Departamento:
Departamento de Ingeniería e Investigaciones Tecnológicas
Programa de acreditación:
PROINCE
Programa de Investigación[1]:
Código del Proyecto:
C223
Título del proyecto
Uso de Minería de Datos para Mejoramiento Genético en la raza Aberdeen Angus
PIDC: ☐
Elija un elemento.
PII: ☐
Elija un elemento.
Director:
SPOSITTO, Osvaldo Mario
Director externo:

Codirector:
BLANCO, Gabriel Esteban
Integrantes:
BOSSERO, Julio César
GARGANO, Cecilia Victoria
LEVI, Marcelo Jorge
MACIAS CORRAL, Patricio Ezequiel
MATTEO, Lorena Romina
RAVINALE, Carolina Mabel
RYCKEBOER, Hugo Emilio

Investigador Externo, Asesor- Especialista, Graduado UNLaM:

Alumnos de grado: (Aclarar si tiene Beca UNLaM/CIN)

Alumnos de posgrado:

Resolución Rectoral de acreditación: N°
350/2019
Fecha de inicio:
01/01/2019
Fecha de finalización:
30/04/2020

[1] Los Programas de Investigación de la UNLaM están acreditados con resolución rectoral, según lo indica la Resolución HCS Nº 014/15 sobre **Lineamientos generales para el establecimiento, desarrollo y gestión de Programas de Investigación a desarrollarse en la Universidad Nacional de La Matanza**. Consultar en el departamento académico correspondiente la inscripción del proyecto en un Programa acreditado.

A. Desarrollo del proyecto (adjuntar el protocolo)

A.1. Grado de ejecución de los objetivos inicialmente planteados, modificaciones o ampliaciones u obstáculos encontrados para su realización (desarrolle en no más de dos (2) páginas)

El balance de los dos años de proyecto es más que positivo. No sólo se logró orientar la investigación 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 si no que este trabajo permitió al equipo de investigación afianzar los conocimientos en el área de la minería de datos, ganando confianza para publicar artículos y participar en Congresos. A pesar de que la segunda etapa del proyecto se vio algo afectada como consecuencia de las medidas excepcionales establecidas por el Gobierno Nacional a raíz del contexto crítico mundial de pandemia COVID-19 hemos aprovechado al máximo un punto significativo que generó la pandemia: la virtualidad.

A lo largo de estos dos años hemos participado en 4 Congresos *(CONAIISI 2019, JAIIO 2019, WICC 2020, CACIC 2020)* con la satisfacción de haber alcanzado una mención especial en uno de dichos eventos como "Trabajo de Investigación Premiado en el área temática de Base de Datos". A su vez publicamos un artículo en la Revista ReDDI del Dpto de Investigaciones Tecnológicas de la Universidad Nacional de La Matanza y recibimos la invitación para publicar en Journal of Computer Science & Technology (JCS&T) de la Universidad Nacional de La Plata.

Por otra parte, el trabajo presentado en el CONAIISI 2019, fue seleccionado para exponer ante representantes del Centro de Innovación Tecnológica, Empresarial y Social (CITES) del Grupo Sancor Seguros, con el fin de dar visibilidad a lo investigado por el grupo. A su vez gracias a ellos tuvimos contacto directo con los Sres Mariano Fernandez Alt, Coordinador del Programa ERA, Asistente del Área Técnica y Coordinador de la Revista Angus, y con el Sr. Horacio Guitou, Médico Veterinario y Consultor en Genética Animal, ambos miembros de la Asociación Argentina de AnGus, quienes nos brindaron sus conocimientos y valiosa ayuda en temas técnicos ganaderos para el proyecto.

En lo que respecta a la transferencia hacia actividades de docencia, los conocimientos adquiridos aplican a la materia Inteligencia de Negocios de la carrera Licenciatura de Gestión de Tecnología de la Información (Formación Continua). A su vez, los nuevos procedimientos, herramientas y metodologías de laboratorio permiten una transferencia directa a las clases experimentales de laboratorio que se implementan en dicha cátedra o de algún curso o taller que surja al respeto, dentro de la universidad y/o en Instituciones con las cuales UNLaM haya firmado Convenios de Cooperación e Intercambio Internacional.

En cuanto la visibilidad del proyecto, los artículos de esta investigación se encuentran publicados en los sitios de divulgación científica como ser Google Scholar y Research Gate. En este último las estadísticas informan cerca de 850 Reads, desde más de 5 países como ser: Colombia, México, Chile y Rusia, entre otros varios hitos alcanzados en los últimos años según los indicadores que se controlan en el sitio.

Finalmente, aún luego de que este proyecto finalice, es deseable continuar con el aporte de conocimiento y resultados de la investigación en el área de la ganadería, avanzando en las tratativas con expertos de la Sociedad Rural para la obtención de datos de mejor calidad y sumando el interés de nuevos establecimientos ganaderos de Aberdeen Angus. Otra acción importante que sería deseable continuar para proyectos futuros son las conversaciones iniciadas con los miembros de la Asociación

Argentina AnGus, para una posible colaboración en la captura de los datos requeridos en la generación de un DEP específico de Facilidad de Parto, los involucrados al 20% restante no relacionado al PN.

