ha avanzado mucho, y el precio ha disminuido, aún los costos son altos para los países subdesarrollados, como México.

En este trabajo, se propone utilizar estrategias de aprendizaje automático para predecir si una vaca será lechera a partir de su información genética.

El análisis de grandes datos genómicos se ve obstaculizado por problemas como un pequeño número de observaciones y un gran número de variables predictivas (comúnmente conocidas como "gran P pequeña N"), alta dimensionalidad o estructuras de datos altamente correlacionadas. Los métodos de aprendizaje automático son famosos por tratar estos problemas [26].

1.7 Aprendizaje automático

El aprendizaje automático (también llamado *Machine Learning*, 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 [27]. ML tiene dos categorías principales: métodos de aprendizaje supervisados [28] y métodos de aprendizaje no supervisados [29]. 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 [30].

En este documento, se muestra que el ML se puede usar en la selección de ganado, como se sugiere en [31], para la estimación del fenotipo a la que pertenece una vaca (lechera / no lechera). Esta predicción se obtiene mediante algoritmos ML aplicados a un conjunto de datos con genotipos y fenotipos conocidos, es decir se utiliza aprendizaje supervisado. En particular, se utilizan las técnicas de árboles de decisión [32] y de red neuronal artificial [33] [34].

1.7.1 Árboles de decisión

Un árbol de decisión (*Decision Tree*, DT) es una herramienta eficiente para la solución de problemas de clasificación [35]. Un DT consiste de nodos que forman un árbol enraizado; es decir, se trata de un árbol dirigido con un nodo llamado “raíz” que no tiene arcos de entrada y donde todos los otros nodos tienen exactamente un arco de entrada. Un nodo “interno” es un nodo con arcos de salida, el resto de los nodos son llamados “hojas” [32]. Durante el proceso de aprendizaje utilizando un DT para aprendizaje categórico, las muestras en cada nodo interno son separadas en subconjuntos con base en un atributo, y el proceso se repite en cada subconjunto de los nodos así obtenidos en una manera recursiva llamada “particionamiento recursivo”. El proceso finaliza cuando se cumple alguna de las siguientes condiciones [36]:

- El subconjunto de una muestra en un nodo tiene el mismo valor objetivo.
- Cuando la división no mejora la clasificación.

- Cuando la separación es imposible debido a restricciones impuestas por el usuario.

Algunas ventajas del uso de árboles de decisión se presentan a continuación:

- Son simples de entender e interpretar y pueden ser visualizados.
- El costo es logarítmico en el número de datos usados para el entrenamiento.
- Son capaces de manejar datos numéricos y categóricos.
- Soportan problemas de salida múltiple.

Entre las desventajas de los árboles de decisión se encuentran:

- Pueden crear árboles demasiado complejos que no generalicen bien, situación conocida como sobreajuste (*overfitting*).
- Los DT pueden ser inestables debido a que pequeñas variaciones en los datos pueden resultar en la generación de un árbol completamente diferente.
- El problema de aprendizaje con árbol de decisión óptimo es NP-completo bajo diversos conceptos de optimalidad y aún sobre conceptos simples [37]. Como consecuencia, los algoritmos están basados en heurísticas tales como algoritmos voraces (*greedy*) donde se toman decisiones de optimalidad locales en cada nodo. Tales algoritmos no garantizan que el árbol global sea óptimo.
- Los DT pueden estar sesgados si las clases están desbalanceadas.

1.7.2 Redes neuronales artificiales tipo *feedforward-backpropagation*

Las redes neuronales artificiales están inspiradas en la capacidad del cerebro para aprender patrones de datos complicados cambiando el valor de las conexiones sinápticas entre las neuronas.

Una red neuronal artificial (*Artificial Neural Networks*, ANN) tipo *feedforward* de los años 60's requerían un ajuste manual de los parámetros de la red, con lo cual el entrenamiento se volvía una tarea tediosa. En 1986 apareció el procedimiento de *backpropagation* con el cual solo se tenía que alimentar a la red con los valores iniciales pues automáticamente se hacían los ajustes necesarios para modelar complicadas relaciones entre la entrada y la salida de la red, esta es una característica especialmente útil si se utilizan varias capas intermedias [38]. En cierto sentido, el entrenamiento depende de los valores iniciales de los pesos de las aristas, pues los valores finales de los pesos de las aristas al terminar la etapa de entrenamiento serán diferentes si se cambian estos valores, esto será evidente en la sección 2.2.

Aunque por un tiempo dejaron de utilizarse, en épocas recientes las redes neuronales han cobrado fuerza debido, entre otras cosas, a la aparición de equipo de cómputo con mejores características en cuanto a velocidad de procesamiento y capacidad de almacenaje; y a que los algoritmos mismos han sido mejorados; con todo esto, el entrenamiento de la red neuronal se ha vuelto más fácil.

Una red neuronal tipo *feedforward*, también llamada perceptrón multicapa, tiene como objetivo el aproximar alguna función $f^*(X, \theta)$ y aprender los valores de los parámetros θ que resultan en la mejor función de aproximación que resuelva el problema considerado. Se trata del modelo de red neuronal más común; es un modelo supervisado pues requiere de un conjunto de datos de entrenamiento para aprender [39]. Una red neuronal tipo *feedforward* consiste de un número n de neuronas, organizadas en m capas, donde $m, n \in \mathrm{Z}^+$, cada una de las cuales tiene una cantidad dada de neuronas. Las capas se conectan para permitir el flujo de información de una capa a la siguiente a través de conexiones con arcos que van de cada neurona en una capa a todas las neuronas en la capa siguiente, de ahí el nombre de

feedforward. El flujo comienza cuando una muestra, $X = (x_1, x_2, \dots, x_{n0})$ que en este caso está formado por los valores SNPs de cada muestra $(SNP_1, SNP_2 \cdots, SNP_{n_0})$, entra a la capa de entrada, pasa a través de cada capa oculta y finaliza cuando alcanza la capa de salida (ver la Figura 5).

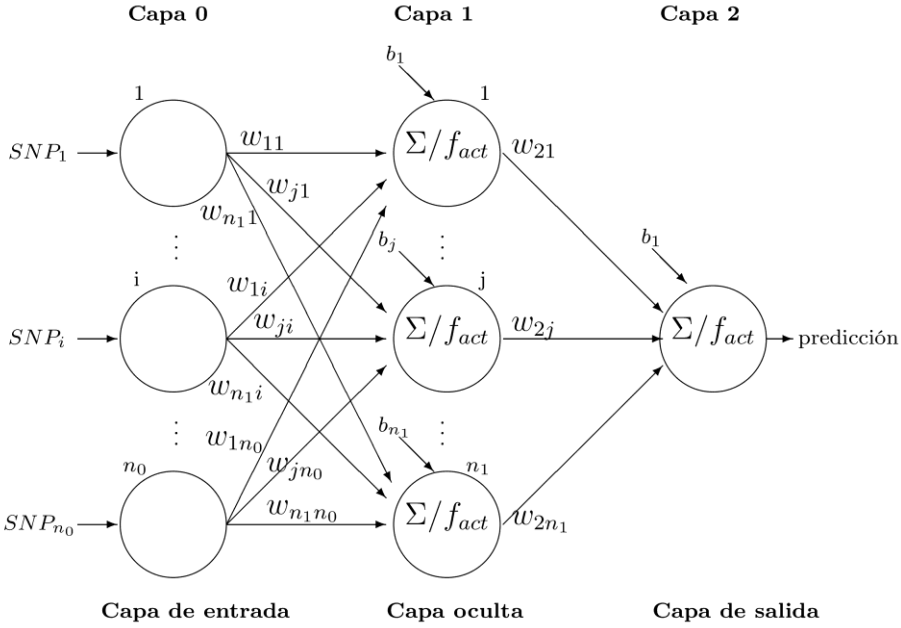


Figura 5. Una red neuronal tipo feedforward con tres capas.

La salida de la neurona i (en una capa diferente a la capa inicial), es el valor que arroja la función de activación para la suma de todos los valores de entrada a la neurona i, más un sesgo, $f_{act}(\sum_{i=1}^{n_j} x_j \cdot w_{ij} + b_i)$. Este valor será enviado a cada una de las neuronas en la capa siguiente. La función de activación usada en las neuronas de las capas ocultas puede ser diferente a la usada en las neuronas de la capa de salida dado que debe cuidarse que la salida esté en el rango de valores requerido. Los valores de $W = [w_{ij}]$ corresponden a los pesos de los arcos desde la neurona *i* a la neurona *j* (en capas consecutivas), y se inicializan con valores aleatorios, al igual que la matriz de sesgos, b.

