Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos No Supervisados

Osvaldo Spósitto[1], Gabriel Blanco[2], Marcelo Levi[3], Patricio Macías Corral[4], Lorena Matteo[5]

Universidad Nacional de La Matanza
Departamento de Ingeniería e Ingestigaciones Tecnológicas

[1] sposito@unlam.edu.ar, [2] g2blanco@unlam.edu.ar, [3] mlevi@unlam.edu.ar
[4] pmacias@unlam.edu.ar, [5] lmatteo@unlam.edu.ar

Resumen

Un programa de mejoramiento genético de una raza contribuye a la mejora en la productividad y el beneficio económico de las explotaciones. Para ello se necesita disponer de información objetiva y precisa que contribuya a la toma de decisiones. Este trabajo es parte de un proyecto de investigación que pretende brindar una herramienta complementaria para la selección animal usando técnicas de Minería de Datos. En particular, se presenta un estudio que pretende encontrar patrones o grupos de características que determinen el peso de los terneros al nacer a través de la agrupación de los casos a partir de los valores genéticos usados en un rodeo de cría de la raza Aberdeen Angus, empleando para tal fin algoritmos de Minería de Datos del tipo No Supervisados. Por otra parte, se busca aplicar algoritmos de clasificación, como método alternativo para reducir la dimensionalidad de la muestra seleccionando un subconjunto de atributos. El análisis realizado se basa en datos provenientes de 360 ejemplares de progenitores proporcionados por un establecimiento ganadero de la localidad de Chascomús, Provincia de Buenos Aires, recolectados durante el periodo 2017-2018. Se concluye que, a partir del modelado construido, este puede colaborar con los criadores, en la toma de decisiones en un programa de mejoramiento genético

Introducción

El propósito de un programa de mejoramiento genético de una raza de carne es conocer y promover los mejores animales basados en registros de comportamiento y evaluación de sus progenitores [1]. Los productores ganaderos se basan en ellos para identificar y procurar aquellos animales que mejor se adapten a las condiciones de producción existentes y que al mismo tiempo conduzcan a un incremento del beneficio económico de la actividad. Para esto es necesario valerse de información objetiva y precisa sobre los reproductores, que permita a los criadores, tomar decisiones de selección y hacer un uso diferencial de los mismos.

Como se sabe, la producción de carne en Argentina es una de las más importantes fuentes de ingreso del país. El stock ganadero bovino alcanzó una importante recomposición en el territorio bonaerense: creció un 1,5% entre marzo de 2018 y marzo de 2019, alcanzando las 19,1 millones de cabezas, el nivel más alto desde 2009 [2]. Las actuales existencias bovinas son las mayores de la última década, cuando se había alcanzado un total de 54.816.050 de animales al 31 de marzo de 2018, el stock ganadero bovino muestra una recomposición del 2,7% con respecto al mismo periodo del año 2017, informa el Servicio Nacional de Sanidad y Calidad Agroalimentaria (SENASA) [3].

La ganadería consta de distintos factores, que se deben considerar para que la misma sea exitosa y rentable, como ser la alimentación, la reproducción, la sanidad y la genética. La reproducción de bovinos mediante la Inseminación Artificial (IA) [4] y [5] es bastante sencilla y tiene muchas ventajas. Esta técnica se está aplicando desde hace bastante tiempo en el país. Como se expresa en [6], *"…el problema genético que enfrenta el criador o productor comercial es seleccionar toros que al ser apareados con sus vientres produzcan progenies superiores a aquellas corrientemente producidas…*". La definición de "*superior*" constituye la dirección o rumbo que el criador a la que pretende llegar en su propio rodeo, en cuanto a la calidad genética de sus animales. La IA es uno de los métodos de reproducción, en el cual el hombre ha sustituido el apareamiento natural entre el macho y la hembra. Para poder realizar dicha técnica se debe extraer semen al macho, diluirlo y conservarlo, para luego, mediante una técnica e instrumental adecuado depositarlo en el lugar y momento preciso del aparato reproductor de la hembra con el fin de fecundarla [7].

Entre las razas bovinas para carne, la raza Aberdeen Angus es considerada una de la más extendida en el país.

De tamaño pequeño, crece mucho en una sola temporada y tiene una alta adaptabilidad y facilidad de parto, lo que significa que las hembras son capaces de parir solas. Originaria de Escocia, es de pelaje negro, aunque también hay colorados (red Angus), y no tiene cuernos. Entre sus ventajas es que su engorde es rápido [8]. Esta raza posee lomos anchos y sus cuartos traseros son largos, anchos y musculosos. Además, infiltra grasa, lo que le permite obtener una carne blanda y apetecida. Pese a que tiene una producción de leche muy inferior a las razas doble propósito, llega al fin de la lactancia al inicio del otoño, en una condición corporal superior que las otras razas de carne [9].

Una de las herramientas utilizadas, por los ganaderos, para realizar las evaluaciones genéticas de los toros son la Diferencia Esperada entre Progenie (DEP), que permite a los productores tomar decisiones de selección en base a información objetiva [1]. Los DEP anticipan cómo será el comportamiento promedio de las futuras crías de un toro en comparación con las que producirán el resto de los padres. Por el lado de las madres, a partir de esta información, la selección de los reproductores a utilizar como padres pasa a ser una de las más importantes decisiones de manejo que tiene el productor, permitiéndole seleccionar aquellos animales acordes a sus propios objetivos, su medio ambiente, su sistema de producción, e ir logrando avances genéticos que son acumulativos dentro del rodeo. Otras alternativas utilizadas para la selección por los criadores, de menor a mayor grado de complejidad, son: apreciación visual (Fenotipo); fallos de jurados en exposiciones; pruebas de comportamiento; pruebas de progenie; índices de selección (Ratios); marcadores moleculares y selección genómica, entre otros. [1]

En otras palabras, los DEP son un indicador numérico que predice la calidad genética de las futuras crías de un toro o una vaca respecto de una base de comparación [10]. Estos parámetros son utilizados por el programa de medición del grupo Evaluación de Reproductores Angus[1] (ERA). Dicho programa es desarrollado por el Instituto Nacional de Tecnología Agropecuaria (INTA) y la Asociación Argentina de Angus, la cual lo promueve entre sus socios. Dichos parámetros se calculan con la información del reproductor más la de sus parientes, los valores pueden ser positivos o negativos, mientras que el valor "cero" coincide con el promedio de cada carácter del rodeo [11]. A continuación, se detallan las descripciones de las siglas que componen los DEP's. Estos parámetros fueron extraídos del Anuario Las Lilas 2017[2]:

