Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de la Matanza
ISSN: 2525-1333
Vol.: 5 - Nro. 1 (AGOSTO-2020)

Artículo original

TECNICAS DE PREPROCESAMIENTO DE DATOS EN MODELOS NO SUPERVISADOS APLICADOS AL ESTUDIO GENETICO DE LA RAZA ABERDEEN ANGUS

DATA PRE-PROCESSING TECHNIQUES IN NON-SUPERVISED MODELS APPLIED TO THE GENETIC STUDY OF THE ABERDEEN ANGUS BREED

Osvaldo SPOSITTO [(1)], Gabriel BLANCO [(2)], Lorena MATTEO [(3)]

[(1)] CeDIT UNLaM, Universidad Nacional de La Matanza
sposito@unlam.edu.ar

[(2)] DIIT UNLaM, Universidad Nacional de La Matanza
g2blanco@unlam.edu.ar

[(3)] DIIT UNLaM, Universidad Nacional de La Matanza
lmatteo@unlam.edu.ar

ISSN: 2525-1333. Vol.: 5 - Nro. 1 (AGOSTO-2020)

Resumen:

Es sabido que la construcción de un buen modelo de minería de datos implica invertir la mayor parte del tiempo y esfuerzo en la fase de preprocesamiento de los datos de entrada.

Uno de los problemas centrales es identificar un conjunto representativo de características adecuadas y de buena calidad para construir el modelo de un caso particular.

En este artículo se explican las tareas de preprocesamiento llevadas a cabo para mejorar el conjunto de datos utilizado en la construcción de modelos no supervisados, mediante los cuales se buscan las características de los progenitores de terneros de la raza Aberdeen Angus con bajo peso al nacer.

A su vez se detallan y comparan los resultados previos y posteriores a la aplicación de estas tareas de preprocesamiento.

Debido a que el mayor obstáculo que se presenta en muchos proyectos de ciencias de datos es precisamente la cantidad y calidad de los datos de entrada, mediante este artículo se motiva a poner énfasis en las etapas iniciales de comprensión y preparación de dicho conjunto, por sobre la premura de interpretación y evaluación de los resultados.

Abstract:

Building a robust data mining model involves investing most of the time and effort in the pre-processing phase of the input data.

One of the central problems is to identify a representative and good quality dataset with adequate variables to build the model of a specific case of study.

This article explains the pre-processing tasks carried out to improve the dataset used in the construction of unsupervised models, by means of which the characteristics of the parents of calves of the Aberdeen Angus breed with low birth weight are sought.

In turn, the results before and after the application of these pre-processing tasks are detailed and compared.

Due to the fact that the biggest obstacle that appears in many data science projects is precisely the quantity and quality of the input data, this article try to motivate you to emphasize the initial stages of understanding and preparing the dataset, instead of finding fast evaluation of the results and deployment.

ISSN: 2525-1333. Vol.: 5 - Nro. 1 (AGOSTO-2020)

Palabras Clave: *Minería de Datos, Preprocesamiento, Modelos No Supervisados, Aberdeen Angus, Ganadería*

Key Words: *Data Minig, Pre-processing, Unsupervised Models, Aberdeen Angus, Cattle Raising*

Colaboradores: *Julio BOSSERO, Carolina RAVINALE, Marcelo LEVI*

ISSN: 2525-1333. Vol.: 5 - Nro. 1 (AGOSTO-2020)

I. CONTEXTO

La selección de padres es una de las decisiones más importantes que tiene un productor ganadero, debiendo elegir aquellos animales acordes a sus propios objetivos, su medio ambiente y su sistema de producción, para así lograr avances genéticos acumulativos dentro del rodeo y un incremento del beneficio económico de su actividad.

La línea de investigación que se presenta en este artículo está enmarcada dentro del Programa de Incentivos para Docentes Investigadores de la Secretaría de Políticas Universitarias (PROINCE) bajo el título: "Uso de Minería de Datos para Mejoramiento Genético en la raza Aberdeen Angus", y es financiado por la Universidad Nacional de La Matanza. En dicho trabajo se pretende brindar una herramienta complementaria que ayude a los criadores ganaderos a la hora de seleccionar los mejores reproductores de dicha raza, en base a la dirección o rumbo del rodeo.

Se destaca el hecho de no haber encontrado trabajos en Argentina que empleen la minería de datos en la selección de padres reproductores, por lo que nos encontramos ante una interesante línea de investigación al respecto, en especial a lo que refiere a la facilidad de parto.