B. Principales resultados de la investigación

B.1. Publicaciones en revistas (informar cada producción por separado)

Artículo 1:	
Autores	*Osvaldo SPOSITTO* *Gabriel BLANCO* *Lorena MATTEO*
Título del artículo	*TECNICAS DE PREPROCESAMIENTO DE DATOS EN MODELOS NO SUPERVISADOS APLICADOS AL ESTUDIO GENETICO DE LA RAZA ABERDEEN ANGUS*
N° de fascículo	*Nro 1*
N° de Volumen	*Volumen 5*
Revista	*ReDDI – Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza*
Año	*2020-08*
Institución editora de la revista	*UNLaM – Universidad Nacional de La Matanza*
País de procedencia de institución editora	*Argentina*
Arbitraje	Elija un elemento.
ISSN:	*2525-1333*
URL de descarga del artículo	*https://reddi.unlam.edu.ar/index.php/ReDDi/article/view/119*
N° DOI	

B.2. Libros

Libro 1	
Autores	
Título del Libro	
Año	
Editorial	
Lugar de impresión	
Arbitraje	Elija un elemento.
ISBN:	
URL de descarga del libro	
N° DOI	

B.3. Capítulos de libros

Autores	*Osvaldo Spositto, Gabriel Blanco, Marcelo Levi, Patricio Macías Corral, Lorena Matteo*
Título del Capitulo	*Trabajos Premiados (Base de Datos): mención especial* *Y* *Trabajos de investigadores* *PESO AL NACER DE TERNEROS ABERDEEN ANGUS MEDIANTE ALGORITMOS NO SUPERVISADOS*
Título del Libro	*Libro de Actas - 7mo CONAIISI 2019 DIIT UNLaM - RIISIC* *2019 CONAIISI: VII Congreso Nacional de Ingeniería Informática: Sistemas de Información*
Año	*2019*
Editores del libro/Compiladores	*Bettina Donadello*
Lugar de impresión	*UNLaM San Justo, Buenos Aires*
Arbitraje	Elija un elemento.
ISBN:	*978-987-4417-73-2*
URL de descarga del capítulo	*https://conaiisi2019.unlam.edu.ar/pdf/2019-CONAIISI-ACTAS-7MA-EDICION.pdf*
N° DOI	

Autores	*Osvaldo Spositto, Gabriel Blanco, Lorena Matteo, Marcelo Levi, Julio Bosero*
Título del Capitulo	*12780 Ciencia de datos aplicada al mejoramiento genético de la raza Aberdeen Angus* *54*
Título del Libro	*Actas del XXII Workshop de Investigadores en Ciencias de la Computación: WICC 2020* *y* *POSTERS WICC 2020 XXII Workshop de Investigadores en Ciencias de la Computación*

Año	*2020*
Editores del libro/Compiladores	*Rodolfo Bertone... [et al.]; compilado por Marta Lasso*
Lugar de impresión	*Universidad Nacional de la Patagonia Austral, Río Gallegos, Santa Cruz*
Arbitraje	Elija un elemento.
ISBN:	*978-987-3714-82-5*
URL de descarga del capítulo	*https://libros.unlp.edu.ar/index.php/unlp/catalog/view/1532/1514/4929-1* y *https://libros.unlp.edu.ar/index.php/unlp/catalog/view/1521/1503/4896-1*
N° DOI	

B.4. Trabajos presentados a congresos y/o seminarios

Autores	*Osvaldo Spósitto, Gabriel Blanco, Marcelo Levi, Patricio Macías Corral, Lorena Matteo*
Título	*Artículo 278. - "Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos No Supervisado"*
Año	*2019*
Evento	*VII Congreso Nacional de Ingeniería Informática/Sistemas de Información (CONAIISI 2019)*
Lugar de realización	*Universidad Nacional de La Matanza (UNLaM), San Justo, Buenos Aires, Argentina*
Fecha de presentación de la ponencia	*14/Nov/2019*
Entidad que organiza	*Universidad Nacional de La Matanza (UNLaM)*
URL de descarga del trabajo (especificar solo si es la descarga del trabajo; formatos pdf, e-pub, etc.)	*https://www.researchgate.net/publication/337445353_Peso_al_Nacer_de_Terneros_Aberdeen_Angus_mediante_Algoritmos_No_Supervisados*

B.5. Otras publicaciones

Autores	*Osvaldo Spósitto, Gabriel Blanco, Marcelo Levi, Julio Bossero*
Año	*2019*
Título	*Clasificación del Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos Supervisados*
Medio de Publicación	*48 JAIIO, Universidad Nacional de Salta (UNSa), Salta, Argentina* *16/Sept/2019*

C. Otros resultados. Indicar aquellos resultados pasibles de ser protegidos a través de instrumentos de propiedad intelectual, como patentes, derechos de autor, derechos de obtentor, etc. y desarrollos que no pueden ser protegidos por instrumentos de propiedad intelectual, como las tecnologías organizacionales y otros. Complete un cuadro por cada uno de estos dos tipos de productos.

C.1. Títulos de propiedad intelectual. Indicar: Tipo (marcas, patentes, modelos y diseños, la transferencia tecnológica) de desarrollo o producto, Titular, Fecha de solicitud, Fecha de otorgamiento

Tipo	Titular	Fecha de Solicitud	Fecha de Emisión

C.2. Otros desarrollos no pasibles de ser protegidos por títulos de propiedad intelectual. Indicar: Producto y Descripción.

Producto	Descripción

D. Formación de recursos humanos. Trabajos finales de graduación, tesis de grado y posgrado. Completar un cuadro por cada uno de los trabajos generados en el marco del proyecto.

D.1. Tesis de grado

Director (apellido y nombre)	Autor (apellido y nombre)	Institución	Calificación	Fecha /En curso	Título de la tesis

D.2 Trabajo Final de Especialización

Director (apellido y nombre)	Autor (apellido y nombre)	Institución	Calificación	Fecha /En curso	Título del Trabajo Final

D.2. Tesis de posgrado: Maestría

Director (apellido y nombre)	Tesista (apellido y nombre)	Institución	Calificación	Fecha /En curso	Título de la tesis

D.3. Tesis de posgrado: Doctorado

Director (apellido y nombre)	Tesista (apellido y nombre)	Institución	Calificación	Fecha /En curso	Título de la tesis

D.4. Trabajos de Posdoctorado

Director (apellido y nombre)	Posdoctorando (apellido y nombre)	Institución	Calificación	Fecha /En curso	Título del trabajo	Publicación

E. Otros recursos humanos en formación: estudiantes/ investigadores (grado/posgrado/ posdoctorado)

Apellido y nombre del Recurso Humano	Tipo	Institución	Período (desde/hasta)	Actividad asignada[2]

F. Vinculación[3]: Indicar conformación de redes, intercambio científico, etc. con otros grupos de investigación; con el ámbito productivo o con entidades públicas. Desarrolle en no más de dos (2) páginas.