- PN (Peso al nacer): Expresado en kilos, indica las diferencias genéticas para el PN de las crías de un padre determinado. El peso ajustado al nacer es un predictor indirecto de la facilidad de parto que transmite un toro padre a su progenie.
- PD (Peso al destete): Expresado en kilos y ajustado a los 210 días de vida, indica el mérito genético de un reproductor en transmitir potencial de crecimiento directo a sus crías hasta el momento del destete.
- CM (Combinado materno): Esta variable combina el peso al destete y la aptitud materna en un solo valor, el cual predice la diferencia heredable total para peso al destete de los padres evaluados. El cálculo se realiza adicionando al DEP de aptitud materna la mitad del DEP para peso al destete.
- CE (Circunferencia escrotal): Expresada en centímetros y ajustada por edad de vida, es un indicador indirecto de la fertilidad de los rodeos. Esta variable expresa el potencial de un toro padre en transmitir diferencias genéticas para el tamaño testicular de sus crías
- PF (Peso final): Expresado en kilos, indica la aptitud que tiene un reproductor en transmitir a su progenie capacidad de crecimiento post-destete.
- AM (Aptitud materna): Es un predictor de la producción lechera y aptitud materna que transmite un toro a sus hijas. En otras palabras, es la proporción del peso al destete de las crías que se atribuye a la producción de leche de la madre.
- AOB (Área del Ojo de Bife): Es la altura de la 12a costilla (superficie transversal del músculo dorsal largo, expresada en centímetros cuadrados), siendo un indicador del peso total y rendimiento de cortes despostados de la res.
- GD (Grasa dorsal): Expresada en milímetros, el espesor de grasa dorsal a la altura de la 12a costilla es un predictor genético de la precocidad y facilidad de terminación de las reses.
- MAR (Grado de Marmoreo): El grado de marmóreo es un indicador del porcentaje de grasa intramuscular del músculo dorsal largo, utilizado como patrón indirecto de la palatabilidad (en especial sabor y jugosidad) de los cortes obtenidos.

Como se menciona también en [12], una posible herramienta a utilizar para tratar los grandes volúmenes de información es la técnica de explotación de información. El termino Explotación de información (Minería de Datos: MD). La MD puede definirse como la manera no trivial de extracción de información no implícita, previamente desconocida y potencialmente útil, de una base de datos. Representa la posibilidad de buscar exhaustivamente dentro de un gran volumen de datos, información y conocimiento que pueden resultar de mucho valor. Es considerada uno de los puntos más importantes de los sistemas expertos de base de datos, y uno de los desarrollos más prometedores en la industria del manejo de la información.

Vale destacar que no se encontraron trabajos en Argentina que empleen la MD en la selección de padres reproductores, por lo que se abre una interesante línea de investigación al respecto. A continuación, se listan algunos trabajos relacionados a esta temática llevados a cabo en países como Brasil, Colombia y España. Estos describen

[1] http://www.angus.org.ar/era.php
[2] http://laslilas.com/pdf/Anuario-Genetica-2017.pdf

implementaciones en las que se emplean algunas técnicas de MD, en distintos aspectos del sector ganadero.

En [13] se presenta un trabajo que tuvo como objetivo descubrir la influencia y la relación de las variables de producción y manejo en la bonificación, peso de hacienda, ganancia media diaria y edad de en el caso de las variables de cría con el uso de aplicaciones de MD. Mientras que en Colombia [14], como técnica de MD se empleó una metodología Bayesiana con muestreos de Gibbs, para ajustar modelos lineales univariados. Los autores demostraron que el conocimiento de parámetros genéticos es indispensable en el establecimiento de programas de mejoramiento genético, los cuales tienen como fin optimizar la productividad. Por último, según Flores en [15], se usaron dos técnicas de MD, para establecer qué relación existe entre variables de manera muy intuitiva y visual, cualquier persona pueda analizar y comprender el dominio de predecir el valor genético de la oveja manchega como complemento a la metodología BLUP (Best Linear Unbiased Prediction), permite obtener el valor genético de los animales para una o más características y así seleccionar como reproductores aquellos con mayor mérito genético.

Este trabajo se realiza bajo la hipótesis que si se aplica una metodología para realizar MD a partir de los datos del material genético de los progenitores machos, además de ciertos datos de las vacas, como la información genética de su progenitor y otros datos propios, como la edad, el historial de partos, etc., es posible, confeccionar un modelo descriptivo, el cual mediante la detección de grupos de individuos con características similares permita encontrar patrones o grupos de características que determinen el peso de los terneros al nacer. Este modelo se puede proponer como una herramienta adicional a las existentes, que ayude a los criadores al momento de la toma de decisiones en la cruza de animales. A su vez, dada la experiencia obtenida en trabajos previos de esta investigación en [16], se busca aplicar algoritmos de clasificación, más precisamente el árbol de decisión J48 y las tablas de decisión PART, como métodos alternativos para reducir la dimensionalidad de la muestra seleccionando un subconjunto de atributos.

Materiales y Métodos

Los datos utilizados para este estudio provienen de los rodeos Aberdeen Angus de la estancia El Doce y de Cabaña Las Lilas[3] ambas ubicadas en la localidad de Chascomús, en la provincia de Buenos Aires. Esta Cabaña perteneciente a la Asociación Argentina de Angus[4], es uno de los caudales genéticos más importantes de la Argentina y del Mercosur, produciendo genética en cinco razas líderes entre las que se encuentra la raza Aberdeen Angus. Esto le permite generar los reproductores que van a cubrir una parte importante de los planteles y rodeos generales. En particular, para este trabajo se han tomado como base los valores correspondientes a dos reproductores: Pucará y Al-Mós.

Introducción a la Explotación de información

El presente trabajo se enmarca en lo que se conoce como proceso de Extracción de Conocimiento o KDD (Knowledge Discovery in Databases) el cual consta de una serie de fases que definen la metodología a utilizar. La secuencia de estas fases, que serán explicadas a continuación, no es estricta y frecuentemente hay movimiento entre ellas, dependiendo del resultado de cada fase dando lugar a un proceso de naturaleza cíclica [17].

Como se menciona en [18], la etapa más relevante de este proceso es la MD, que provee mecanismos para la extracción no trivial de información implícita, previamente desconocida a partir de una base de datos y así descubrir reglas y/o patrones significativos de información que puedan ayudar tanto en el diagnóstico correcto del problema como en la formulación de estrategias de solución.

Continuando con la línea de investigación iniciada en [16], en la cual se efectúa una evaluación y comparación de diferentes algoritmos de clasificación del tipo Supervisado, a fin de predecir el PN de los terneros de la raza Aberdeen Angus, en el presente trabajo se pone énfasis en métodos No Supervisados, usando algoritmos de Agrupamiento y aprovechando el conocimiento de los datos adquirido anteriormente.

Según [18], una vez que los datos han sido pre-procesados, la información es considerada una vista minable y está preparada para ser sometida a la técnica que permita establecer el modelo buscado. La MD presenta un amplio espectro de técnicas. La claridad de los resultados va a depender en gran medida de la técnica elegida, es por eso que un análisis previo resulta relevante. Se debe tener presente que la simple aplicación de una técnica de MD a una vista minable y el conocimiento previo del problema, no garantizan patrones expresivos, novedosos y útiles. Los algoritmos muchas veces ofrecen malos resultados debido a causas ajenas a su efectividad, ya sea porque no existe patrón en los datos o porque no se está usando la herramienta adecuada o porque el patrón es realmente difícil de encontrar. Existen dos tipos de tareas, las predictivas y las descriptivas, este estudio se centra en las segundas.