El entrenamiento de una red neuronal requiere de un número de iteraciones (llamadas épocas), donde cada época consta de una etapa de *feedforward* y una de *backpropagation*. En la etapa de *backpropagation*, se calcula el gradiente del error obtenido entre el valor estimado y el valor esperado respecto al vector de parámetros θ el cual se utiliza para modificar la matriz de pesos y la matriz de sesgos desde la capa de salida hasta la primera capa oculta, de ahí el nombre de *backpropagation*. El entrenamiento termina cuando se alcanza el número de épocas especificado o cuando el error de entrenamiento es menor a un cierto umbral. Un bosquejo del algoritmo de aprendizaje en NN se ilustra en la Figura 6.

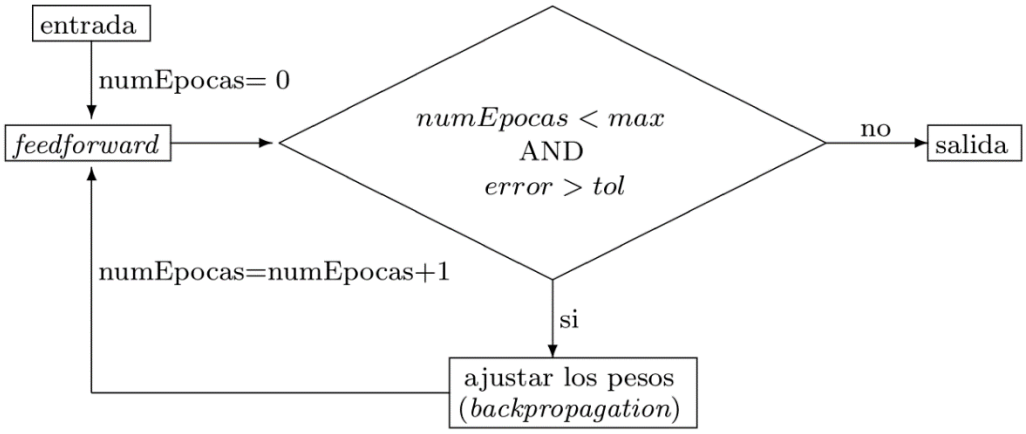


Figura 6. Bosquejo del algoritmo de entrenamiento de una red neuronal.

La ventaja de utilizar una red neuronal es su capacidad de aprender modelos no-lineales. Entre las desventajas se encuentran que las redes neuronales multicapa tienen una función de pérdida no convexa donde pueden existir más de un mínimo local, por lo tanto, al inicializar las matrices de los pesos y los sesgos con diferentes valores puede obtenerse resultados diferentes.

Al utilizar las redes neuronales multicapa se deben ajustar diversos parámetros, tales como el número de capas ocultas, el número de neuronas en cada una de las capas ocultas y el número de iteraciones. La selección hecha de estos y otros parámetros repercute en el tiempo necesario para realizar el entrenamiento.

1.8 Planteamiento del problema y objetivos

La producción lechera en México es insuficiente para satisfacer las necesidades de la población. Más aún, el estado de Baja California produce apenas un ~8% de la producción lechera del estado de Jalisco, el estado con mayor producción lechera en el país. Es necesario realizar acciones que apoyen a la ganadería con miras a incrementar la producción lechera en el estado. Si bien, hay diversas maneras de apoyar, una estrategia es aprovechar el desarrollo de la tecnología permite obtener el genotipo de los ejemplares con un número de SNP cada vez mayor, además de que se ha desarrollado hardware y software capaces de procesar esa información cada vez más rápido. La motivación es apoyar a la industria ganadera a reunir un hato donde la propensión genética de cada animal sea la de un alto productor de leche.

Mientras que los GWAS tienen como objetivo identificar marcadores (o regiones de marcadores) que puedan explicar un fenotipo dado, y que la selección genómica asigna un GEBV a cada ejemplar e identifica así a los que tienen puntuación más alta para un fenotipo de interés, en este trabajo se tiene una perspectiva diferente: el objetivo es predecir la clase de una vaquilla a partir de su genotipo, basando la clase de la vaquilla en la producción lechera para clasificarla como alta productora de leche o baja productora de leche, usando técnicas de aprendizaje automático supervisado.

El aprendizaje automático supervisado tiene como objetivo construir un modelo genotipo-fenotipo mediante el aprendizaje de dichos patrones genéticos a partir de un conjunto etiquetado de ejemplos de entrenamiento que también proporcionarán predicciones fenotípicas precisas en casos nuevos con antecedentes genéticos similares [40].

Con ello, este trabajo contribuye al estado del arte sobre el uso de técnicas de aprendizaje automático para la tipificación temprana del fenotipo lechero de vaquillas, a partir del uso de marcadores genéticos, con el fin de 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.

Objetivo general

El objetivo general es investigar el potencial del aprendizaje automático para el análisis genético con el cual se puedan realizar predicciones en ganado bovino raza *Holstein*, específicamente sobre la producción lechera, partiendo de marcadores SNP a lo largo del genoma completo.

Los **objetivos específicos** son los siguientes:

1. Obtener un panel de marcadores SNPs del genoma completo de bovinos.
2. Aplicar la técnica de árboles de decisión para clasificar las muestras.
3. Aplicar la técnica de redes neuronales artificiales para clasificar las muestras.
4. Documentar el desempeño de las dos técnicas de aprendizaje automático para resolver el problema de clasificación.

2 Métodos y resultados

En este capítulo se presenta una descripción del conjunto de datos utilizados, los filtros de calidad aplicados al conjunto y una descripción de la implementación realizada. También se presentan los resultados obtenidos.

2.1 Métodos

2.1.1 Conjunto de datos utilizado

El conjunto de datos usado[2] en este trabajo fue obtenido de [11]. Los datos consisten de muestras genotípicas de 1092 vacas Holstein, en un panel de 164 312 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. En la Tabla 2 se muestra el contenido del archivo con información parcial del cromosoma 14; la columna V1 indica el cromosoma, la columna V2 indica la posición en el genoma, las columnas V3 y V4 indican los nucleótidos que conforman el SNP, a partir de V5 se encuentra una matriz de valores 0, 1 y 2, que sirve entrada para los métodos utilizados. Se cuenta con medidas de fenotipos para diferentes rasgos y se seleccionó el EBV de producción lechera promedio a 305 días ('*milk_le_ave_305*').

[2] El *dataset* está disponible en https://doi.org/10.5061/dryad.cs133

Tabla 2. Muestra del conjunto de datos con información parcial del cromosoma 14.

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
chr14	14915	C	T	0	0	0	0	0	0	0	0	0	0
chr14	91936	G	T	1	0	1	0	1	1	1	0	0	1
chr14	111990	T	C	1	0	1	0	1	1	1	2	1	1
chr14	114632	T	G	2	0	2	1	1	2	1	2	1	2

El conjunto de datos contiene 29 cromosomas 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 *Bovine QTL database*[3]. A la fecha hay reportados un total de 45 525 QTL como asociados con la producción lechera, de los cuales 4 226 están ligados con la cantidad de producción lechera a lo largo de todo el genoma. La Tabla 3 presenta, para cada cromosoma, el número de QTLs relacionado con la producción lechera y el número de SNPs muestreados en el panel. 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 descritas en la sección Resultados.

[3] CattleQTLdb [2019]. QTL/Associations distribution analysis tool [online], search for traits with a key word(s): *milk yield*. [consultado: 25-mayo-2019].
Website https://www.animalgenome.org/cgi-bin/QTLdb/BT/oview?submit=Search&qtype=QTL&srchtr=milk+yield

Tabla 3. Número de QTLs relacionados con la producción lechera y número de SNPs, por cromosoma.