G. Otra información. Incluir toda otra información que se considere pertinente.

Premiado como uno de los dos mejores Trabajos de Investigadores en el área temática de Bases de Datos del Congreso 7mo CONAIISI DIIT UNLaM – RIISIC, publicado en ISSN: 2347-0372 - ISBN 978-987-4417-73-2 ©

Título: PESO AL NACER DE TERNEROS ABERDEEN ANGUS MEDIANTE ALGORITMOS NO SUPERVISADOS
Autores: Osvaldo Spositto, Gabriel Blanco, Marcelo Levi, Patricio Macías Corral, Lorena Matteo
Institución: Departamento de Ingeniería e Investigaciones Tecnológicas – Universidad Nacional de La Matanza

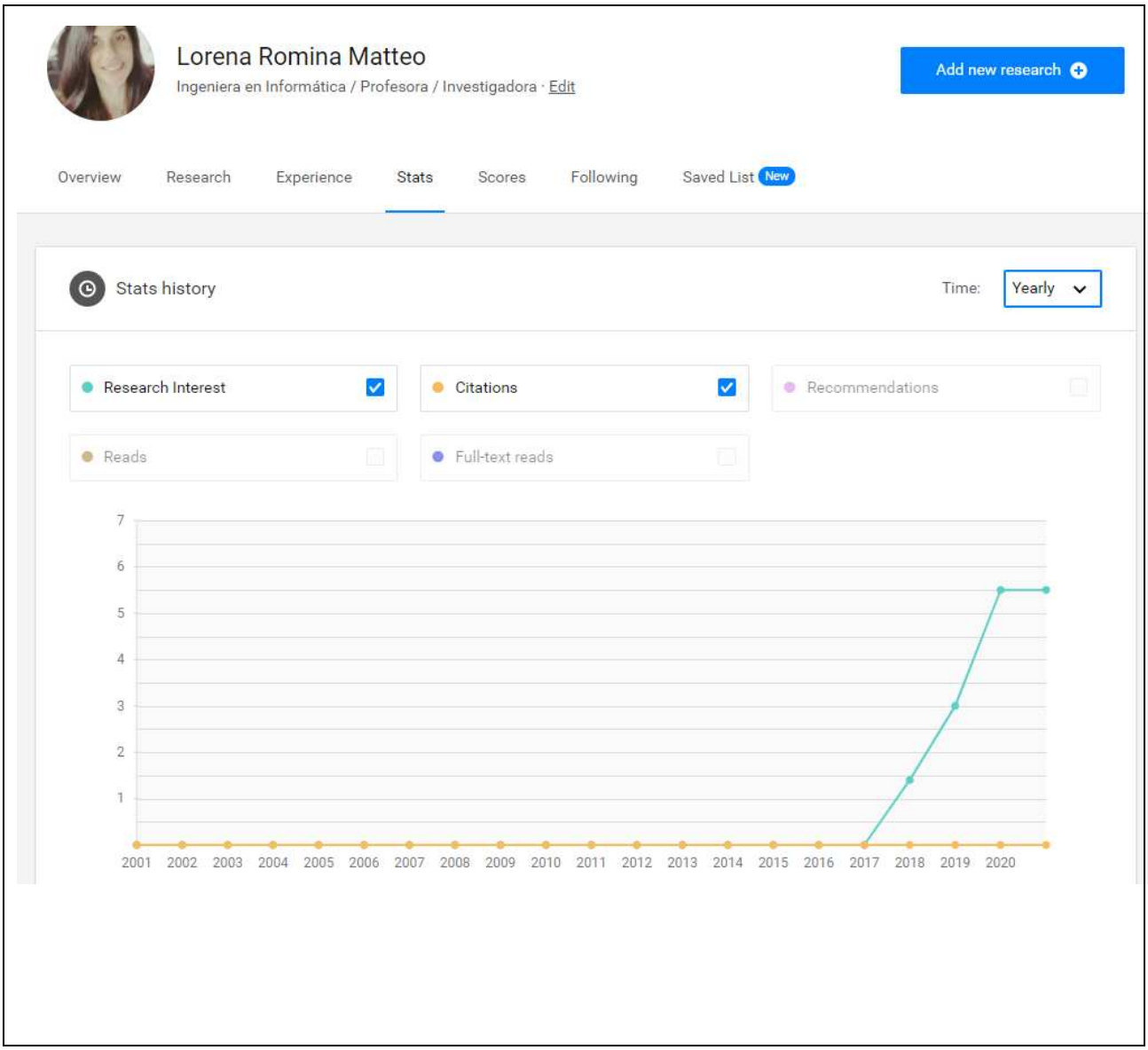
***RESEARCH GATE y GOOGLE SCHOLAR**. – Para lograr visibilidad de trabajos de la UNLaM los artículos de esta investigación se encuentran publicados en los sitios de divulgación científica como ser Google Scholar y Research Gate. En este último, las estadísticas informan cerca de 850 Reads, desde más de 5 países como ser: Colombia, México, Chile y Rusia, entre otros varios hitos alcanzados en los últimos años según los indicadores que se controlan en el sitio.*