Métodos Descriptivos - No Supervisados

Las tareas descriptivas buscan mostrar nuevas relaciones entre las variables y generalmente son utilizadas para mejorar el modelo. Su objetivo es describir los datos existentes. Entre las tareas descriptivas más frecuentes, puede mencionarse el agrupamiento (clustering), cuyo objetivo es obtener grupos o conjuntos entre los ejemplos, de manera que los elementos asignados al mismo grupo sean similares [17]. A priori no se sabe ni cómo son los grupos ni cuantos hay, eso se determina con el proceso de

[3] http://laslilas.com
[4] https://www.angus.org.ar/

aprendizaje. Es decir, se diferencia de la clasificación ya que no se conocen ni las clases ni su número (aprendizaje no supervisado), con lo que el objetivo es determinar grupos o racimos (clústeres) diferenciados del resto. Una utilidad del agrupamiento reside en que utilizando la función obtenida con nuevos ejemplos se puede determinar a qué grupo pertenece el nuevo elemento y con eso indicar su comportamiento. Otras tareas descriptivas son las correlaciones y factorizaciones, su objetivo es detectar si dos atributos numéricos están correlacionados linealmente o relacionados de algún otro modo. Su utilidad es la detección de atributos redundantes o dependientes y analizar la relevancia de atributos para hacer una selección entre ellos. También encontramos dentro de las tareas descriptivas a las reglas de asociación que realizan un estudio similar al de correlaciones, pero para atributos nominales.

Particularmente para el caso de estudio que se presenta en este trabajo, se dispone de los datos del material genético de los progenitores machos, más ciertos datos de las vacas, como la información genética de su progenitor y otros datos propios, como la edad, el historial de partos, etc., esta información podría ayudar a conocer los perfiles de los animales a cruzar, cuyo modelo describa las características que tienen los progenitores de terneros con bajo PN.

Metodología

Entre las distintas metodologías para llevar a cabo un proyecto de MD existente [19], se optó por Cross Industry Standard Process for Data Mining (CRISP-DM) [20] y [21]. Esta tecnología interrelaciona las diferentes fases del proceso entre sí, de tal manera que se consolida un proceso iterativo y recíproco. A continuación, se desarrolla una breve descripción de las fases utilizadas para la construcción del modelo.

1. Comprensión del negocio

Esta fase inicial se enfocó en la comprensión de los objetivos del proyecto. Después se convirtió este conocimiento de los datos en la definición de un problema de MD y en un plan preliminar diseñado para alcanzar los objetivos. El modelo buscó explorar los datos de los progenitores, para llegar a conocer, de qué manera el material genético, influye en el peso a nacer de las crías.

2. Comprensión de los datos

Esta fase comenzó con la recolección de datos iniciales y continuó con las actividades de integración de datos, que permitieron familiarizarse con los mismos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos relevantes para formular hipótesis en cuanto a la información oculta. Cabe recordar que tomando como referencia la información que se encuentra en el programa genético de la Cabaña La Mariana[5], el PN es uno de los componentes más importantes para evaluar la facilidad de parto, este depende en gran medida del padre utilizado y de la edad de la madre, siendo que cuando la edad de ella se incrementa de 2 a 7 años, también lo hace el PN.

3. Preparación de los datos

Para el desarrollo del modelo, se dispusieron de los datos históricos de los anteriores rodeos y sus resultados. Además de utilizaron los datos provenientes de los DEPs que se describieron en el apartado anterior. Como ya se mencionó, los datos proceden del establecimiento El Doce y corresponden a los años 2017 y 2018. La muestra total se conformó de 360 animales hembras las cuales fueron inseminadas por dos reproductores de la Cabaña Las Lilas. La nómina de variables utilizas se muestra en la siguiente tabla.

Tabla 1. Descripción de las variables del conjunto de datos.

Nomenclatura	Tipo de Dato	Descripción
ID	Numérico	Identificación de la instancia
PAN_Padre	Numérico	Peso al Nacer
PAD_Padre	Numérico	Peso al Destete
PAF_Padre	Numérico	Peso Final
PAdulto_Padre	Numérico	Peso Real
CEsc_Padre	Numérico	Circunferencia Escrotal
Frame_Padre	Numérico	Altura
Certiv_Padre	Numérico	Edad promedio de los vientres primerizos
PNDEP_Padre	Numérico	DEPs del Toro Progenitor (Padre)
PDDEP_Padre	Numérico	
AMDEP_Padre	Numérico	
CMDEP_Padre	Numérico	
PFDEP_Padre	Numérico	
CEDEP_Padre	Numérico	
AOBDEP_Padre	Numérico	
GDDEP_Padre	Numérico	
MARDEP_Padre	Numérico	
EdadMeses_Madre	Numérico	Datos de la Vaca (Madre)
PAN_Madre	Numérico	
PAD_Madre	Numérico	
UltPeso_Madre	Numérico	
CantNac_Madre	Numérico	
CantAbortos_Madre	Numérico	
CantCesareas_Madre	Numérico	
MuertesAntesDestete_Madre	Numérico	
CEsc_AbueloM	Numérico	DEPs del Toro Progenitor de la Vaca (Abuelo Materno)
Frame_ABueloM	Numérico	
Certiv_AbueloM	Numérico	
PNDEP_AbueloM	Numérico	
PDDEP_AbueloM	Numérico	
AMDEP_Abuelo	Numérico	
CMDEP_AbueloM	Numérico	
PFDEP_AbueloM	Numérico	
CEDEP_AbueloM	Numérico	
AOBDEP_AbueloM	Numérico	
GDDEP_AbueloM	Numérico	
MARDEP_AbueloM	Numérico	
Pnacer_Hijo	Texto	Variable Objetivo en [16], una más en este estudio.

Para optimizar los algoritmos, tanto los No Supervisados como los Supervisados propuestos para este

[5] http://www.estancialamariana.com/lacabana.html

proyecto, se normalizaron las variables de entrada a los algoritmos. Normalizar significa, en este caso, comprimir o extender los valores de la variable para que estén en un rango definido. Se empleó la Normalización mínimo-máximo que transforma linealmente los datos a un intervalo, para este caso, entre 0 y 1, donde el valor mínimo se escala a 0 y el máximo a 1[20], que se define como:

$$X_{normalizada} = X - X_{min} / X_{max} - X_{min} \quad (1)$$

3.1. Selección de atributos o características

Tal lo mencionado en [18], uno de los problemas centrales en la MD es identificar un conjunto representativo de características adecuadas para construir un modelo para una tarea en particular. Hay muchos factores que afectan el éxito de una tarea de DM, la calidad de los datos de ejemplo es uno muy importante. En teoría, tener más características debería resultar en una mayor potencia descriptiva o predictiva, sin embargo, la experiencia práctica ha demostrado que no siempre es éste el caso. Los problemas con una alta dimensionalidad, cantidad limitada de ejemplos disponibles y mucha información redundante o irrelevante son difíciles de tratar. Como explica en su libro Hernández Orallo [21], si tenemos muchas dimensiones (atributos) respecto a la cantidad de instancias, pueden existir demasiados grados de libertad, por lo que los patrones extraídos pueden ser poco robustos. Este problema se conoce popularmente como "la maldición de la dimensionalidad" ("*the curse of dimensionality*"). Una manera de intentar resolver este problema es mediante la reducción de dimensiones.