Las tareas de preprocesamiento de datos que se detallarán en el presente artículo forman parte de uno de los experimentos llevados a cabo en dicha investigación, titulado "Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos No Supervisados" y publicado en CoNaIISI 2019 [1].

II. INTRODUCCIÓN

El presente trabajo se enmarca en lo que se conoce como proceso de Extracción de Conocimiento o KDD por sus siglas en inglés (Knowledge Discovery in Databases) el cual consta de una serie de fases que definen la metodología a utilizar.

Según [2], una etapa relevante en este proceso es la minería de datos, ya que brinda mecanismos para la extracción no trivial de información implícita, previamente desconocida a partir de una fuente de datos para así descubrir reglas y/o patrones significativos de información.

Como ya se mencionó, esta investigación se orienta a la búsqueda de una herramienta complementaria que ayude a los productores ganaderos a la hora de seleccionar progenitores dentro de un rodeo bovino de la raza Aberdeen Angus (AnGus).

Para ello se probaron distintas técnicas de minería de datos en base a información obtenida de dos cabañas de la localidad de Chascomús, en la Provincia de Buenos Aires. Una de estas cabañas tiene como objetivo obtener crías con bajo peso al nacer (PN). Este indicador numérico, expresado en kilogramos (kg), es un predictor indirecto de la facilidad de parto, siendo los valores más bajos los más favorables. El 80% de los problemas de parto están relacionados con el peso al nacer, lo cual está documentado en la literatura.[1]

El PN forma parte de un conjunto mayor de características conocidas como DEPs, Diferencia Esperada de Progenie, dentro del Programa ERA impulsado por el INTA y la Asociación Argentina AnGus, y son una de las herramientas utilizadas por los ganaderos para realizar las evaluaciones genéticas. Los DEPs predicen el verdadero mérito genético de un toro, basando su cálculo no sólo sobre su propia performance

[1] http://www.angus.org.ar/

ISSN: 2525-1333. Vol.: 5 - Nro. 1 (AGOSTO-2020)

sino también sobre la información de performance disponible sobre sus progenies y parientes. Otros DEPs son peso al destete, peso final, área de ojo de bife, grasa dorsal, grado de marmóreo, combinado materno, etc.

Tal lo expresado en [3], dentro de las fases del proceso KDD, una vez que los datos han sido preprocesados, la información es considerada una vista minable y está preparada para ser sometida a las técnicas que permitan establecer el modelo buscado. Muchas veces, con el afán de obtener resultados tempranos, las etapas iniciales del proceso se subestiman.

La claridad de los resultados puede depender de la técnica elegida, pero es importante tener presente que el conocimiento previo del problema y la simple aplicación de una técnica de minería de datos a una vista minable, no garantizan patrones expresivos, novedosos y útiles. Los algoritmos muchas veces no ofrecen buenos resultados debido a causas ajenas a su efectividad, ya sea porque no existe patrón en los datos o este es difícil de encontrar, porque no se está usando la herramienta adecuada o bien debido a la calidad de los datos del conjunto a analizar.

En el caso de estudio, el mayor obstáculo que se presentó fue la cantidad y calidad de los datos de entrada, si bien se logró encontrar tendencias y verificar ciertos resultados basados en el conocimiento del dominio, fue necesario trabajar en la muestra original.

En el presente artículo se detallarán las tareas de preprocesamiento de datos más importantes que fue necesario aplicar para mejorar el conjunto de datos de entrada:

- discretización de la variable Peso al Nacer
- normalización mínimo-máximo de las variables de entrada
- reducción de dimensionalidad

III. MÉTODOS

Entre las distintas metodologías existentes para llevar adelante un proyecto de minería de datos, se optó por Cross Industry Standard Process for Data Mining (CRISP-DM) dado que esta tecnología interrelaciona las diferentes fases del proceso entre sí, de tal manera que se consolida un proceso iterativo y recíproco, lo cual lo hace flexible.

El análisis realizado se basó en datos provenientes de El Doce, un establecimiento ganadero de la localidad de Chascomús, Provincia de Buenos Aires, recolectados durante el periodo 2017-2018. La muestra total se conformó de 360 animales hembras de dicho rodeo, las cuales fueron inseminadas por dos reproductores machos de Cabaña Las Lilas.