Cromosoma	Número de SNPs	Número de QTLs	Cromosoma	Número de SNPs	Número de QTLs
1	7 338	23	**16**	5 269	10
2	7 049	16	**17**	4 750	22
3	8 064	21	**18**	7 579	10
4	7 572	12	**19**	6 108	16
5	7 733	31	**20**	3 395	25
6	5 312	36	**21**	5 871	20
7	7 465	18	**22**	3 765	5
8	6 088	8	**23**	4 548	18
9	5 273	10	**24**	3 622	1
10	5 952	14	**25**	5 773	6
11	7 120	7	**26**	3 536	22
12	6 640	8	**27**	3 492	8
13	6 736	13	**28**	3 680	7
14	4 004	51	**29**	5 004	9
15	5 574	4			

2.1.2 Filtros de control de calidad

Como estrategia de control de calidad (*Quality Control*, QC), se aplicaron ciertos filtros a los datos iniciales 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 más de 20% de genotipos faltantes.
2. Todos aquellos SNPs que violen la distribución de frecuencias de Hardy-Weinberg, como se aplica en [1].
3. Todos aquellos SNPs que tienen una MAF < 5%.

2.1.3 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). Una de las ventajas de este lenguaje es que está desarrollado bajo el esquema *open source*, es decir, permite que el software sea libremente usado, modificado y distribuido. Además, incluye el paquete *scikit-learn* [41] con herramientas simples y eficientes para el aprendizaje automático.

2.1.4 Análisis de componentes principales

Una de las herramientas más valiosas del álgebra lineal aplicada es el análisis de componentes principales (*Principal Component Analysis*, PCA) [42]. Formalmente, PCA está definido como una transformación lineal ortogonal que transforma los datos a un nuevo sistema de coordenadas de tal forma que la mayor varianza en alguna de las proyecciones de

los datos está en la primera coordenada (llamada componente principal), la segunda mayor varianza de los datos está en la segunda coordenada, y así sucesivamente.

Teóricamente, PCA es la transformación óptima de los datos dados en términos de mínimos cuadrados. PCA puede ser usado para reducción de la dimensionalidad en un conjunto de datos reteniendo aquellas características que tienen una mayor varianza, manteniendo los componentes principales de bajo orden e ignorando los de mayor orden. Frecuentemente, los componentes de bajo orden contienen los aspectos "más importantes" de los datos [17], [43].

El procedimiento para realizar un PCA puede resumirse como sigue: Dada una matriz X^T de m muestras con n características (dimensiones),

$$X = \begin{pmatrix} x_1^{(1)} & \cdots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(m)} & \cdots & x_n^{(m)} \end{pmatrix}$$

cuyo vector promedio M y covarianza C son descritos por $E(X) = [m_1, m_2, \dots, m_n]^T$ y $C = E((X - M)(X - M)^T)$, respectivamente. Se calculan los eigenvalores $\lambda_1, \lambda_2, \dots, \lambda_n$ y los eigenvectores $P_1, P_2, \dots, P_d$ de la matriz de covarianza C; se ordenan los eigenvectores de acuerdo a la magnitud de los eigenvalores; se seleccionan los primeros d eigenvectores para representar las n variables, con $d < n$. Entonces los vectores $XP_1, XP_2, \dots, XP_d$ son llamados los "componentes principales".

Se inspeccionó visualmente la complejidad de los datos aplicando una modalidad de PCA para datos categóricos llamada MCA, a cada subconjunto de datos considerado. La Figura 7 muestra la gráfica de los dos componentes principales para el cromosoma 14. Las marcas azules (+) corresponden a vacas con alta producción de leche, mientras que las marcas rojas (*) corresponden a vacas con baja producción. Se puede observar que los datos están

mezclados, lo que significa que no existe una distinción visual entre los grupos representados. Así que deben buscarse métodos alternativos para procesar los datos que permitan discernir entre los grupos de vacas.

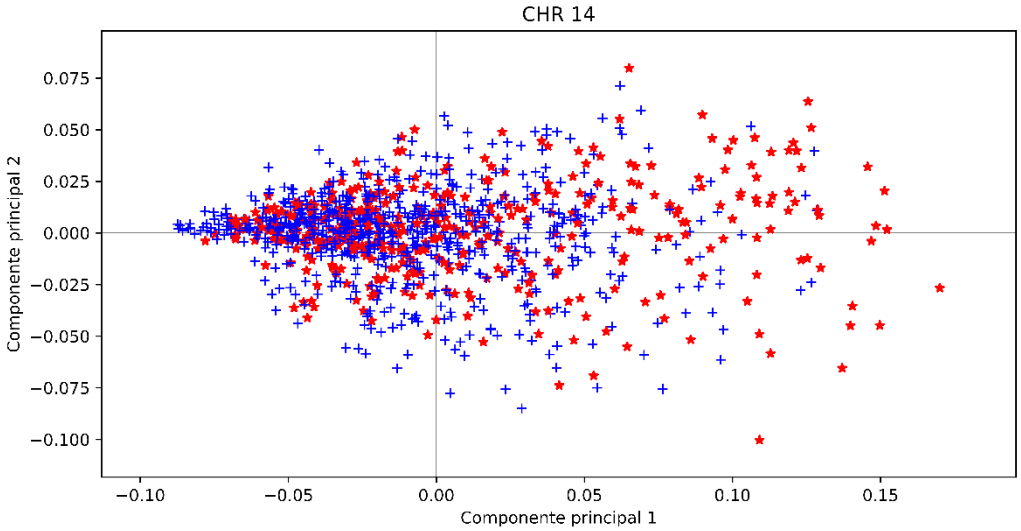


Figura 7. Gráfica MCA con datos del cromosoma 14. Las marcas azules representan a los ejemplares con alta producción lechera, y las marcas rojas, a los bajos productores.

2.1.5 Procesamiento de los datos

A continuación, se describe el procesamiento realizado a cada conjunto de datos una vez implementado el procedimiento de QC descrito en la sección 2.1.2

2.1.5.1 Preprocesamiento tipo *one-hot encode*

La entrada al clasificador es una codificación que representa si se trata de un homocigoto alelo menor, heterocigoto u homocigoto alelo mayor, con valores 0, 1 y 2, respectivamente; es decir, son valores categóricos en lugar de valores continuos. No es deseable confundir al modelo al hacer parecer que existe cierto orden o jerarquía entre los valores, cuando no la hay. Para evitarlo, se utiliza la codificación de tipo *one-hot encode* en cada SNP. El

procedimiento para realizar esta codificación es el siguiente: Suponer que las columnas de la matriz representan a los SNP y se usan *n* valores distintos, cada uno de ellos se codifica para utilizar *n* columnas en lugar de una; así pues, se tendrán tantas columnas como valores diferentes existan. En este caso, al ser 3 valores diferentes, se utilizarán una terna. El valor de entrada 0, se convierte en (1,0,0); el valor 1, se convierte en (0,1,0) y el valor 2, se convierte en (0,0,1). De esta manera, la distancia entre cualquier pareja de valores en la misma así que se quita una posible fuente de confusión. La librería *scikit-learn* tiene una serie de funciones que facilitan la transformación en el paquete *sklearn.preprocessing*.

2.1.5.2 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).

La función de Python utilizada para separar al conjunto de datos en entrenamiento/prueba es train_test_split() y se utilizó como sigue

```
X_train, X_test, y_train, y_test = train_test_split(genotipo, fenotipo, test_size
= 0.1, random_state = semilla)
```

La función regresa el conjunto de genotipos dividido en el conjunto de entrenamiento (X_train) y el de pruebas (X_test) y el conjunto de fenotipos dividido como y_train, y_test. Se reservó el 10% de los datos para la etapa de pruebas, dejando el 90% para la etapa de entrenamiento.

El siguiente paso es seleccionar el clasificador a utilizar. En este trabajo los clasificadores utilizados son el árbol de decisión (la función en Python se llama DecisionTreeClassifier) y la red neuronal o *multi-layer perceptrón* (la función en Python se llama MLPClassifier). Cada uno de estos clasificadores utiliza un conjunto de parámetros y hay que buscar cuales son los mejores para el conjunto de datos donde se aplicará el método.

Cuando se evalúa un conjunto de parámetros en el clasificador, aún hay riesgo de sobreajuste debido a que el entrenamiento puede llevarse a cabo hasta que se aprenden todos los valores utilizados lo que no permitirá una buena generalización (es decir, no funcionará bien con el conjunto de pruebas). Para evitarlo, se divide el conjunto de entrenamiento en dos: una parte para el entrenamiento y otra, para el conjunto de validación. El entrenamiento se lleva a cabo con el conjunto de entrenamiento y se evalúa con el conjunto de validación, cuando se obtienen buenos resultados, se aplica el conjunto de pruebas para una evaluación final del modelo.