Este caso de estudio claramente presenta estas características: existe un número importante de atributos, 38 en total, y la cantidad de ejemplos se ve limitada por la reducida cantidad de ejemplares del rodeo, disponiéndose tan solo de 360 instancias. Más adelante se explica cómo se aborda este problema, basados en la experiencia obtenida en trabajos previos de esta investigación [16].

4. Modelado

También denominada MD, por ser la más característica del KDD, es la fase en la que se seleccionan y aplican diferentes técnicas de modelado, configurando sus parámetros para la obtención de resultados. Aquí es donde se produce conocimiento nuevo, construyendo modelos a partir de los datos recopilados.

Se utilizó el software WEKA[6] (acrónimo de Waikato Environment for Knowledge Analysis, en español «entorno para análisis del conocimiento de la Universidad de Waikato») que permite ejecutar y evaluar distintos algoritmos de los tipos mencionados anteriormente.

En esta fase, se seleccionaron y aplicaron las técnicas de modelado. En general, se utilizaron las configuraciones por defecto, que propone el software [17]; el cual ha sido previamente recomendado en [21].

Las técnicas utilizadas para realizar la segmentación (clusterización) fueron del tipo No Supervisadas o Descriptivas. A continuación, se las describe brevemente, de acuerdo con [22]:

- **EM (Expectation Maximization):**

Este algoritmo pertenece a una familia de modelos que se conocen como Finite Mixture Models, los cuales se pueden utilizar para segmentar conjuntos de datos. Está clasificado como un método de particionado y recolocación, o sea, clustering (agrupación) probabilístico. Se trata de obtener la FDP (Función de Densidad de Probabilidad) desconocida a la que pertenecen el conjunto completo de datos. El algoritmo EM, procede en dos pasos que se repiten de forma iterativa:

  - Expectation: Utiliza los valores de los parámetros, iniciales o proporcionados por el paso Maximization, obteniendo diferentes formas de la FDP buscada.
  - Maximization: Obtiene nuevos valores de los parámetros a partir de los datos proporcionados por el paso anterior. Finalmente se obtendrá un conjunto de clústeres que agrupan el conjunto de proyectos original. Cada uno de estos clústeres estarán definidos por los parámetros de una distribución.

EM puede decidir cuántos clústeres crear mediante validación cruzada, o puede especificar a priori cuántos clústeres generar. Para el presente experimento se configuró la cantidad de clústeres en 2, a fin de poder comparar con el resto de los algoritmos empezando por un número de grupos reducido.

- **FarthestFirst:**

Este algoritmo de aprendizaje No Supervisado pertenece a los modelos de agrupamiento numérico. Funciona como un agrupador simple, rápido y aproximado. Modelado a partir de SimpleK-Means, podría ser un inicializador útil para él. Selecciona aleatoriamente una instancia que pasa a ser el centro (centroide) del clúster. Se calcula la distancia entre cada una de las instancias y el centro. La distancia que se encuentre más alejada del centro más cercano es seleccionada como el nuevo centro del clúster. Este proceso se repite hasta alcanzar el número de clústeres buscado.

- **Simple K-Means:**

Este algoritmo de aprendizaje no supervisado también pertenece a los modelos de agrupamiento numérico, debe definir el número de clústeres que se desean obtener, lo que lo convierte en un algoritmo voraz para particionar. Los pasos básicos a aplicar son muy simples. Primeramente, se determina la cantidad de clústeres en los que se quiere agrupar la información. Luego se asume de forma aleatoria los centros por cada clúster. Una vez encontrados los primeros centroides el algoritmo efectuará los pasos siguientes:

  - Determina las coordenadas del centroide.
  - Determina la distancia de cada objeto a los centroides. Puede usar la distancia euclidiana (predeterminada) o la distancia de Manhattan. Si se

[6] www.cs.waikato.ac.nz/~ml/weka/

utiliza la distancia de Manhattan, los centroides se calculan como la mediana de los componentes en lugar de la media.
- Agrupa los objetos basados en la menor distancia.
- El proceso se itera, objeto a objeto, hasta que todos los objetos se mantienen en el mismo centroide.

Finalmente quedarán agrupados por clústeres, los grupos de simulaciones según la cantidad de clústeres que el investigador definió en el momento de ejecutar el algoritmo.

- **Mapas Auto Organizados (Redes SOM)**

Según [23], los mapas autoorganizativos de Kohonen son un algoritmo, a veces agrupado dentro de las redes neuronales, que, a partir de un proceso iterativo de comparación con un conjunto de datos y cambios para aproximarse a los mismos, crea un modelo de esos mismos datos que puede servir para agruparlos por criterios de similitud; adicionalmente, este agrupamiento se produce de forma que la proyección de estos datos sobre el mapa distribuya sus características de una forma gradual.

De acuerdo con [24], la idea básica del modelo es crear una imagen de un espacio multidimensional de entrada en un espacio de salida de menor dimensionalidad.

Se trata de un modelo con dos capas de neuronas, una de entrada y otra de procesamiento. Las neuronas de la primera capa se limitan a recoger y canalizar la información. La segunda capa está conectada a la primera a través de los pesos sinápticos y realiza la tarea importante: una proyección no lineal del espacio multi dimensional de entrada, preservando las características esenciales de estos datos en forma de relaciones de vecindad.

El resultado final es la creación del llamado mapa autoorganizado donde se representan los rasgos más sobresalientes del espacio de entrada. Como explica en [25] la diferencia con clustering es que el interés no está puesto en encontrar clústeres de datos similares, sino en cuantizar el espacio de entrada. La densidad de las neuronas y en consecuencia de los subespacios es más alta en aquellas áreas en las cuáles los inputs tienen mayor densidad de probabilidad. En la versión 3.8.2 de WEKA este algoritmo se agrega como un paquete adicional.

5. Interpretación y evaluación de Resultados

Este apartado coincide con la fase Evaluación de la metodología CRISP-DM. En esta etapa, una vez construida la vista minable, se utiliza WEKA para su análisis. Se efectuaron verificaciones con TANAGRA[7], otro paquete gratuito de software de aprendizaje automático para fines académicos y de investigación. desarrollado por Ricco Rakotomalala en la Universidad Lumière Lyon II, Francia, pero dichos resultados serán objeto de otro informe.

Volviendo a WEKA; como conjunto de datos de entrada se utiliza la vista minable generada en la sección de Preparación de datos. Una vez que ingresa este archivo en formato .CSV (del inglés Comma-Separated Values), el software se encarga de darle un formato propietario: ARFF (del inglés Attribute-Relation File Format), que es un archivo de texto ASCII que describe una lista de instancias que comparten un conjunto de atributos y los datos [17].