La vista minable inicial se generó con ciertos datos de las vacas más indicadores provenientes de las evaluaciones genéticas de los toros: conocidos como la Diferencia Esperada entre Progenie (DEP) los cuales permiten a los productores tomar decisiones de selección en base a información objetiva.

La nómina de variables utilizadas para generar la vista minable se muestra en la Tabla 1.

ISSN: 2525-1333. Vol.: 5 - Nro. 1 (AGOSTO-2020)

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.

- ***Discretización de la variable Peso al Nacer (PN)***

Desde un primer comienzo fue necesario categorizar la variable Peso al Nacer, la cual se definió como “Alta” y “Baja”. Un PN promedio es de 38 kg.; por tal motivo, a los pesos mayores o iguales a 38 kg. se los clasificó como Alto y al resto como Bajo.

- ***Normalización Mínimo-Máximo de las variables de entrada***

Para optimizar los algoritmos fue necesario normalizar las variables de entrada mediante la función de normalización mínimo-máximo.

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[A], definiéndose como:

$$X_{\text{normalizada}} = X - X_{\text{min}} / X_{\text{max}} - X_{\text{min}}$$

- ***Reducción de Dimensionalidad***

En general, se suele pensar que cuanto mayor sea la cantidad de características de un caso de estudio mayor será la potencia de predicción o descripción del modelo resultante, sin embargo, la experiencia demuestra que esto no siempre es así. Tener muchas dimensiones (atributos) respecto a la cantidad de instancias, genera demasiados grados de libertad, lo cual deriva no solo en la obtención de patrones poco robustos si no también en bajo rendimiento computacional (más coste y más tiempo), aumentando la complejidad lo cual dificulta la comprensión del modelo y de sus resultados.

Este caso de estudio claramente presentaba estas características, un número importante de atributos, 38 en total, y la cantidad de ejemplos se veía limitada por la reducida cantidad de ejemplares del rodeo, disponiéndose tan solo de 360 instancias.

Es decir, nos encontramos frente a un problema popularmente conocido como “la maldición de la dimensionalidad”, el cual provoca que los patrones

ISSN: 2525-1333. Vol.: 5 - Nro. 1 (AGOSTO-2020)

extraídos sean poco robustos, tal como Hernández Orallo describe en su libro [B].

Detectar esta situación llevó a replantearse la calidad del conjunto de datos con el que se contaba hasta ese momento, decidiendo seguir adelante utilizando esa vista minable como entrada para los algoritmos no supervisados que se habían elegido para el estudio.

El plan se basó en analizar los resultados obtenidos en ese contexto inicial para posteriormente poder compararlos con los que se obtuviesen una vez efectuadas las mejoras en la muestra de datos.

Gracias a la selección de CRISP-DM como metodología de procesos iterativa, fue posible volver a la fase de comprensión y preparación de datos para efectuar una de las tareas de pre-procesamiento más difundidas, la selección de atributos a fin de lograr una reducción de dimensiones.

En ningún momento se perdió el rumbo del objetivo principal del experimento, el cual buscaba identificar patrones o grupos de características en los valores genéticos de los animales reproductores que determinen el peso de los terneros al nacer de la cría de la raza Aberdeen Angus

En lo que respecta a las herramientas de minería de datos, tanto para las tareas de selección de atributos como para el entrenamiento y las pruebas de las técnicas no supervisadas, se utilizó WEKA, acrónimo de Waikato Environment for Knowledge Analysis, un software de código abierto desarrollado por la Universidad de Waikato de Australia, con fines educativos y de investigación.

Siguiendo el plan propuesto, en primer lugar, se analizaron los resultados proporcionados por los algoritmos de segmentación/ agrupamiento del tipo no supervisado elegidos previamente para el análisis, a saber:

- Expectation Maximization (EM)
- FarthestFirst
- Simple K-Means
- Mapas Auto Organizados (Redes SOM)

Trabajar con métodos no supervisados implica no tener un atributo etiqueta o clase con valores predefinidos en el conjunto de datos, los algoritmos agrupan los datos dentro de un número de clases preestablecidas o no, partiendo de criterios de distancia o similitud. De esta forma se agrupan las clases que sean similares entre sí y distintas con las otras clases.

Inicialmente y utilizando WEKA, se procedió a trabajar con los 38 atributos para realizar las tareas de segmentación, ignorando solamente el atributo ID, ya que se sabía podía distorsionar los resultados.

Como era de esperar fue notoria la falta de claridad en la correlación de las variables, tal como se muestra en los histogramas de la Figura 1.