Así que el conjunto de datos original se divide en tres partes, esto reduce el número de muestras usadas en el conjunto de entrenamiento y el resultado puede variar dependiendo de la selección aleatoria particular de los conjuntos de entrenamiento/validación. Para solucionar este problema, se usa un procedimiento llamado validación cruzada (*cross-validation*, CV). Una parte del conjunto de datos se sigue reservando para la evaluación final, aunque el conjunto de validación no necesita ser especificado. En la técnica básica, llamada *k-fold* CV, el conjunto de entrenamiento se divide en *k* partes más pequeñas. El

entrenamiento es realizado usado $k - 1$ de las partes, y el modelo resultante es validado con la parte restante del conjunto de datos. Se repite el procedimiento variando la parte utilizada para el conjunto de validación cada vez, y el resultado que se reporta es el promedio del desempeño calculado en cada iteración del ciclo. Este procedimiento puede considerarse computacionalmente costoso, pero su uso está muy recomendado, sobre todo, cuando se tiene un número pequeño de muestras [44].

Cada uno de los clasificadores considerados tiene una serie de parámetros, llamados hiper-parámetros, que pueden cambiar el desempeño del mismo. Hay que realizar una búsqueda de la mejor combinación de hiper-parámetros que arroje los mejores resultados al utilizar CV. Esto se conoce como ajuste de los hiper-hiperpámetros del estimador. Python, proporciona la clase GridSearchCV para considerar exhaustivamente todas las combinaciones de los hiper-parámetros especificados.

Cuando se llama a la función, se especifica cuál es el clasificador a utilizar y cuál es el conjunto de parámetros a considerar en la búsqueda, así como el número de *folds* a considerar en el CV.

En la Figura 8 se presenta un bosquejo del procedimiento descrito.

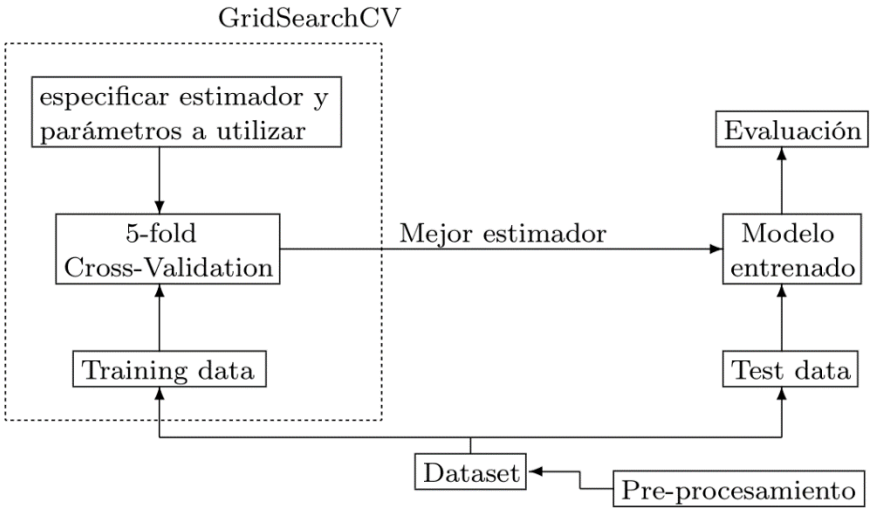


Figura 8. Bosquejo del entrenamiento usando GridSearchCV.

La clase contiene los métodos *fit* con el que se entrena el modelo buscando los mejores hiper-parámetros, y *score*, que evalúa la precisión de la clasificación realizada. También incluye el atributo *best_estimator_* que regresa el estimador que obtuvo la mayor puntuación.

Otra de las estrategias utilizadas para evitar el sobreajuste es la regularización, empleada en las redes neuronales. La regularización consiste en añadir una penalización a la función de costos.

Suponer que la función de costo, J, está basada en el error cuadrático medio (MSE), así que la función tiene la forma

$$J = \frac{1}{M}\sum_{i=1}^{M}(real_i - estimado_i)^2$$

donde M, es el número de muestras, $real_i$ y $estimado_i$ representan el valor real y el valor estimado por el modelo para la muestra *i*, respectivamente. Al considerar regularización debe agregarse un término que penaliza la complejidad del modelo

$$J = \frac{1}{M}\sum_{i=1}^{M}(real_i - estimado_i)^2 + \alpha C$$

donde C es el modelo para medir la complejidad del modelo. Python utiliza el modelo de regularización Ridge, también llamado L2, donde la complejidad se mide como la media del cuadrado de los coeficientes. La función de costo queda como

$$J = \frac{1}{M}\sum_{i=1}^{M}(real_i - estimado_i)^2 + \alpha \frac{1}{2N}\sum_{i=1}^{N} {w_i}^2$$

El factor α es llamado el coeficiente de regularización.

La penalización de tipo L2, se utiliza cuando la mayoría de los atributos son relevantes y existe cierta correlación entre ellos. Cuando se utiliza regularización se minimiza la complejidad del modelo y la función de costo. Estos modelos más simples tienden a generalizar mejor.

A continuación, se describe la implementación realizada de la búsqueda de los mejores parámetros para el clasificador.

2.1.5.3 DecisionTreeClassifier

Los parámetros con los cuales se va a evaluar el clasificador son:

Parámetro	**Valores**
criterion	'gini', 'entropy'
splitter	'best', 'random'
random_state	1, 2, 3, …, 1000

La clase DecisionTreeClassifier tiene el atributo *tree_* con el cual se puede conocer el número de nodos que conforman el árbol resultante.

Como puede observarse se usaron diferentes valores para los criterios de paro (Gini [32] y entropy [42]) y 2 diferentes métodos de separación. También, se usaron diferentes valores en la variable aleatoria (*random_state*), en el rango de 1 a 1000. Se utilizaron un total de 4 000 combinaciones diferentes de parámetros, buscando aquella combinación que arrojara mejores resultados.

2.1.5.4 MLPClassifier

En el caso de usar una red neuronal como clasificador, se usaron los parámetros siguientes:

Parámetro	**Valores**
hidden_layer_sizes	1, 2, …,25
activation	identity, logistic, tanh, relu
solver	lbfgs, sgd, adam
alpha	1.0, 0.1, 0.01, 0.001
max_iter	3500, 4000, 4500
random_state	1, 2, …, 15

Se utilizó una red neuronal con tres capas: una capa de entrada con tantas neuronas como SNPs del *dataset* usado, una capa oculta con un número variable de neuronas y una capa de salida con una neurona. La matriz de pesos y la de sesgos se inicializaron aleatoriamente. Se usan cuatro opciones para la función de activación f_{act} (logistic [34] y ReLu [33], entre otras). También se consideró un número de valores que va de 1 a 25 para la

capa oculta. Otros parámetros utilizados fueron el método de optimización aplicado, *solver* (3 opciones), la tasa de aprendizaje α (4 valores de 0.1 a 1), el número máximo de épocas tiene 3 valores posibles, y el número aleatorio, semilla, toma valores de 1 a 15. En total se estudiaron 54 000 combinaciones de hiper-parámetros para seleccionar aquella que arroje los mejores resultados. Esto explica por qué el entrenamiento requiere de un tiempo considerable.

2.2 Resultados

El fenotipo utilizado es el EBV de la producción promedio de leche a 305 días ('*milk_le_ave_305*'), con un rango de valores que va desde -8.576 hasta 16.838. Los EBV se calcularon utilizando un modelo de regresión aleatoria con muestreo diario, con efectos fijos del día de muestreo del rebaño y coeficientes de regresión fijos [11]. Para propósitos de clasificación, se transforman los valores del fenotipo a valores categóricos, asignándole la clase “no-lechera”, si el valor del fenotipo fue menor o igual a 0, y la clase “lechera”, si el fenotipo fue mayor que 0. Como resultado, se tienen 670 vacas catalogadas como “no-lechera” y 422, como “lechera”.