Inicialmente, se procedió a trabajar con los 38 atributos para realizar las tareas de segmentación, ignorando solamente el atributo ID, ya que se sabe puede distorsionar los resultados. Es notoria la falta de claridad en la correlación de las variables, tal como se muestra en los histogramas de la Figura 1.

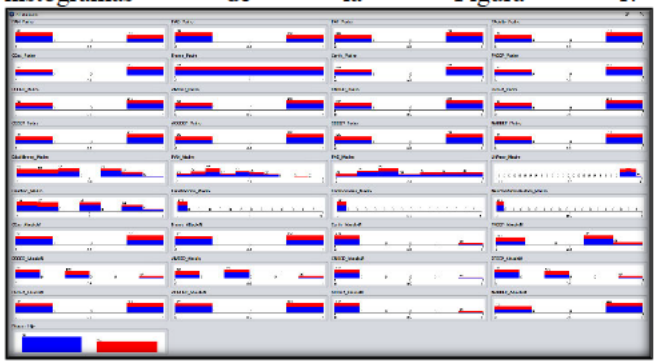

Figura 1. Opción Visualize All una vez cargado el archivo. arff empleado en el estudio.

Una vez cargados los datos, se procedió a entrenar de manera no supervisada los distintos algoritmos con el mismo lote de datos. En la pestaña Clúster (Segmentación), se cuenta con una lista desplegable, de donde se selecciona el tipo de algoritmo a utilizar. Una vez realizado esto, se pueden configurar diferentes parámetros.

En esta oportunidad se ha optado por usar los ofrecidos por la herramienta por defecto, salvo para el algoritmo EM, para el cual se modifica la opción default de Num. Clúster, -1 (automática) por el valor 2. Y para SOM que se modifica el valor de Width (of lattice) de 2 a 1, para agrupar en 2 Clústeres también, con fines de comparación de los distintos algoritmos.

En la misma ventana, es posible elegir las opciones Modo de Agrupación, es decir, la manera de probar las agrupaciones. Las 4 opciones que dispone el producto son:

- **Use Training Set:** Es el usado para este trabajo. Para hacer la prueba se usa el mismo conjunto de datos que el de entrenamiento.
- **Supplied Test Set:** Si se tiene un fichero con datos de prueba distintos a los de entrenamiento, aquí es donde se puede seleccionar.
- **Porcentage Split:** En este caso, se dividirá el conjunto de entrenamiento en dos partes: los primeros 66% de los datos para construir el clasificador y el 33% finales, para hacer la prueba. Se puede modificar el valor de estos porcentajes.
- **Classes to clusters evaluation:** En este modo WEKA primero omite el atributo de clase y genera la agrupación en clústeres. A continuación, durante la fase de prueba asigna clases a los clústeres, en función del valor mayoritario del atributo de clase dentro de cada clúster. A continuación, calcula el error de

[7] http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html

clasificación, en función de esta asignación y también muestra la matriz de confusión correspondiente.

Otra opción disponible siempre es "Ignore atributes", en donde se puede señalar algunas variables de la base de datos que no formarán parte de las variables predictoras para generar el modelo.

5.1. Análisis de los Grupos Resultantes

El *data set* utilizado está conformado por los 38 atributos detallados en la Tabla 1. En primer lugar, se aplican los algoritmos EM, FarthestFirst, SOM y K-Means para agrupar los 360 registros originarios en 2 clústeres. Así se obtiene una primera caracterización de los clústeres.

A continuación, Figura 2, se muestra la Cantidad, en valor absoluto y porcentual, de los casos de la muestra asignado a cada Clúster (Grupo) según cada algoritmo no supervisado.

	NumClust: 2							
	SimpleKMeans		EM		FarthestFirst		SOM	
	Cant.	Cant. %	Cant.	Cant. %	Cant.	Cant. %	Cant.	Cant. %
Cluster 0	210	58%	210	58%	210	58%	210	58%
Cluster 1	150	42%	150	42%	150	42%	150	42%
General	360	100%	360	100%	360	100%	360	100%

NumClust: 2
ClusterMode: Use Full Training Set

Figura 2. Cantidad en valor absoluto y porcentual de los casos de la muestra asignado a cada clúster (grupo) según cada algoritmo no supervisado

Dado el mismo comportamiento en todos estos algoritmos de segmentación, se procedió a continuar las pruebas tomando al algoritmo SimpleK-Means como referencia de agrupación para comprender las características de cada grupo (Clúster). Siguiendo el razonamiento de [12], si los clústeres fueran irrelevantes, se esperaría encontrar una proporción aproximada de 42% azul (Clúster 0) y 58% rojo (Clúster 1) en cada variable de cada atributo. Si bien en algunos atributos esta proporción se cumple, en otros existen interacciones significativas (por ejemplo, Clúster 1 con EdadMeses_Madre y CantNac_Madre). Los atributos donde se observa más interacción entre las variables y los clústeres son, además de los anteriores: PAN_Madre, PAD_Madre, PAD_Madre, UltPeso_Madre, CEsc_ AbueloM, Frame_ABueloM, PDDEP_AbueloM, PFDEP_ AbueloM, Pnacer_Hijo.

Al igual que en [18], luego se enfrentó la tarea de describir los grupos obtenidos a través de sus centroides y calcular los valores frecuentes en los grupos [26]. En todos los casos se utilizó el conocimiento del dominio para guiar la descripción, pero los resultados no fueron satisfactorios dada la cantidad de atributos intervinientes. Las pruebas realizadas aplicando k-medias pusieron en evidencia la gran dimensionalidad del problema, que oscurece la interpretación de los agrupamientos obtenidos. Dada la cantidad de atributos involucrados (todos los de la vista minable inicial), no es posible encontrar un conjunto de clústeres descriptivo de los datos de entrada. La selección de atributos puede ser guiada por el conocimiento del dominio o por técnicas específicas de MD. En este punto surgió la necesidad de utilizar las herramientas de la MD para guiar la selección de un subconjunto de características (atributos) que sean relevantes para el problema. Por esta razón se dejó en suspenso la aplicación de los métodos de agrupamiento y se enfocó la tarea en la utilización de técnicas que permitieran visualizar el conjunto de atributos adecuado para la aplicación de dichos métodos.

5.1.1. Selección de características relevantes

El objetivo en este punto era encontrar un subconjunto de atributos del conjunto total inicial, que incluyera aquellos relevantes para la tarea de agrupamiento.

A partir del conocimiento del dominio y aplicando algunos de los métodos de Selección de Atributos de la herramienta WEKA, se han encontrado tres subconjuntos de atributos significativos, que redujeron la dimensionalidad de los datos. En esta etapa fue clave la experiencia obtenida en trabajos previos de esta investigación [16], donde se aplicaron técnicas supervisadas de clasificación, el árbol de decisión J48 (C4.5), y se agregaron reglas PART, como métodos de validación de la transformación del espacio de características. Se determinó que los datos relevantes para agrupar a los terneros según su PN involucraban principalmente las características de su Madre y del Abuelo Materno. Un dato sobresaliente fue la ausencia de atributos del Padre en el grupo de relevancia, lo cual coincide con el conocimiento del caso, dado que se cuenta con datos de sólo dos toros progenitores; de todos modos, se conservó el atributo PAN_Padre.