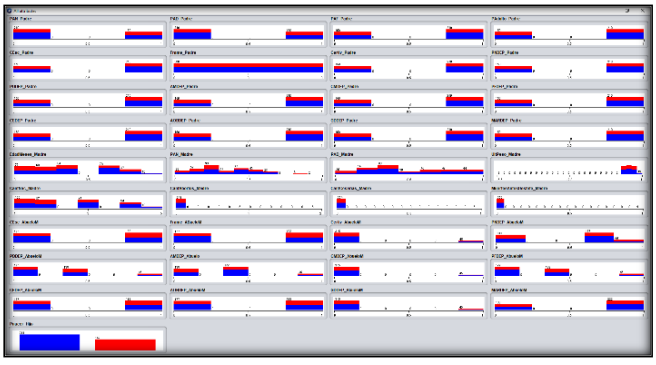

Fig. 1. Opción Visualize All una vez cargada la vista minable empleada en el estudio.

ISSN: 2525-1333. Vol.: 5 - Nro. 1 (AGOSTO-2020)

A continuación, 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 (Figura 2).

	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

Fig. 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 considerando al algoritmo SimpleK-Means como referencia de agrupación para comprender las características de cada grupo (Clúster).

Siguiendo el razonamiento de [3], si los clústeres fueran irrelevantes, se hubiese esperado 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 cumplió, en otros hubo interacciones significativas (por ejemplo, Clúster 1 con EdadMeses_Madre y CantNac_Madre). Los atributos donde se observó una mayor interacción entre variables y clústeres, además de los anteriores, fueron: PAN_Madre, PAD_Madre, PAD_Madre, UltPeso_Madre, CEsc_ AbueloM, Frame_ABueloM, PDDEP_AbueloM, PFDEP_ AbueloM, PNacer_Hijo.

En este momento, se enfrentó la tarea de describir los grupos obtenidos a través de sus centroides y calcular los valores frecuentes en los grupos [4].

En todos los casos se utilizó el conocimiento del dominio para guiar la descripción, pero como se esperaba los resultados no fueron satisfactorios dada la cantidad de atributos intervinientes. Las pruebas realizadas aplicando K-Means pusieron en evidencia la gran dimensionalidad del problema, lo cual complicó la interpretación de los agrupamientos obtenidos. En otras palabras, debido a la cantidad de atributos involucrados en la vista minable inicial, no fue posible encontrar un conjunto de clústeres descriptivo de los datos de entrada.

- ***Selección de características relevantes***

Continuando con el plan mencionado, se procedió a llevar adelante una de las tareas de preprocesamiento de datos que se utiliza frecuentemente para mejorar la calidad de una muestra: la selección de atributos. Esta tarea puede ser guiada por el conocimiento del dominio o por técnicas específicas de minería de datos.

En este punto surgió la necesidad de utilizar las herramientas de minería de datos para guiar la selección de un subconjunto de características (atributos) que sean relevantes para el problema. Es decir, hallar un subconjunto de atributos del conjunto total inicial, que incluya 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 encontraron 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 [5], donde se aplicaron técnicas supervisadas de clasificación por lo que aquí también se generó un el árbol de decisión J48

ISSN: 2525-1333. Vol.: 5 - Nro. 1 (AGOSTO-2020)

(C4.5) y reglas PART, como métodos de validación de la transformación del espacio de características.

Como resultado de ello se determinó que los datos relevantes para agrupar a los terneros según su Peso al Nacer 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 en particular de la calidad de este conjunto de datos, dado que se cuenta con datos de sólo dos toros progenitores; de todos modos, se conservó el atributo PAN_Padre.

En total se consideraron 3 selecciones diferentes para comparar resultados, a saber:

(1) Select Attributes PAN Padre + Todos Madre y Abuelo Materno
(2) Select Attributes Classification J48
(3) Select Attributes Classification PART

IV. RESULTADOS Y OBJETIVOS

Siguiendo lo propuesto, y bajo el nuevo escenario, una vez efectuada la reducción de dimensionalidad del conjunto de datos, se volvió a probar el comportamiento del algoritmo de agrupación SimpleK-Means. 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 será necesario 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

Fig. 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 las Selecciones 1 y 3 coincidieron en cuanto a la distribución de los casos en grupos, se consideraron los clústeres resultantes de "Selección Prueba 3" (Classification Reglas PART), ya que, si bien tendía a generalizar, los atributos relevantes se acercaban a los conocidos en el dominio del problema.