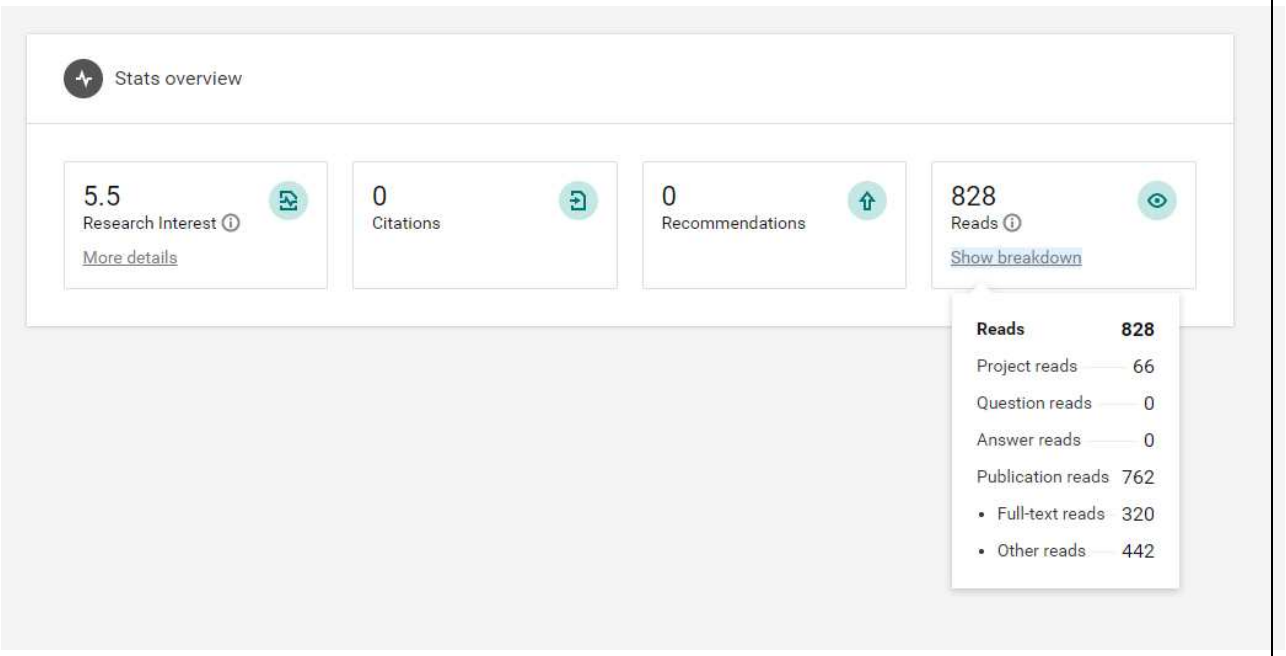
Research Gate

Lorena Romina Matteo | Stats (researchgate.net)

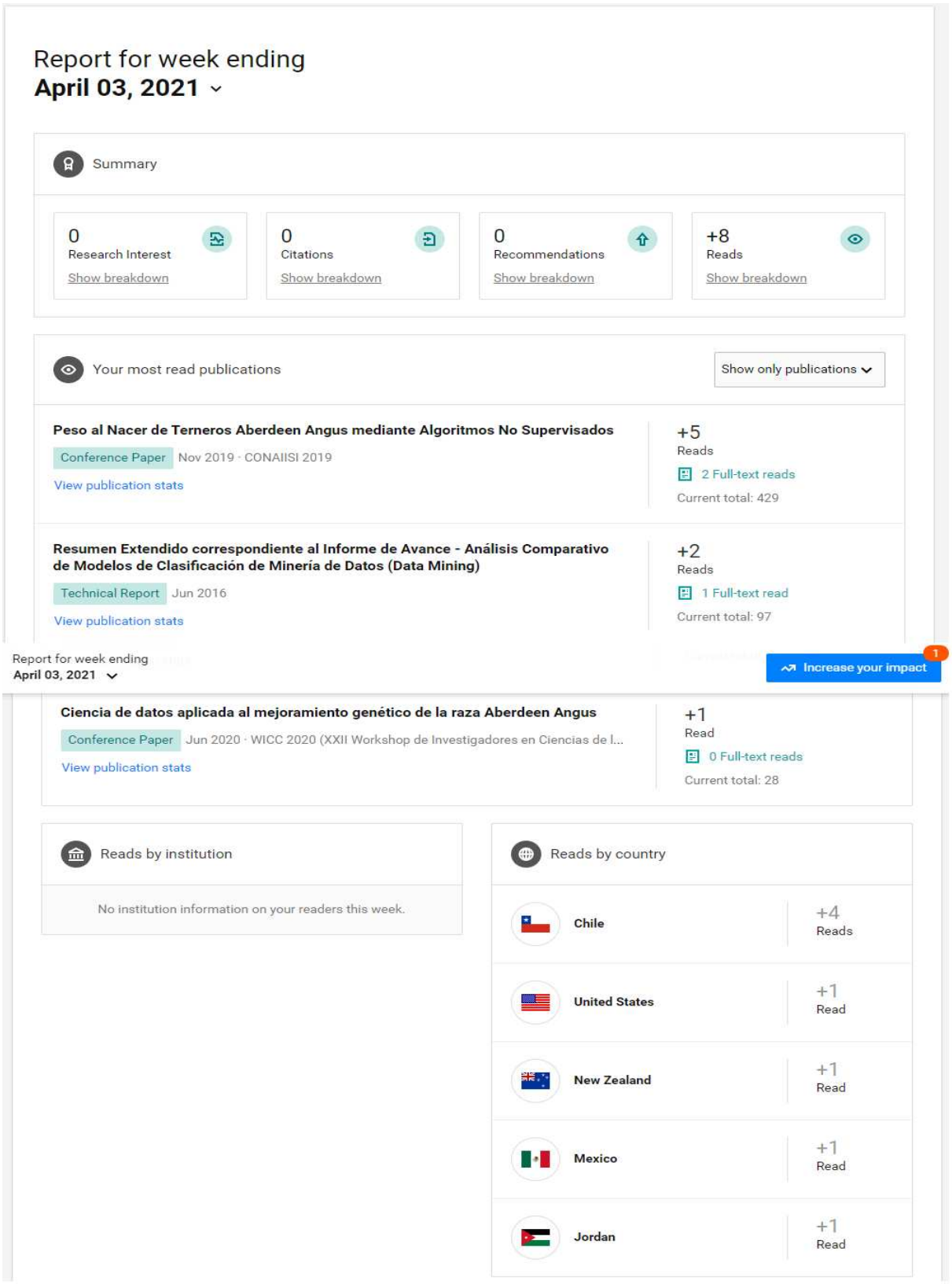
Estadísticas – Crecimiento Reads desde 2017

[2] Descripción de la/s actividad/es a cargo (máximo 30 palabras)
[3] Entendemos por acciones de "vinculación" aquellas que tienen por objetivo dar respuesta a problemas, generando la creación de productos o servicios innovadores y confeccionados "a medida" de sus contrapartes.