Con este nuevo escenario se volvió a probar el comportamiento del algoritmo de agrupación SimpleK-Means. Una vez aplicado el método, el resultado de la asignación a grupos de esta ejecución se comparó con el resultado de la ejecución anterior, determinando que menos de un 10% de los ejemplos se movieron de grupo, lo que indicó que el criterio de agrupamiento se conservó a pesar de la reducción de características, lo cual denota que aún se debe trabajar en la etapa de Preparación de datos. (Figura 3)

	NumClust: 2								NumClust: 3	
	SimpleKMeans		SimpleKMeans(*1)		SimpleKMeans(*2)		SimpleKMeans(*3)		SimpleKMeans	
	Cant.	Cant. %	Cant.	Cant. %	Cant.	Cant. %	Cant.	Cant. %	Cant.	Cant. %
Cluster 0	210	58%	177	49%	245	68%	177	49%	210	58%
Cluster 1	150	42%	183	51%	115	32%	183	51%	75	21%
Cluster 2									75	21%
General	360	100%	360	100%	360	100%	360	100%	360	100%

ClusterMode: Use Full Training Set
(*1): Select Attributes PAN_Padre + Todos Madre y AbueloM
(*2): Select Attributes Classification J48
(*3): Select Attributes Classification PART

Figura 3. Cantidad en valor absoluto y porcentual de los casos de la muestra asignado a cada clúster (grupo) sólo pruebas SimpleK-Means

En cuanto a las agrupaciones, si bien la Selección 1 y 3 coinciden en cuanto a la distribución de los casos en grupos, se han tomado los clústeres resultantes de "Selección Prueba 3" (Classification Reglas PART), si bien tiende a generalizar, los atributos relevantes se acercan a los conocidos en el dominio del problema.

5.1.2. Descripción de perfiles obtenidos

Para interpretar los resultados de los algoritmos de agrupación en base a la "Selección Prueba 3", se regresó a los valores originales de los datos, aplicando la siguiente fórmula:

$$X = (X_{\text{normalizada}} * (X_{max} - X_{\text{min}}) + X_{\text{min}}) \quad (2)$$

La Figura 4 muestra el agrupamiento obtenido por el algoritmo Simple-KMeans con "Selección Prueba 3" (Classification Reglas PART), y su correspondiente vuelta a los valores originales.

Simple K-Means Norm K=2 Solo Attribs PART

ELEGIDO PARA SELECCIÓN DE ATRIBUTOS Y DESCRIBIR LAS CARACTERÍSTICAS DE LOS GRUPOS

Attribute	Full Data (360)	0 (177)	1 (183)	C0 Xorig	C1 Xorig
PAN_Padre	0,4167	0	0,8197	36	36,8197
				230	230
				441	441
				925	925
				40	40
				4,2	4,2
				15	15
				0,4	0,4
				1,8	1,8
				0,2	0,2
				2,9	2,9
				7,9	7,9
				0,4	0,4
				-2,7	-2,7
				0,2	0,2
				-0,1	-0,1
EdadMeses_Madre	0,3933	0,6282	0,1661	61,692	33,966
PAN_Madre	0,3273	0,3854	0,2711	36,9372	34,8798
PAD_Madre	0,4963	0,5085	0,4845	190,51	189,07
UltPeso_Madre	0,4332	0,3933	0,4718	437,3636	444,821
CantNac_Madre	0,3701	0,6356	0,1134	3,5424	1,4836
CantAbortos_Madre	0,025	0,0508	0	0,1016	0
CantCesareas_Madre	0,0167	0,0226	0,0109	0,0226	0,0109
MuertesAntesDestete_Mad	0,0278	0,0452	0,0109	0,0452	0,0109
CEsc_AbueloM	0,5083	0	1	41	42
				4,2	4,2
Certiv_AbueloM	0,126	0	0,2459	15	15,7377
PNDEP_AbueloM	0,5183	0,8	0,2459	0,5	0,22295
				-2,7	-2,7
				-3,5	-3,5
				-3,1	-3,1
				-12,2	-12,2
				0,3	0,3
				-0,2	-0,2
				0,1	0,1
MARDEP_AbueloM	0,5167	1	0,2459	0,1	0,02459
Pnacer_Hijo	Alto	Alto	Bajo	Alto	Bajo

Figura 4. Agrupamiento obtenido por el algoritmo Simple-KMeans con "Selección Prueba 3" (Classification Reglas PART) + vuelva a valores originales

A continuación, en la Figura 5, se muestra la distribución de los clústeres resultantes con la reducción de dimensionalidad:

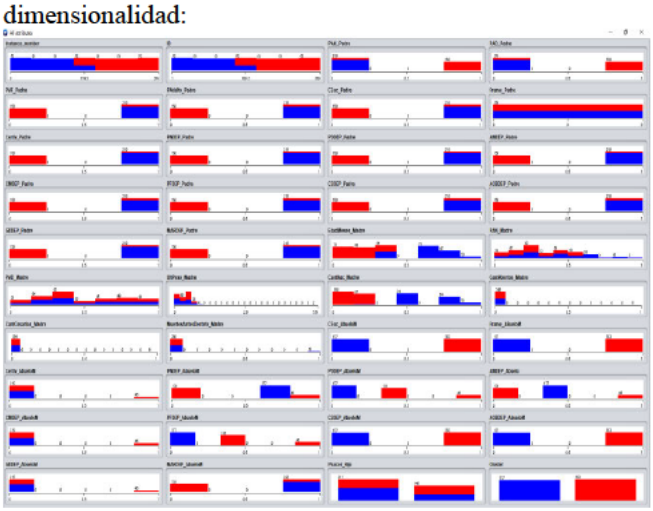

Figura 5. Opción Visualize All una vez cargado el archivo. arff que contenía la prueba (*3): "Selección Prueba 3" (Classification Reglas PART)

En la Tabla 2 se muestran los grupos obtenidos, a través de los valores que pueden tomar los atributos seleccionados:

Tabla 2. Descripción grupos en base a sus características relevantes

# Selección Prueba 3 (CLASSIFICATION REGLAS PART)	Clúster 0 (177/49%)	Clúster 1 (183/51%)
PAN_Padre	36kg	37kg
EdadMeses_Madre	62m	34m
PAN_Madre	37kg	35kg
PAD_Madre	191kg	189kg
UltPeso_Madre	437kg	445kg
CantNac_Madre	3,54	1,45
CantAbortos_Madre	0,10	0,00
CantCesareas_Madre	0,02	0,01
MuertesAntesDestete_Madre	0,05	0,01
CEsc_AbueloM	41cm	42cm
Certiv_AbueloM	15m	16m
PNDEP_AbueloM	0,50	0,22
MARDEP_AbueloM	0,10	0,02
Pnacer_Hijo	Alto	Bajo