- ***Descripción de perfiles obtenidos***

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

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

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.

Revista Digital del Departamento de
Ingeniería e Investigaciones
Tecnológicas de la Universidad
Nacional de la Matanza
ISSN: 2525-1333. Vol.: 5 - Nro. 1 (AGOSTO-2020)

Simple K-Means Norm K=2 Solo Attribs PART

ELEGIDO PARA SELECCIÓN DE ATRIBUTOS Y DESCRIBIR LAS CARACTERISTICAS 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	437,3635	444,821
CantNac_Madre	0,3701	0,6356	0,1134	3,5424	1,4536
CantAbortos_Madre	0,025	0,0508	0	0,1016	0
CantCesareas_Madre	0,0167	0,0226	0,0109	0,0226	0,0109
MuertesAntesDestete_Mac	0,0278	0,0452	0,0109	0,0452	0,0109
CEsc_AbueloM	0,5083	0	1	41	42
				4,2	4,2
Certiv_AbueloM	0,125	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	0,6167	1	0,2459	0,1	0,02459
Pnacer_Hijo	Alto	Alto	Bajo	Alto	Bajo

Fig. 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:

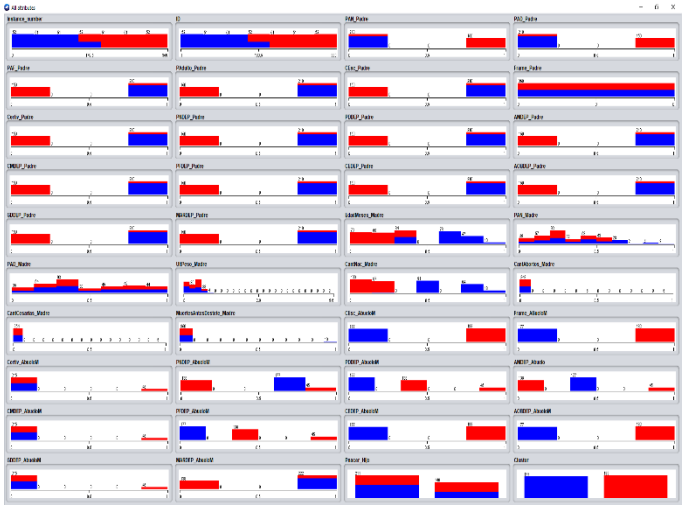

Fig. 5. Opción Visualize All una vez cargada la nueva vista minable que contenía la prueba (*3): "Selección Prueba 3" (Classification Reglas PART)

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

TABLA 2. DESCRIPCIÓN DE LOS 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 [2], se puede afirmar que 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

ISSN: 2525-1333. Vol.: 5 - Nro. 1 (AGOSTO-2020)

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.

Algunas de estas características coinciden con el conocimiento de dominio que se tenía en la etapa de Comprensión de los Datos de la metodología usada en el proyecto.

De todos modos, es importante tener en cuenta que la reducción excesiva de variables puede llevar a la generalización de los casos perdiendo características interesantes de los mismos, por lo que debe aplicarse con cierta cautela.

Estos resultados muestran que aún es necesario trabajar en los datos de entrada, por no contar con suficientes registros que permitan detectar patrones muy marcados en el comportamiento de estos, pero esto no debería ser un escollo, sino más bien una motivación para continuar con la investigación en un futuro, ahondando en los siguientes puntos detectados gracias a este experimento:

- luego de reducir la dimensionalidad del conjunto de datos analizado, se notó una leve mejora en la distribución de los casos entre los 2 grupos, aproximadamente el 10% de los ejemplos se movieron de grupo. Esto indica que el criterio de agrupamiento se conserva a pesar de la reducción de características; pero da la pauta de que aún se debe mejorar la calidad de datos, donde es probable que algunas instancias se hayan definido (al generar la muestra) 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.
- 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. Por lo que es evidente que otra de las causas que pudo provocar la dificultad para encontrar distintos grupos no sólo estuvo relacionada a la selección de los atributos relevantes, sino también a los valores que toma cada uno de ellos.
- 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 se confirma que es necesario diseñar una aplicación que permita una captura ágil de los datos, lo cual conllevará a mejorar la calidad de la muestra y desde el inicio del proyecto de investigación fue uno de sus objetivos.