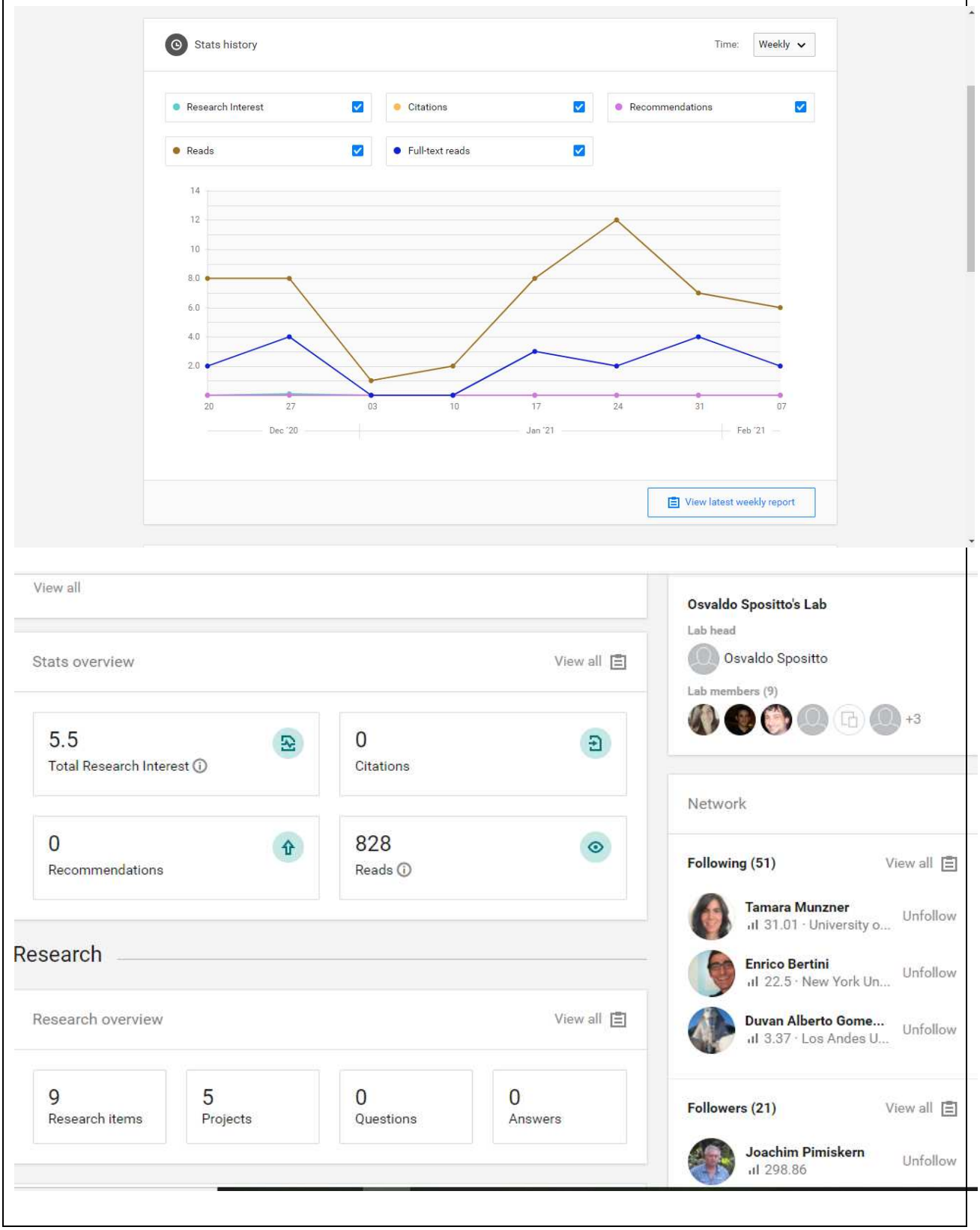

Lorena Romina Matteo
Ingeniera en Informática / Profesora / Investigadora · Edit
Add new research
Overview
Research
Experience
Stats
Scores
Following
Saved List New
Stats history
Time:
Yearly
Research Interest
Citations
Recommendations
Reads
Full-text reads
7
6
5
4
3
2
1
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020