Finalmente, como en [18], la segmentación en estos grupos permitió tener mayor conocimiento de los subgrupos que componen la clase de interés (PNacer_Hijo), bajo esta premisa se los pudo describir en base a las características relevantes:

- **(Clúster 0) – Peso al Nacer ALTO (49% - 177 casos)**. Son aquellos donde el PN del Padre fue 36Kg (el menor para este caso de estudio), la Edad en meses de la Madre fue de 62m, el PN y al Destete de la Madre son leventemente superiores que los del Clúster 1, 37 y 192kg respectivamente, no así el Ultimo Peso de la Madre que es un tanto menor, 437kg. Las Madres de los terneros con Alto Peso al Nacer se caracterizan por haber tenido más de 3 nacimientos previos, algún índice de abortos, cesáreas y muertes antes del destete. En cuanto al Abuelo Materno tiene una Circunferencia escrotal de 41cm y un Certiv (Edad promedio de los vientres primerizos) de 15 meses.
- **(Clúster 1) – Peso al Nacer BAJO (51% - 183 casos)**. Son aquellos donde el PN del Padre fue 37Kg (el mayor para este caso de estudio), la Edad en meses de la Madre fue de 33m, mucho menor que el Clúster 0, el PN y al Destete de la Madre son leventemente inferiores que los del Clúster 0, 35 y 189kg respectivamente, no así el Ultimo Peso de la Madre que es un tanto mayor, 445kg. Las Madres de los terneros con Bajo Peso al Nacer se caracterizan por haber tenido un índice de nacimientos previos bajo, lo que también disminuye la cantidad de abortos, cesáreas y muertes antes del destete. En cuanto al Abuelo Materno tiene una Circunferencia escrotal de 42cm y un Certiv (Edad promedio de los vientres primerizos) de 16 meses.

Como puede notarse, algunas de estas características coinciden con el conocimiento de dominio mencionado en la sección de Comprensión de los Datos.

6. Difusión y uso de los resultados

La creación del modelo no implica la finalización del proyecto. El conocimiento obtenido debe ser organizado y presentado de manera que pueda ser comprendido y utilizado por el usuario final. La tarea importante de esta fase consiste en que el usuario entienda los resultados y pueda utilizar los modelos creados. Los trabajos de esta investigación están siendo difundidos en diversos Congresos y Jornadas de Computación.

7. Implementación de medidas basadas en el conocimiento obtenido

Cuando la fase de uso de resultados genera una clase de conocimiento que habilita al usuario a ejecutar acciones en pos de resolver el problema planteado originalmente, se produce una etapa de implementación de medidas que debe llevar a cabo la organización. Estas medidas tendrán como objetivo mejorar o corregir la realidad descubierta a través del modelado, actuando directamente sobre la organización.

8. Medición de resultados

Luego de la implementación de las medidas de la fase anterior, es posible la utilización del DM para medir los resultados alcanzados por esas acciones. En esta fase se pueden volver a ejecutar los modelos para compararlos con los obtenidos en la primera iteración y de esa manera conseguir mediciones concretas del éxito o fracaso de las medidas tomadas.

Conclusiones

De acuerdo con los resultados obtenidos, y en algunos puntos coincidiendo con [12,16,18], se determina que:

- Con este agrupamiento o clustering se pudo efectuar una descripción inicial de los grupos con características comunes alcanzando una visión más clara de los mismos, apuntando principalmente a comprender las relacionadas con las clases del Peso al Nacer de los Terneros.
- Se encontraron métodos híbridos, basados en el conocimiento del dominio y en los métodos de selección de atributos, que ayudaron a la reducir la dimensionalidad de los datos.
- No fue necesario ser un experto ni en temas de sistemas inteligentes ni en el tema a explorar para un análisis inicial. Sin embargo, es condición necesaria contar con la experiencia de especialistas en la temática a analizar a fin de que validen las conclusiones extraídas de los modelos de explotación de información.

Por otro parte, como ya se había notado en [16], confirmamos que si bien los resultados de los modelos, en este caso no supervisados, son aceptables aún se requiere trabajar en los datos de entrada, dado que no contamos con suficientes registros que permitan detectar patrones muy marcados en el comportamiento de los mismo. Esto puede argumentarse en las siguientes razones:

- Luego de reducir la dimensionalidad del conjunto de datos analizado, notamos una leve mejora en la distribución de los casos entre los 2 grupos, aproximadamente el 10% de los ejemplos se movieron de grupo. Esto indicó que el criterio de agrupamiento se conservó a pesar de la reducción de características; pero nos dio la pauta de que puede deberse a un tema de calidad de datos, donde es probable que algunas instancias se hayan definido como pertenecientes a un valor de la variable Peso al Nacer, por ej. “Bajo” cuando en realidad por sus características pertenecía a otro, “Alto”, y viceversa.
- Es sabido que en general, la recolección de datos por parte de los criadores de los rodeos ganaderos se efectúa de manera rudimentaria, por lo que este experimento ayuda a confirmar que es necesario continuar en la línea general propuesta entre los objetivos de este trabajo de investigación, la cual incluye el diseño de una aplicación que permita una ágil captura de los datos de los rodeos.
- A su vez, la reducción excesiva de variables puede llevar a la generalización de los casos perdiendo características interesantes de los mismos.
- Otra causa que pudo provocar la dificultad para encontrar distintos grupos pudo deberse no sólo a la selección de los atributos relevantes, sino también a los valores que toman cada uno de ellos. Se comprobó que, en la muestra de datos, las instancias son muy similares entre sí, por lo cual las medidas de distancia usadas por los algoritmos de segmentación también devuelven valores muy cercanos para sus atributos, lo que provoca que sean agrupados como objetos semejantes.

Trabajos Futuros

Se espera contar con datos de rodeos ganaderos de mayor cantidad de individuos, con toros, vacas y abuelos de características variadas, tales que al cruzarse permitan obtener distintas combinaciones en el PN de los Terneros, con valores más diferenciados entre sus variables. Al tener mayor cantidad de reproductores, aumentarán los valores de sus DEPs. En la medida que se vayamos mejorando la calidad y cantidad de registros:

- Se espera poder generar mayor cantidad de grupos y patrones en base a las características de estos.
- Se podría experimentar utilizando diferentes herramientas computacionales, por ejemplo, probar con software más sofisticados como Matlab[8], SPSS[9], etc.