Se realizó una reducción de atributos usando el procedimiento descrito en la sección 2.1.2 y después se aplica la codificación descrita en 2.1.5.1 con lo cual se triplicó el tamaño del conjunto obtenido en la reducción de atributos. Para evaluar el desempeño de cada algoritmo, se generaron diferentes subconjuntos conteniendo SNPs de grupos de cromosomas que contienen el número más alto de QTLs relacionados a la producción lechera. En la Tabla 4 se describen los conjuntos de datos generados a partir del conjunto de

datos original. Considerar el número que le corresponde al dataset en la lista para futuras referencias.

Tabla 4. Conjuntos de datos utilizados.

Conjunto	Cromosomas incluidos
1	1, 2, …, 29
2	14
3	6, 14
4	5, 6, 14
5	5, 6, 14, 20
6	1, 5, 6, 14, 20
7	1, 14

En la Tabla 5 se muestra el número de SNPs en cada subconjunto antes y después del filtrado de control de calidad (no se incluye el tamaño después de la codificación *one-hot encode*). Obsérvese que se obtiene un conjunto de ∼30% del tamaño original.

Tabla 5. Número de SNPs contenidos en cada conjunto, antes y después del filtrado de QC.

Conjunto de datos	Número de SNPs antes del filtrado de QC	Número de SNPs después del filtrado de QC
1	164 312	52 475
2	4 004	1 230
3	9 316	2 736
4	17 049	4 866
5	20 444	5 804
6	27 782	7 901
7	11 342	3 327

2.2.1 Árboles de decisión

Se seleccionaron los nueve subconjuntos de SNPs del conjunto de datos según se explicó anteriormente. En cada caso, se aplicó la técnica de CV descrita en la sección 2.1.5.2 y se aplicó el clasificador según se describe en el apartado 2.1.5.3 para entrenar al clasificador.

La Tabla 6 presenta los resultados, donde la columna 1 indica el conjunto utilizado, la columna 2 muestra la precisión de la clasificación, la columna 3 muestra en número de nodos en el árbol resultante y la columna 4 contiene el número de muestras que se clasificaron incorrectamente (de un total de 110). Como puede observarse, los mejores resultados de la clasificación se obtuvieron para el conjunto_7 (incluye los cromosomas 1 y 14), con una precisión de 94.5% y 97 nodos. La precisión más baja fue de 90.9% se obtuvo con dos conjuntos, según puede verse en la tabla mencionada.

Tabla 6. Precisión de la clasificación usando árboles de decisión.

Conjunto	**Clasificación usando árboles de decisión**		
	Precisión del conjunto de pruebas (%)	Núm. de nodos en el árbol resultante	Número de muestras clasificadas incorrectamente
1	93.6	67	7
2	91.8	117	7
3	91.8	107	9
4	90.9	93	10
5	90.9	95	10
6	93.6	93	7
7	94.5	97	6

En la Tabla 7 se especifica cuáles fueron los parámetros utilizados[4] para obtener la puntuación más alta, en este caso para el conjunto_7.

Tabla 7. Mejores parámetros para el árbol de decisión.

Parámetro	**Valores**
criterion	entropy
splitter	random
random_state	232

[4] Para consultar el listado completo de los parámetros utilizados, ir al apéndice A.

La matriz de confusión de la Tabla 8 indica que se etiquetaron mal 6 muestras de un total de 110, cuando se usó el conjunto_7 *(*chr1 y chr14).

Tabla 8. Matriz de confusión arrojada por el clasificador DT para el conjunto_7.

		Clase esperada	
		No-lechera	Lechera
Clase predicha	No-lechera	62	4
	Lechera	2	42

En cuanto al tiempo de procesamiento, con el conjunto de datos que incluye todos los cromosomas, se requirieron, aproximadamente, 32 horas de procesamiento. Cuando se utilizó el conjunto_2 (cromosoma 14), el tiempo de ejecución fue de, aproximadamente, 1.4 horas

Obsérvese que al cambiar el valor de la variable *random_state* la precisión de la clasificación cambia, por lo que se hicieron varias llamadas al programa cambiando únicamente este argumento y en la Figura 9 se puede ver los distintos valores para la precisión que se obtienen considerando el conjunto_7.

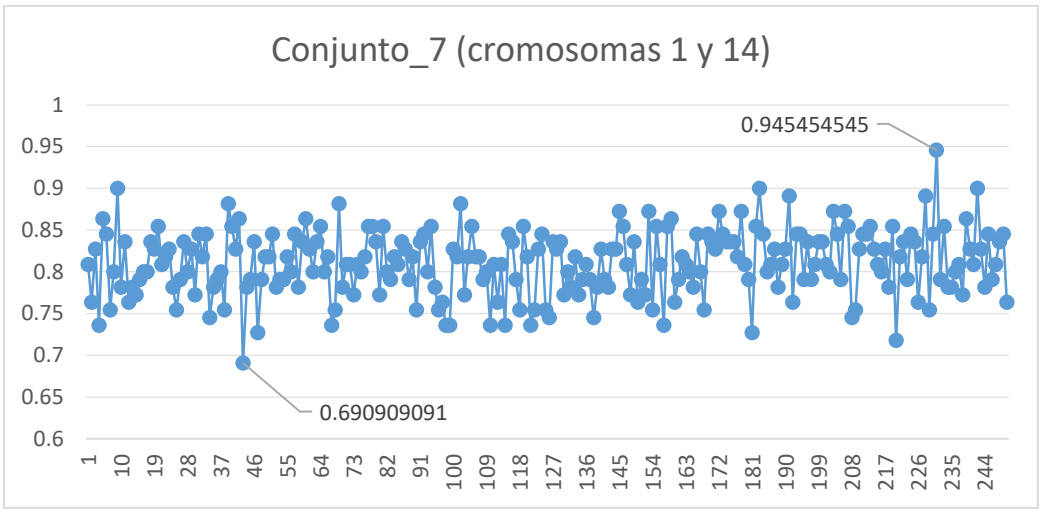


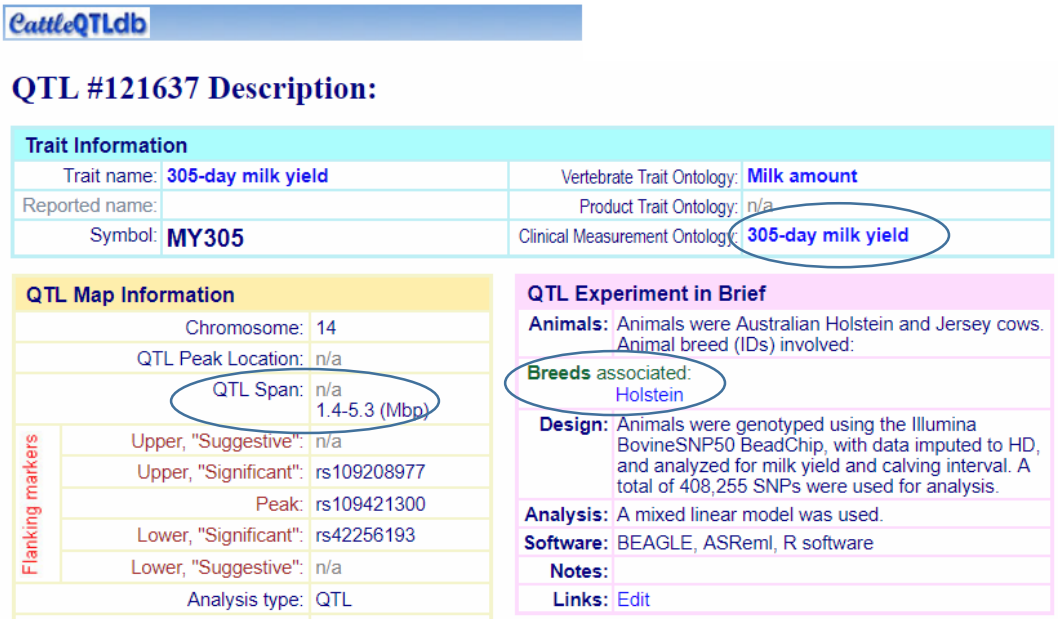
Figura 9. Variación del resultado al cambiar random_state *en el árbol de decisión.*

En los árboles de decisión, al realizar la toma de decisiones por el análisis jerárquico de las variables (tomadas de forma individual), es relativamente sencillo calcular la importancia de cada una de las variables (SNP) en el proceso de toma de decisiones. Con cada uno de los conjuntos de datos, el algoritmo selecciona el SNP con posición 1 455 997, 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 1 455 997 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[5]. El SNP en cuestión está dentro del QTL #121 637

[5] Cattle QTL database, https://www.animalgenome.org/cgi-bin/QTLdb/BT/index

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 10.