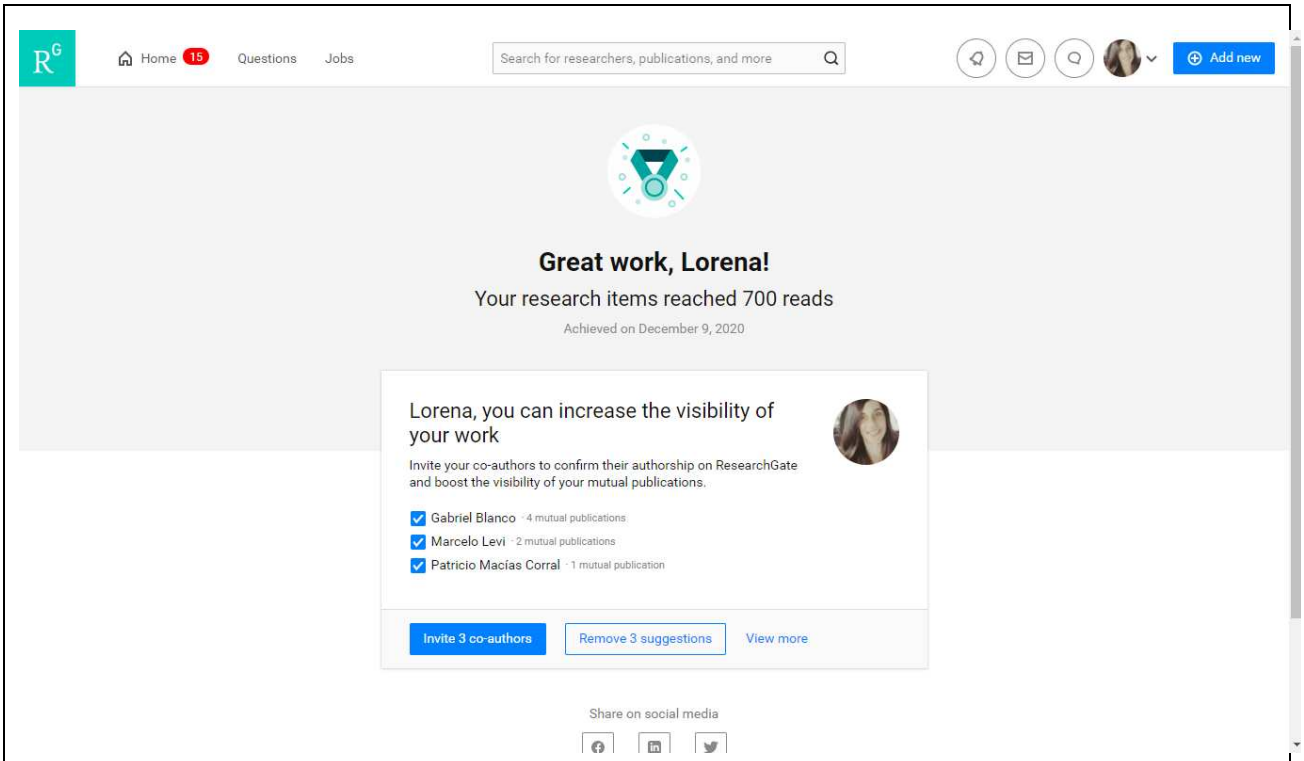


Lorena Romina Matteo | Weekly Report for 2021-01-31 (researchgate.net)



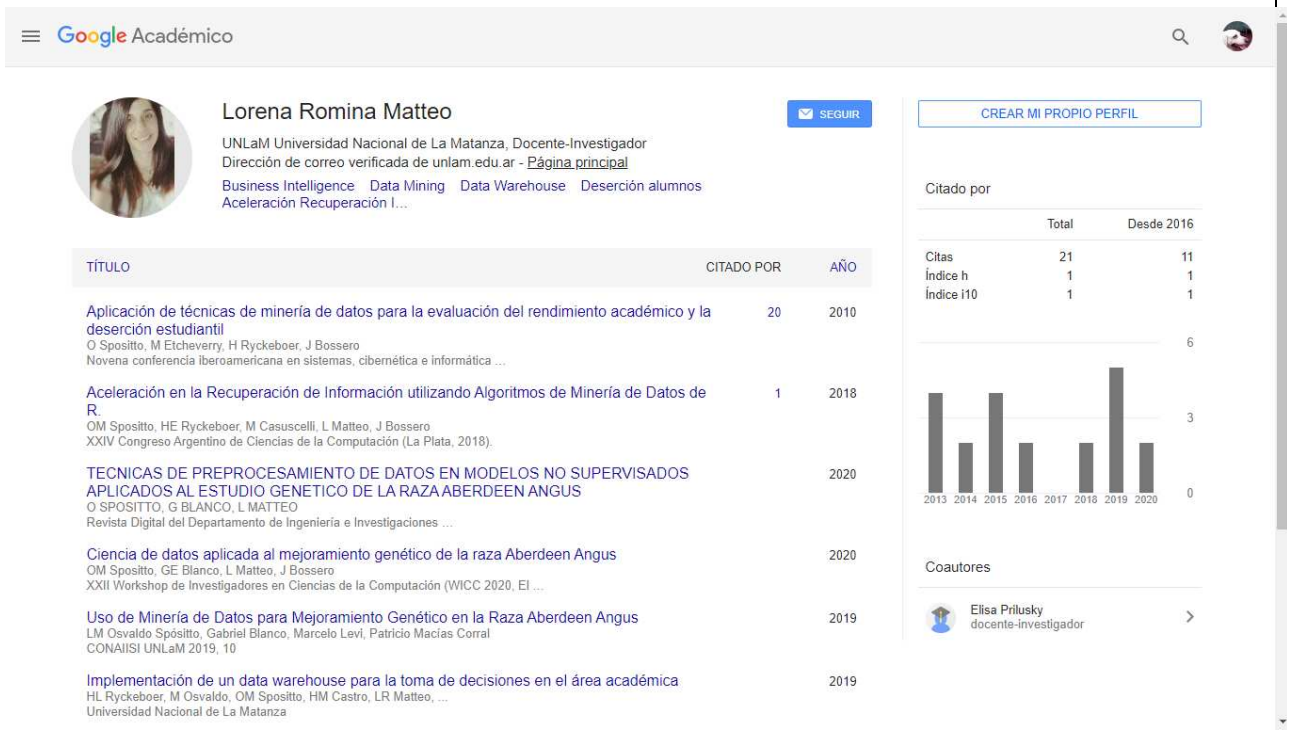
Reads última semana Enero 2021


Stats history
Time: Weekly
Research Interest
Citations
Recommendations
Reads
Full-text reads
14
12
10
8.0
6.0
4.0
2.0
20
27
03
10
17
24
31
07
Dec '20
Jan '21
Feb '21
View latest weekly report
View all
Stats overview
View all
5.5
Total Research Interest
0
Citations
0
Recommendations
828
Reads
Research
Research overview
View all
9
Research items
5
Projects
0
Questions
0
Answers
Osvaldo Spositto's Lab
Lab head
Osvaldo Spositto
Lab members (9)
+3
Network
Following (51)
View all
Tamara Munzner
31.01 · University o...
Unfollow
Enrico Bertini
22.5 · New York Un...
Unfollow
Duvan Alberto Gome...
3.37 · Los Andes U...
Unfollow
Followers (21)
View all
Joachim Pimiskern
298.86
Unfollow



Google Scholar

Lorena Romina Matteo - Google Académico



Cursos, talleres y/o seminarios de perfeccionamiento:

Algunos integrantes del equipo realizaron los siguientes cursos, talleres y/o seminarios:

- Curso de Escritura Científica I (30 hs UNLaM Vía Teams y MIeL)
- Curso de Escritura Científica II (20 hs UNLaM Vía Teams y MIeL)

- Curso Introductorio de SPSS (20 hs UNLaM Vía Teams y MIeL)
- Trayecto Formativo: La enseñanza desde la "no presencialidad" (4 encuentros UNLaM Vía Teams y MIeL)
- VIII JCC-BD&ET - Big Data UNLP 2020 (Vía Streaming YouTube)
- Webinars de MicroStrategy (Vía Zoom)
- Webinars Amazon Web Services (AWS) (Vía GoToWebinar)
- Seminario "Inteligencia artificial y ciencia de datos: potencial y desafíos para la gestión de crisis sanitarias" Ciclo de encuentros: Inteligencia artificial y ciencia de datos, potencial y desafíos para la gestión de crisis sanitarias | Argentina.gob.ar vía Zoom y YouTube, los jueves 1, 8, 15, 22 y 29 de octubre de 2020, de 16 a 18 horas.

H. Cuerpo de anexos:

- Anexo I: Copia de cada uno de los trabajos mencionados en los puntos B, C y D, y certificaciones cuando corresponda.[4]
- Anexo II:
  - FPI-013: Evaluación de alumnos integrantes. (si corresponde)
  - FPI-014: Comprobante de liquidación y rendición de viáticos. (si corresponde)
  - FPI-015: Rendición de gastos del proyecto de investigación acompañado de las hojas foliadas con los comprobantes de gastos.
  - FPI-035: Formulario de reasignación de fondos en Presupuesto.
- Anexo III: Alta patrimonial de los bienes adquiridos con presupuesto del proyecto (FPI 017)
- Nota justificando baja de integrantes del equipo de investigación.

Spositto Osvaldo

Firma y aclaración
del director del proyecto.

Lugar y fecha: San Justo 28 de febrero de 2021

- Presentar una copia impresa firmada del presente documento junto con los Anexos, y enviar todo en archivo PDF por correo electrónico a la Secretaría de Investigación Departamental. **Límite de entrega: 28 de febrero de 2020**

[4] En caso de libros, podrá presentarse una fotocopia de la primera hoja significativa o su equivalente y el índice.

Anexo I

B.1. Publicaciones en revistas

Certificado Publicación Revista ReDDI UNLaM

Certificado Publicación Vol. 5-Nro.1 ReDDI

RD

Revista Digital DIIT
Vie 23/10/2020 21:58
Para: Osvaldo Mario Spositto; Gabriel Blanco; LORENA MATTEO

pdf Técnicas de preprocesamient...
335 KB

Estimado autor, agradeciendo nuevamente su interés en publicar en la Revista Científica del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de La Matanza, le enviamos el certificado que acredita la publicación de su artículo en nuestro primer número del año (Vol.5-Nro.1).
Quedando a su disposición le envío un saludo cordial,
Ing. Andrea Vera
Coordinadora Editorial
ReDDI E-journal

-

Piense si es necesario imprimir este correo.
Todos somos responsables por el cuidado del medio ambiente

B.4. Trabajos presentados a congresos y/o seminarios

2019 – Certificado Autores y Publicación Artículo 48º JAIIO, Universidad Nacional de Salta (UNSa), Salta, Argentina

Clasificación del Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos Supervisados v3. doc

Clasificación del Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos Supervisados

Osvaldo Spositto[1], Gabriel Blanco[2], Marcelo Levi[3], Julio Bossero[4]

[1]Universidad Nacional de La Matanza. Departamento de Ingeniería e Investigaciones Tecnológicas.
spositto@unlam.edu.ar
[2]Universidad Nacional de La Matanza. Departamento de Ingeniería e Investigaciones Tecnológicas.
g2blanco@unlam.edu.ar
[3]Universidad Nacional de La Matanza. Departamento de Ingeniería e Investigaciones Tecnológicas.
mlevi@unlam.edu.ar
[4]Universidad Nacional de La Matanza. Departamento de Ingeniería e Investigaciones Tecnológicas.
jbossero@unlam.edu.ar

Resumen. Se presenta un estudio que tiene un doble propósito, por un lado conocer si a partir de los valores genéticos usados en un rodeo de cría de la raza Aberdeen Angus es posible obtener buenas predicciones sobre la clasificación del peso de los terneros al nacer, y por otro lado, determinar con qué algoritmo de minería de datos (MD) del tipo supervisado se obtienen mejores porcentajes de instancias bien clasificadas. Estos algoritmos fueron implementados en WEKA, una plataforma de software para el aprendizaje automático y la minería de datos, de libre distribución. Para el estudio se analizaron datos concretos de 360 ejemplares de progenitores proporcionados por un establecimiento ganadero de la localidad de Chascomús, Provincia de Buenos Aires, los mismos corresponden al 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.

Palabras claves: Minería de Datos, Ganadería, Peso al Nacer, Aberdeen Angus, Algoritmos Supervisados.

1 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 los 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 pasado, 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 es bastante sencilla y tiene muchas ventajas, se está aplicando desde hace bastante tiempo en el país. Como se expresa en [4], *"...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 selección, es decir la elección de padres, es la principal herramienta que poseen los criadores y productores comerciales para conseguir mejoras dentro de sus rodeos. Consecuentemente, la evaluación objetiva de los reproductores y la posterior selección de los mismos es uno de los pilares básicos para lograr los objetivos de cualquier programa genético.

La forma rápida de incorporar genética de toros probados es la Inseminación Artificial (IA) [5] y [6]. La IA puede definirse como la biotecnología para la aplicación de semen en el tracto genital de una hembra en el momento efectivo para la fecundación. Respecto al origen de la misma, existen historias indocumentadas de la obtención por los árabes de esperma a partir de yeguas servidas pertenecientes a grupos rivales, y su uso en la inseminación de sus propias yeguas. Es innegable la relevancia de esta técnica en el mejoramiento de los parámetros reproductivos y productivos de la ganadería mundial [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].