ISSN: 2525-1333. Vol.: 5 - Nro. 1 (AGOSTO-2020)

V. DISCUSIÓN

Los resultados arrojados por los modelos no supervisados fueron aceptables, confirmando la importancia de poner en práctica técnicas de preprocesamiento de datos que evidencian mejoras en la interpretación de los grupos de características buscados.

Se puede afirmar que con el agrupamiento/segmentación obtenido fue posible 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.

A su vez se encontraron métodos híbridos, basados en el conocimiento del dominio y en los métodos de selección de atributos, que ayudaron a reducir la dimensionalidad de los datos.

VI. CONCLUSIONES

A lo largo del artículo se exhibieron los resultados obtenidos en un experimento presentado y publicado en CoNaIISI 2019 poniendo énfasis en la etapa del prepocesamiento de datos a fin de mejorar la calidad de datos de entrada y la interpretación de los resultados esperados. El análisis corroboró que el mayor obstáculo estuvo relacionado a la cantidad y calidad de la muestra inicial, y que si bien fue posible encontrar tendencias y verificar ciertos resultados basados en el conocimiento del dominio, aún es necesario trabajar en tales datos, siendo lo deseado:

- obtener nuevas cifras de establecimientos ganaderos similares
- incorporar nuevas variables como ser el tipo de alimentación de los animales, el factor climático, entre otros.
- sumar algún software más sofisticado.
- aplicar otras técnicas de preprocesamiento de datos como ser SMOTE.

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 lo que se espera que los resultados logrados de forma automática por los algoritmos puedan compararse con los criterios de los expertos basados en los datos de los terneros a nacer en años subsiguientes.

Por último, se están realizando tratativas con expertos de la Sociedad Rural de Chascomús para la obtención de datos de mejor calidad y de nuevos establecimientos ganaderos. También con la Asociación Argentina AnGus, en una posible colaboración para la captura de los datos requeridos para generar un DEP específico de Facilidad de Parto, los involucrados al 20% restante, no relacionados al Peso al Nacer. [6]

VII. REFERENCIAS Y BIBLIOGRAFÍA

A. Referencias bibliográficas:

[1] O. Spositto, G. Blanco, L. Matteo, “Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos No Supervisados”, 2019. [En línea]: https://www.researchgate.net/publication/337445353_Peso_al_Nacer_de_Terneros_Aberdeen_Angus_mediante_Algoritmos_No_Supervisados [Último acceso: 27/07/2020]

ISSN: 2525-1333. Vol.: 5 - Nro. 1 (AGOSTO-2020)

[2] S. Fomia, L. Lanzarini, "Evaluación de técnicas de Extracción de Conocimiento en Bases de Datos y su aplicación a la deserción de alumnos universitarios" [En línea] http://sedici.unlp.edu.ar/bitstream/handle/10915/27523/Documento_completo.pdf?sequence=1&isAllowed=y [Último acceso: 28/07/2020].

[3] P. Britos, E. Fernández, H. Merlino, MF. Pollo-Cataneo, et.al, "Explotación de Información Aplicada a Inteligencia Criminal en Argentina", 2008. [En línea]: http://laboratorios.fi.uba.ar/lsi/rgm/comunicaciones/CACIC-2008-1866.pdf [Último acceso: 29/07/2020]

[4] B. Liu, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Data-Centric Systems and Applications", 2011, Springer. [En línea]: http://sirius.cs.put.poznan.pl/~inf89721/Seminarium/Web_Data_Mining__2nd_Edition__Exploring_Hyperlinks__Contents__and_Usage_Data.pdf [Último acceso: 28/07/2020].

[5] O. Spositto, G. Blanco, M. Levi, J. Bossero, "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.

[6] "Resumen de Padres ANGUS 2019", 23. Apéndice D: DEP de Facilidad de Parto, página 226, [En línea]: https://www.angus.org.ar/finder/files/toros/Resumen_de_Padres_Angus_1-72.pdf [Último acceso: 29/07/2020]

B. Bibliografía:

[A] H. Jiawei, *Data Mining: Concepts and Techniques,* 3ra. Edición, 2011

[B] J. Hernández Orallo, *Introducción a la minería de datos,* Ed. Pearson, Edición I, 2004

Recibido: 2020-07-30
Aprobado: 2020-08-07
Hipervínculo Permanente: http://www.reddi.unlam.edu.ar
Datos de edición: Vol. 5-Nro. 1-Art. 7
Fecha de edición: Formato: 2020-08-15