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
Flanking markers — Upper, "Significant":	rs109208977
Flanking markers — Peak:	rs109421300
Flanking markers — Lower, "Significant":	rs42256193
Flanking markers — 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 10. El SNP más influyente está en un QTL relacionado con la producción lechera a 305 días en ganado Holstein.

Más aún, la anotación del SNP fue verificada en los recursos NCBI[6], buscando los genes que están dentro del rango de los 110 Kbp a ambos lados del SNP, para investigar la asociación con la funcionalidad biológica. Se encontraron nueve genes, uno de los cuales no está caracterizados El gen no caracterizado es LOC104973964, mientras que los caracterizados son genes codificadores de proteína: LOC112441461, LOC787628L, Y6H, GPIHBP1, LY6E, LY6L, GML y CYP11B1, según puede verse en la Figura 11. En [45] se relaciona al gen GPIHBP1 con el atributo de grasa en leche.

[6] NCBI, https://www.ncbi.nlm.nih.gov/nuccore/NC_037341.1?report=graph

Figura 11. Listado de genes cercanos a la posición 1 455 997 en un rango de 110 000 bp.

Para verificar que el cromosoma 14 es el que tiene más influencia en el resultado obtenido, se hicieron pruebas considerando los conjuntos mencionados sin incluir al cromosoma 14. En la Tabla 9 se muestran los resultados obtenidos.

Tabla 9. Resultados de la clasificación usando árboles de decisión, sin considerar al cromosoma 14.

Conjunto de datos con los cromosomas	**Clasificación usando árboles de decisión**	
	Precisión del conjunto de pruebas (%)	Núm. de nodos en el árbol resultante
Todos, excepto el cromosoma 14	73.6	149
6	70.9	231
5, 6	72.7	229
5, 6, 20	71.8	223
1, 5, 6, 20	72.7	193
1	70	221

En todos los casos, los resultados están muy por debajo de los obtenidos en la Tabla 6, lo que indica que la información del cromosoma 14 es realmente de importancia para el fenotipo considerado.

2.2.2 Redes neuronales

Al aplicar el algoritmo de redes neuronales artificiales para clasificación se utilizaron los mismos grupos de SNPs utilizados con el DT y el mismo procedimiento buscando la mejor combinación de parámetros. La topología de la red neuronal implementada fue un perceptrón multicapa, con una capa de entrada, una capa oculta y una capa de salida con una neurona. El número de neuronas de la capa oculta se fue incrementando gradualmente de 1 a 25 para encontrar el tamaño de la capa oculta que produjera la mejor clasificación para el conjunto

de entrenamiento. La Tabla 10 presenta los resultados, la columna 2 muestra la exactitud de la clasificación mientras que la columna 3 muestra el número de neuronas en la capa escondida, para los siete subconjuntos de SNPs. La precisión alcanzada fue 87.3% obtenida con una capa oculta de 1 neurona, con el cromosoma 14. La peor precisión de la clasificación fue de 79%, obtenida con una capa oculta de 8 neuronas, en el conjunto$_1$ (los 29 cromosomas).

Tabla 10. Precisión de la clasificación usando NN.

Conjunto de datos	**Clasificación usando Redes Neuronales**	
	Precisión (%)	Número de neuronas en la capa oculta
1	79	8
2	87.3	1
3	83.6	7
4	80.9	23
5	81.8	18
6	82.7	22
7	80	1

Para el conjunto_2, la mejor estrategia encontrada fue con los parámetros especificados en la Tabla 11[7].

[7] Para consultar el listado completo de los parámetros utilizados, ir al apéndice B.

Tabla 11. Mejores parámetros para la red neuronal.

Parámetro	**Valores**
hidden_layer_sizes	1
activation	tanh
solver	lbfgs
alpha	0.1
max_iter	3500
random_state	2

En cuanto al tiempo de procesamiento, considerando al conjunto_2, el tiempo requerido fue de aproximadamente 384 horas, mientras que al considerar con el conjunto_1 fue de 10 000 horas, aproximadamente. Hay que señalar que se realizaron las corridas del programa por separado para cada semilla, cada uno de ellos requirió de un mes de procesamiento; sin embargo, se utilizó con clúster con diversos procesadores y se asignó un trabajo a cada uno de ellos, para su ejecución simultánea. El tiempo total de procesamiento fue el equivalente a 15 meses (pues se usaron 15 semillas).

Los métodos de ML permiten explorar diferentes soluciones (o aproximaciones a la mejor solución). Cuando se ejecutan, es necesario inicializar un conjunto de variables antes de iniciar la ejecución, tales como el valor de la semilla usada para inicializar las variables aleatorias, el número de capas escondidas y el número de neuronas en cada una, entre otras. La Figura 12 muestra el espacio de soluciones que se encontraron cuando se barrieron tanto la semilla de inicialización, de 1 a 15 y el número de neuronas en la capa oculta (*nnxl*) de 1 a 25, cuando se aplica el algoritmo de redes neuronales al conjunto_2 (cromosoma 14).

Obsérvese la variedad de los resultados obtenidos. La mejor predicción corresponde al valor más alto de la gráfica, resaltado mediante una marca roja.

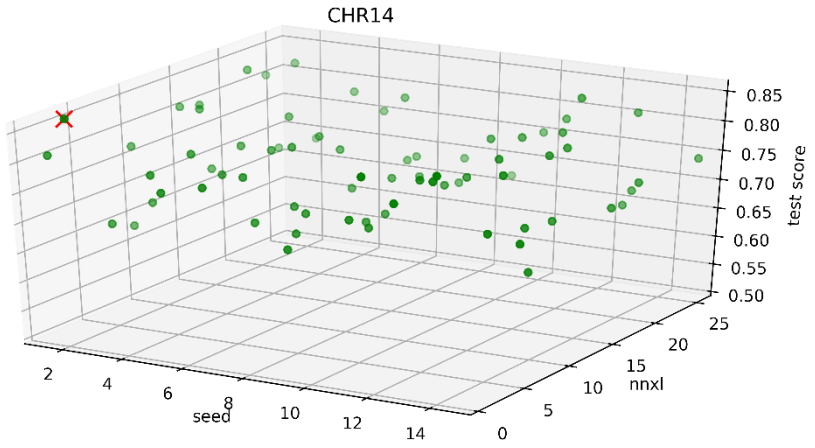


Figura 12. Espacio de soluciones obtenidas (score) con redes neuronales. Los ejes representan la semilla, el número de neuronas por capa y la exactitud de la predicción obtenida por el algoritmo.

3 Discusión

Los resultados obtenidos en la sección anterior muestran la capacidad de predicción temprana del fenotipo lechero a partir de SNPs de genoma completo utilizando dos tipos diferentes de algoritmos de clasificación. En este documento, se probaron dos algoritmos de ML para clasificación con el propósito de predecir cuál vaquilla será una buena productora de leche y cuál no lo será. Los datos consisten de 52 475 SNPs después del filtrado de control de calidad, utilizando los 29 cromosomas autosomales, de 1092 vacas de raza *Holstein*. Para entrenar y probar la precisión de los métodos, se usó como fenotipo el EBV de producción lechera promedio a 305 días, incluido con el conjunto de datos. Los algoritmos probados fueron los modelos de árboles de decisión y redes neuronales para clasificación. La implementación del entrenamiento y predicción se desarrolló en el lenguaje Python usando la librería scikit.

Los resultados obtenidos están concentrados en la Tabla 12Tabla 12.

Tabla 12. Concentrado de los resultados obtenidos con los métodos utilizados.

Conjunto de datos	**Clasificación con árboles de decisión**		**Clasificación con redes neuronales**	
	Precisión (%)	Núm. de nodos en el árbol	Precisión (%)	# neuronas en la capa oculta
1	93.6	67	79	8
2	91.8	117	87.3	1
3	91.8	107	83.6	7
4	90.9	93	80.9	23
5	90.9	95	81.8	18
6	93.6	93	82.7	22
7	94.5	97	80	1

Después de aplicar los algoritmos a los diferentes subconjuntos de SNPs generados a partir de los cromosomas con un alto número de QTL asociados con una alta producción lechera, la precisión de la clasificación va de 90.9% a 94.5% usando árboles de decisión, y de 79% a 87.3% usando redes neuronales.