[8] https://matlabacademy.mathworks.com/es

[9] https://spss.softonic.com/

Referencias

1. Ing. Agr. Luis María Firpo Brenta y el Ing. en Prod. Agropec. Ricardo Firpo. (2012) Selección genética y mejoramiento animal. Revista Angus, Bs. As., 257:38-46. Último acceso: http://www.produccion-animal.com.ar/genetica_seleccion _cruzamientos/bovinos_en_general/24-Seleccion_ genetica.pdf. Último acceso: 06/09/2019.
2. "Destacan la recuperación de stock bovino en Buenos Aires". Disponible en: http://www.revistachacra.com.ar/nota/27133-destacan-la-recuperacion-de-stock-bovino-en-buenos-aires/. Último acceso: 06/09/2019.
3. "SENASA Comunica" Disponible en: http://www.senasa. gob .ar/senasa- comunica/ noticias/el-stock-ganadero-bovino-alcanzo-los-548-millones-de-animales. Último acceso: 06/09/2019.
4. "Curso Teórico Práctico de Inseminación Artificial en Bovinos". Último acceso: https://www.engormix.com /ganaderia-carne/articulos/inseminacion-artificial-en-bovinos-t26957.htm. Último acceso: 06/09/2019.
5. Díaz, P. Fonseca, V. Martínez P. y Rey A. Inseminación Artificial en bovinos. (2003). Biblioteca Digita, U. de Chile. Disponible en: http://www.biblioteca.org.ar/libros/8913.pdf. Último acceso: 06/09/2019.
6. Guitou H. y Monti A. Interpretación y uso correcto de los DEPs como herramienta de selección. (1998). Unidad de Genética Animal, INTA Castelar. Disponible en: https://es.scribd.com/document/337947981/20-Interpretacion-Deps. Último acceso: 06/09/2019.
7. Dr. Daniel Jaime. Manual de Inseminación Artificial en Bovinos. Instituto Nacional de Investigación Agropecuaria. (1993). Disponible en: www.ainfo.inia.uy/ digital/bitstream/item/2737/1/111219240807155445.pdf. Último acceso: 06/09/2019.
8. "Cómo seleccionar la mejor raza bovina de carne". Disponible en: https://www.elmercurio.com/Campo/Noticias/Noticias/ 2012/04/16/Como-seleccionar-la-mejor-raza-bovina-de-carne.aspx. Último acceso: 25/09/2019.
9. "Historia y desarrollo". http://www.angus.org.ar/historia.php. Último acceso: 25/06/2019.
10. "¿En qué consiste la Diferencia Esperada en Progenie? Disponible en: https://www.contextoganadero.com/ganaderia-sostenible/en-que-consiste-la-diferencia-esperada-en-progenie. Último acceso: 06/09/2019.
11. Ing. Agr. Pravia, M. Mejoramiento genético y selección en ganado de carne. Inst. Nac. de Investigación Agropecuaria. (2004). Mejoramiento Genético Animal, INIA Las Brujas, Uruguay. Disponible en: http://www.inia.org.uy/prado/2004/ mejoramiento_genetico_y_seleccio.htm Último acceso: 06/09/2019.
12. Britos, P., Fernández, E., Merlino, H., Pollo-Cataneo, F., Rodríguez, D., Procopio, C., Rancan, C., García-Martínez, R. (2008). *Explotación de Información Aplicada a Inteligencia Criminal en Argentina*. Proceedings del XIV Congreso Argentino de Ciencias de la Computación, Workshop de Ingeniería de Software y Bases de Datos, Artículo 1866. ISBN 978-987-24611-0-2. Disponible en: http://laboratorios.fi.uba.ar/lsi/rgm/comunicaciones/CACIC-2008-1866.pdf. Último acceso: 06/09/2019.
13. da Silva, R., Maciel, T., do N. Lampert, V. Previsão de Indicadores de Qualidade de Carcaças na Pecuária de Corte Através de Aplicações de Mineração de Dados. (2018). CAI, Congreso Argentino de AgroInformática. http://47jaiio.sadio.org.ar/sites/default/files/CAI-37.pdf. Último acceso: 06/09/2019.
14. Taborda, J., Cerón-Muñoz, M., Barrera, D., Corrales, J. y Agudelo, D. Inferencia bayesiana de parámetros genéticos para características de crecimiento en búfalos doble propósito en Colombia. Livestock Research for Rural Development. Volume 27, Article #196. Disponible desde http://www.lrrd.org/lrrd27/10/cero27196.html. Último acceso: 06/09/2019.
15. Flores, M., Gámez, J., Mateo, J. y Puerta, J. Selección genética para la mejora de la raza ovina manchega mediante técnicas de Minería de Datos. Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial, ISSN 1137-3601, Nº. 29, 2006 (Ejemplar dedicado a: Minería de Datos), pags. 69-77. 10. 10.4114/ia. v10i29.879.
16. Spositto, O., Blanco, G., Levi, M. y Bossero, J; *Clasificación del Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos Supervisados*. Trabajo aceptado en la 48º JAIIO 2019. Departamento de Informática de la UNSa, Universidad Nacional de Salta. Septiembre de 2019
17. Witten, Ian H. Eibe, Frank, Mark, Hall y Pal, Christopher. Data Mining. Practical Machine Learning Tools and Techniques. 2nd ed. (2005). Morgan Kaufmann. ISBN: 0-12-088407-0.
18. *Ing.* Fomia, Sonia (Sede Atlántica – UNRN) y Lic. Lanzarini, Laura (Facultad de Informática - UNLP) Evaluación de técnicas de Extracción de Conocimiento en Bases de Datos y su aplicación a la deserción de alumnos universitarios Disponible en: http://sedici.unlp.edu.ar/bitstream/handle/ 10915/ 27523/5442.pdf?sequence=1. Último acceso: 11/09/2019
19. Moine, Juan M., Haedo, A. y Gordillo, Silvia E. (2011). Estudio comparativo de metodologías para minería de datos. WIIC 2011. Disponible en: http://sedici.unlp.edu.ar/ handle/10915/20034. Último acceso: 11/09/2019
20. Han Jiawei. Data Mining: Concepts and Techniques.3ra. Edición. (2011). ISBN 978-0-12-381479-1. Disponible en: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf
21. Hernández Orallo, J. y otros. "Introducción a la minería de datos". Editorial: Pearson. Edición: I. Año 2004
22. Garré, M, Cuadrado, J.J., Sicilia, M.A., *"Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software"*, Dto. de Ciencias de la Computación ETS Ingeniería Informática Universidad de Alcalá, Madrid (2005) Disponible en: http://www.sc.ehu.es/jiwdocoj/remis/docs/GarreAdis05.pdf Ultimo acceso: 11/09/2019
23. Merelo J.J., *Mapa autoorganizativo de Kohonen* Disponible en: http://geneura.ugr.es/~jmerelo/tutoriales/bioinfo/Kohonen .html Ultimo acceso: 11/09/2019
24. Matuk ,Rosana, *Mapas Auto-Organizados Redes Neuronales, DC-FCEyN-UBA Primer Cuatrimestre 2018* Disponible en: https://campus.exactas.uba.ar/pluginfile.php/100845/course/section/15680/MapasAutoorganizados.pdf Ultimo acceso: 11/09/2019
25. Catedra Data Mining Universidad Nacional del Centro (2012), *"Mapas Autoorganizados"*, Disponible en: http://www.exa.unicen.edu.ar/catedras/dmining/clases/Clase_ 9.ppt Ultimo acceso: 11/09/2019
26. Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Data-Centric Systems and Applications*. Springer Disponible en: http://sirius.cs.put.poznan.pl/~inf89721/Seminarium/Web_Data_Mining__2nd_Edition__Exploring_Hyperlinks__Contents__and_Usage_Data.pdf Ultimo acceso: 11/09/2019