Una ventaja del modelo de árboles de decisión es que identifica el SNP que tiene una mayor influencia al hacer la predicción. Este SNP está localizado en la posición 1 455 997 del cromosoma 14 y cae en un QTL relacionado con la producción promedio de leche a 305 días. Además, al buscar los genes cercanos al SNP, se encontraron genes caracterizados que no tienen una asociación documentada con la producción lechera. Sin embargo, se

encontraron dos genes no caracterizados que podrían estar asociados. Para verificar estas asociaciones es necesario realizar más análisis experimental, lo cual está fuera del alcance de este estudio.

El hecho de que con los árboles de decisión se obtenga una mayor exactitud en la predicción del fenotipo lechero que con las redes neuronales, los hace más apropiados para implementar una herramienta de clasificación específica para este problema.

Si bien, por la naturaleza inductiva de los métodos de aprendizaje supervisado, no hay garantía de que alcanzará los mismos resultados con un conjunto de datos diferente, en el estudio existen evidencias de que la distribución general de los datos del fenotipo lechero a partir de los SNPs, es compatible con el modelo de árboles de decisión. Es importante resaltar que, de acuerdo a la teoría estadística del aprendizaje, los resultados son más probablemente aproximadamente correctos conforme aumenta el número de evidencias, en este caso, el número de muestras en el conjunto de entrenamiento. Si bien es necesaria una prueba más exhaustiva, en general es recomendable inicial con un modelo de árboles de decisión en el desarrollo de un producto de predicción temprana.

Una característica importante del modelo de árboles de decisión es que no solo arroja el valor de la predicción, sino que ayuda a encontrar los SNPs más influyentes para la clasificación, los cuales pueden ser usados como una indicación de una posible asociación de los SNPs con rasgos de interés. Finalmente, los algoritmos ML son capaces de manejar información de SNPs para realizar análisis, lo que los hace adecuados para la implementación de herramientas de predicción para rasgos económicos en la industria ganadera.

Los resultados obtenidos son alentadores, sobre todo al considerar que al utilizar el conjunto_7, incluso con el conjunto_2, arroja casi los mismos resultados que utilizando el conjunto de datos completo cuando se utilizan árboles de decisión, mientras que cuando se

utilizan redes neuronales la exactitud de la predicción es superior cuando se considera solo el cromosoma 14. Puede concluirse entonces que para los datos que se manejan, el uso del *dataset* conformado por el cromosoma 14 es suficiente para obtener una clasificación con una precisión apenas por debajo a la obtenida con el conjunto completo. Esto puede indicar grandes ahorros en la genotipificación de las muestras y en el tiempo de procesamiento. En cuanto al tiempo requerido para realizar el entrenamiento, hay una gran diferencia en los dos métodos, siendo el modelo de árboles de decisión el que menor tiempo requiere.

El procedimiento utilizado puede apoyar a la selección de los ejemplares que se destinarán a la producción lechera, y con ello mejorar la producción. Si esto se replica en cada rancho, aun cuando el tamaño del hato no sea muy grande debido a las restricciones climáticas de la zona, la producción lechera puede incrementarse en el estado. Cabe señalar que con el paso del tiempo estos resultados pueden mejorar al incluirse un mayor número de muestras conforme se va obteniendo el genotipo de nuevos ejemplares.

4 Conclusiones y trabajo futuro

4.1 Conclusiones

En este trabajo se implementaron, ajustaron y analizaron dos modelos basados en aprendizaje supervisado para la tipificación temprana del fenotipo lechero de vaquillas, a partir del uso de marcadores genéticos. El aporte de esta tesis tiene como fin contribuir al desarrollo de métodos propios de selección de animales genéticamente superiores en menor tiempo, para hacer más eficiente el proceso de reproducción asistida y con ello aportar al desarrollo del sector lechero del país y la región.

Para el desarrollo del trabajo, se utilizó la información de 1 092 vacas lecheras de raza Holstein. Una vez tratados los datos se retuvieron 52 475 SNPs después del filtrado de control de calidad, utilizando los 29 cromosomas autosomales.

Usando esta información, se pudo demostrar en el trabajo que tanto la técnica de redes neuronales para clasificación binaria, como los árboles de decisión son eficientes en la identificación temprana de vacas con alta producción lechera.

El modelo de árboles de decisión utiliza la información de los marcadores genéticos en forma jerárquica y selectiva, mientras que las redes neuronales utilizan el conjunto de marcadores en forma global. Estas diferencias son importantes, ya que pueden dar luz a la relación entre el fenotipo lechero y los diferentes SNPs.

Entre los resultados importantes se encontró que la exactitud en la predicción con redes neuronales disminuía si se utilizaba el conjunto completo de SNPs, al compararlo con el uso de un subconjunto de indicadores del cromosoma 14. Este resultado puede ser evidencia que existen algunos SNP que no se encuentran correlacionados con el fenotipo lechero.

Los árboles de decisión tuvieron un desempeño mejor que las redes neuronales para este problema, lo que podría indicar que existen marcadores con más influencia que otros en la predicción de vacas lecheras. De acuerdo a los resultados se puede concluir que el cromosoma CH14 y el CHR1 tienen una influencia mayor en la determinación del fenotipo lechero que otros marcadores. Esto puede conducir a ahorro económicos pues solo se requiere genotipificar a los cromosomas 14 y 1 para obtener predicciones con una exactitud mayor al 94%.

En la Tabla 3 se menciona que el cromosoma 14 es el que tiene más QTL asociados con la producción lechera, en este trabajo se ratificó esa relación pues se ha demostrado que los SNPs de este cromosoma son muy influyentes en la clasificación.

Más aún, el cromosoma 14 en conjunto con el cromosoma 1, mejoraron ligeramente la precisión de la clasificación alcanzada, usando DT, utilizando todos los cromosomas. Este resultado indica que vale la pena estudiar más a fondo al cromosoma 1 y su posible asociación con el fenotipo de producción lechera.

Si bien, existen estudios sobre la relación del cromosoma 14 con la producción lechera, no se había estudiado si éste pudiera ser suficiente para la detección temprana del fenotipo lechero.

Esta técnica demuestra que es posible seleccionar a una vaca para destinarla a la producción lechera desde el momento de su nacimiento (o incluso antes), sin necesidad de conocer el desempeño de sus padres o a esperar a que exprese su propio fenotipo. Debido a ello, esta técnica puede emplearse para ahorrarle al ganadero gastos innecesarios.

La metodología utilizada puede llevarse a la utilización de otros fenotipos que también incidan en la producción lechera, como el tamaño de la ubre o la predisposición a contraer enfermedades tales como mastitis.

4.2 Trabajo futuro

En base a los resultados obtenidos puede considerarse la creación de un chip que permita la detección temprana del fenotipo lechero basado solo en la información genómica del cromosoma 14 y del cromosoma 1. Puede usarse este chip para medir las características del ganado de le región, buscando depurar el modelo hasta obtener un producto capaz de identificar el fenotipo de interés en el ganado regional.

Una alternativa sería basarnos en que los árboles de decisión pueden detectar los SNPs más influyentes en la clasificación, y seleccionar a los mejores 100 SNPs, por ejemplo, para la creación del chip mencionado y empezar a medir su eficiencia.

Es posible que lleve un tiempo adecuar el modelo para que identifique el fenotipo de interés en el ganado regional, que debido al medioambiente puede tener características genómicas diferentes a la muestra estudiada en este trabajo, sin embargo, una vez adecuado el modelo, podría apoyar a la conformación de un hato con mejor rendimiento en la producción lechera, logrando con esto incrementar la producción de Baja California.

Incluso puede pensarse en utilizar el modelo para predecir otros fenotipos relacionados con la producción lechera o con la salud animal, como su predisposición a contraer mastitis.

Referencias

[1] Cámara Nacional de Industriales de la Leche, “Estadísticas del sector lácteo,” México, 2019.

[2] U. Singh *et al.*, “Molecular markers and their applications in cattle genetic research: A review,” *Biomarkers Genomic Med.*, vol. 6, no. 2, pp. 49–58, 2014.

[3] M. E. Goddard and B. J. Hayes, “Genomic selection-OLD,” *J. Anim. Breed. Genet.*, vol. 124, pp. 323–330, 2007.

[4] R. Oliva, J. Oriola, F. Ballesta, J. Clària, and L. Mengual, *Genética médica*. Barcelona: Elsevier España, 2011.

[5] T. H. Emigh, “A Comparison of Tests for Hardy-Weinberg Equilibrium,” *Int. Biometric Soc.*, vol. 36, no. 4, pp. 627–642, 1980.

[6] M. G. Smaragdov, “Genomic selection of milk cattle. The practical application over five years,” *Russ. J. Genet.*, vol. 49, no. 11, pp. 1089–1097, 2013.

[7] P. M. VanRaden *et al.*, “Invited review: reliability of genomic predictions for North American Holstein bulls.,” *J. Dairy Sci.*, vol. 92, no. 1, pp. 16–24, 2009.

[8] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, “A primer on deep learning in genomics,” *Nat. Genet.*, vol. 51, no. 1, pp. 12–18, 2019.

[9] T. Meuwissen, B. Hayes, and M. Goddard, “Prediction of total genetic value using genome-wide dense marker maps,” *Genetics*, vol. 157, no. 4, pp. 1819–1829, 2001.

[10] 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.

[11] 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, p. e0192695, 2018.

[12] 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.

[13] B. K. Meredith, J. F. Kearney, E. K. Finlay, D. G. Bradley, and D. P. Berry, “Genome-wide Associations for Milk Production and Somatic Cell score in Holstein-Friesian Cattle in Ireland,” *BMC Genet.*, vol. 13, no. 21, 2012.

[14] K. L. Verbyla, B. J. Hayes, P. J. Bowman, and M. E. Goddard, “Accuracy of genomic selection using stochastic search variable selection in Australian holstein friesian dairy cattle,” *Genet. Res. (Camb).*, vol. 91, no. 5, pp. 307–311, 2009.

[15] T. Meuwissen, B. Hayes, and M. Goddard, “Accelerating Improvement of Livestock with Genomic Selection,” *Annu. Rev. Anim. Biosci.*, vol. 1, no. 1, pp. 221–237, 2013.

[16] T. B. H. Consortium, “Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds,” *Science (80-.).*, vol. 324, no. 5926, pp. 528–532, 2009.

[17] R. Villa-Angulo, L. K. Matukumalli, C. A. Gill, J. Choi, C. P. Van Tassell, and J. J. Grefenstette, “High-resolution haplotype block structure in the cattle genome,” *BMC Genet.*, vol. 10, no. 1, p. 19, 2009.

[18] L. K. Matukumalli, J. J. Grefenstette, D. L. Hyten, I. Y. Choi, P. B. Cregan, and C. P. Van Tassell, “SNP-PHAGE- High throughput SNP discovery pipeline,” *BMC Bioinformatics*, vol. 7, no. 1, p. 468, 2006.

[19] L. K. Matukumalli *et al.*, “Development and Characterization of a High Density SNP

Genotyping Assay for Cattle," *PLoS One*, vol. 4, no. 4, p. e5350, 2009.

[20] P. M. VanRaden, "Efficient Methods to Compute Genomic Predictions," *J. Dairy Sci.*, vol. 91, no. 11, pp. 4414–4423, 2008.

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

[22] N. S. Yudin, K. I. Lukyanov, M. I. Voevoda, and N. A. Kolchanov, "Application of reproductive technologies to improve dairy cattle genomic selection," *Russ. J. Genet. Appl. Res.*, vol. 6, no. 3, pp. 321–329, 2016.

[23] D. Boichard, V. Ducrocq, P. Croiseau, and S. Fritz, "Genomic selection in domestic animals: Principles, applications and perspectives," *Comptes Rendus - Biol.*, vol. 339, no. 7–8, pp. 274–277, 2016.

[24] J. E. Pryce and H. D. Daetwyler, "Designing dairy cattle breeding schemes under genomic selection: A review of international research," *Anim. Prod. Sci.*, vol. 52, no. 3, pp. 107–114, 2012.

[25] B. J. Hayes, P. J. Bowman, A. J. Chamberlain, and M. E. Goddard, "Invited review: Genomic selection in dairy cattle: Progress and challenges," *J. Dairy Sci.*, vol. 92, no. 2, pp. 433–443, 2009.

[26] 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.

[27] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science (80-.).*, vol. 349, no. 6245, pp. 255–260, 2015.

[28] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, “Supervised Machine Learning : A Review of Classification Techniques,” in *Emerging artificial intelligence applications in computer engineering*, vol. 160, IOS Press, 2007, pp. 3–24.

[29] Z. Ghahramani, “Unsupervised Learning,” in *Advanced lectures on machine learning*, 2004, pp. 72–112.

[30] M. W. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics,” *Nat. Rev. Genet.*, vol. 16, no. 6, p. 321, 2015.

[31] D. R. Schrider and A. D. Kern, “Supervised Machine Learning for Population Genetics: A New Paradigm,” *Trends Genet.*, vol. 34, no. 4, pp. 301–312, 2018.

[32] L. Rokach and O. Z. Maimon, *Data mining with decision trees, theory and applications*. World Scientific Publishing Company, 2014.

[33] I. Goodfellow, Y. Bengio, and A. Courville, “Deep Feedforward Networks,” in *Deep Learning*, no. 1, MIT Press, 2016.

[34] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” *Calif. Univ San Diego La Jolla Inst Cogn. Sci.*, no. ICS-8506, 1985.

[35] M. Xu, P. Watanachaturaporn, P. K. Varshney, and M. K. Arora, “Decision tree regression for soft classification of remote sensing data,” *Remote Sens. Environ.*, vol. 97, no. 3, pp. 322–336, 2005.

[36] K. Kim, “A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree,” *Pattern Recognit.*, vol. 60, pp. 157–163, 2016.

[37] L. Jjyafjl and R. L. Rjvest, “Constructing optimal binary decision tree is NP-complete,”

Inf. Process. Lett., no. 1, pp. 1–3, 1976.

[38] G. Hinton, "Deep learning—a technology with the potential to transform health care," *Jama*, vol. 320, no. 11, pp. 1101–1102, 2018.

[39] S. Sugiono, R. Soenoko, and L. Riawati, "Investigating the impact of physiological aspect on cow milk production using artificial intelligence," *Int. Rev. Mech. Eng.*, vol. 11, no. 1, pp. 30–36, 2017.

[40] S. Okser, T. Pahikkala, A. Airola, T. Salakoski, S. Ripatti, and T. Aittokallio, "Regularized Machine Learning in the Genetic Prediction of Complex Traits," *PLoS Genet.*, vol. 10, no. 11, p. e1004754, 2014.

[41] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[42] J. Shlens, "Shlens2006_PCATutorial," *Measurement*, vol. version 2, pp. 1–13, 2005.

[43] M. Alwakeel and Z. Shaahban, "Face Recognition Based on Haar Wavelet Transform and Principal Component Analysis via Levenberg-Marquardt Backpropagation Neural Network," *Eur. J. Sci. Res.*, vol. 42, no. 1, pp. 25–31, 2010.

[44] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[45] J. Yang *et al.*, "Functional validation of GPIHBP1 and identification of a functional mutation in GPIHBP1 for milk fat traits in dairy cattle," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017.

Apéndices

A

Listado de los parámetros considerados para el entrenamiento con DecisionTreeClassifier.

Parámetro	Valores
criterion	gini, entropy
splitter	best, random
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0
max_features	None
random_state	1, 2, 3, …, 1000
max_leaf_nodes	None
min_impurity_decrease	None
min_impurity_split	0.
class_weight	None
presort	False

B

Listado de los parámetros considerados para el entrenamiento con MLPClassifier.

Parámetro	Valores
hidden_layer_sizes	1, …, 25
activation	identity, logistic, tanh, relu
solver	lbfgs, sgd, adam
alpha	1.0, 0.1, 0.01, 0.001
batch_size	auto
learning_rate	constant
learning_rate_init	0.001
power_t	0.5
max_iter	3500, 4000, 4500
shuffle	True
random_state	1, …, 15
tol	1e-4
verbose	False
warm_start	False
momentum	0.9
nesterovs_momentum	True
early_stopping	False
validation_fraction	0.1
beta_1	0.9

beta_2	0.999
epsilon	1e-8
n_iter_no_change	10