Según la Asociación Argentina de Angus, "*…que llevó un relevamiento de los ingresos al Mercado de Liniers, sobre la base de un recuento que ya supera las 200.000 cabezas de las distintas categorías que concurren a esa plaza, más del 50% son Angus (puros y "mestizos"), y considerando las distintas cruzas, la "influencia Angus" superó el 70%. Asimismo, un segundo estudio tuvo como objeto evaluar la magnitud y composición racial de la oferta de reproductores anunciada a través de las publicidades de remates de uno de los principales diarios de circulación nacional, usualmente utilizado como medio para su promoción; este relevamiento, repetido dos años, mostró que el 55/57% de los reproductores ofrecidos correspondió a animales Angus puros. Si bien en ambos casos debe tenerse en cuenta las áreas de influencia de las respectivas fuentes de información -fuera de las cuales seguramente los resultados serían diferentes-, ambos trabajos pueden ser considerados como una buena muestra, representativa de la zona ganadera de la Pampa Húmeda, que concentra más del 70% de la población bovina nacional…*"

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]. Cabe aclarar que los rasgos fenotípicos cuentan con rasgos tanto físicos como conductuales. Es importante destacar que el fenotipo no puede definirse exclusivamente como la "manifestación visible" del genotipo, pues a veces las características que se estudian no son visibles en el individuo, como es el caso de la presencia de una enzima. En donde Fenotipo= Genotipo/Ambiente [10].

Los DEP son indicadores numéricos que estiman el desempeño promedio esperado de los hijos de un determinado reproductor en relación a una base de comparación (promedio de la raza o promedio de la cabaña) [4]. Los DEP se expresan como desvíos positivos o negativos, en relación a esta base, y se obtienen de procedimientos conocidos como evaluaciones genéticas poblacionales. Son parámetros utilizados por el programa de medición del grupo Evaluación de Reproductores Angus [5] (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 DEPS. Estas fueron extraídas del Anuario Las Lilas 2017 [12]:

- 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

[5] http://www.angus.org.ar/era.php

- 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.

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 implementaciones en donde 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 minería de datos. 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 existen 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 al sistema BLUP.

Este trabajo se realiza bajo la hipótesis que si se aplica una metodología para realizar Minería de Datos (MD) a partir 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., es posible, confeccionar un modelo predictivo, que sirva de ayuda a los criadores al momento de la toma de decisiones. Y una vez obtenido el modelo, utilizarlo para realizar un análisis comparativo de la capacidad de clasificación de diferentes técnicas de MD.

2. 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[6] ambas ubicadas en la localidad de Chascomús, en la provincia de Buenos Aires. Esta Cabaña perteneciente a la Asociación Argentina de Angus[7], es uno de los caudales genéticos más importantes de la Argentina y del Mercosur, produciendo genética en cinco razas líderes entre las que se encuentran la raza Aberdeen Angus, entre otras. 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 2 (dos) reproductores: Pucará y Al-Mós. En la Figura 1 se muestran los valores de los DEPs de los mismos. Los demás datos se encuentran disponibles en el catálogo del centro de genética Cabaña Las Lilas[8].

Genética	PN	PD	AM	CM	PF	CE	AOB	GD	MAR
DEP	+0,1	+5,8	-2,8	+0,1	+17,7	+1,0	+2,10	+0,30	+0,00
PREC	0,91	0,86	0,43		0,77	0,91	0,82	0,82	0,82
Ranking							15%	30%	25%
Promedio	+0,1	+5,5	+0,5	+3,3	+16,9	+1,1	+0,40	+0,20	+0,00

Fuente: ERA 2017

a) DEPs Toro Pucará

Genética	PN	PD	AM	CM	PF	CE	AOB	GD	MAR
DEP	+0,4	+1,9	+1,5	+2,5	+8,3	+0,4	-2,70	+0,20	-0,10
PREC	0,82	0,70	0,55		0,60	0,82	0,73	0,74	0,73
Ranking			35%						
Promedio	+0,1	+5,5	+0,5	+3,3	+16,9	+1,1	+0,40	+0,20	+0,00

Fuente: ERA 2017

b) DEPs Toro Al-Mós

Figura 1. Valores correspondientes a los 2 reproductores utilizados en la investigación.

[6] http://laslilas.com

[7] https://www.angus.org.ar/

[8] En [12] se encuentra la infografía de los reproductores Al-Mós y Pucará en las páginas 32 y 39 respectivamente.

2.1 La Variable Dependiente: Peso al Nacer

Tomando como referencia la información que se encuentra en el programa genético de la Cabaña La Mariana [16], 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. El PN está positivamente correlacionado con el Peso al Destete (PD), al año y Peso Final (PF). Es decir, que en general, seleccionar por bajo PN va en detrimento del PD y PF.

El objetivo de la modelización supervisada consiste en explicar el comportamiento de una variable a partir del conocimiento de otras. Subyacente al concepto de modelización está la idea de que una variable tiene una cierta variabilidad que está relacionada con el comportamiento de otras variables. Por tal motivo para esta experimentación se tomaron los PN de las progenies del rodeo del establecimiento La Doce, y según la explicación del criador, se tomó para el PN, un peso promedio de 38 kilogramos. Por tal motivo, a los pesos mayores e iguales a 38 kg., se los clasificó como Altos y al resto como Bajos. Este atributo será el atributo "*Clase*" o "*Etiqueta*" [17], que se utilizará para entrenar los algoritmos de clasificación.

2.2 Metodología

Entre las distintas metodologías para llevar a cabo un proyecto de MD existente [18], se optó por Cross Industry Standard Process for Data Mining (CRISP-DM) [19] y [20]. La misma cuenta de seis etapas, como se muestra en la Figura 2. Además de ello, esta tecnología interrelaciona las diferentes fases del proceso entre sí, de tal manera que se consolida un proceso iterativo y recíproco.

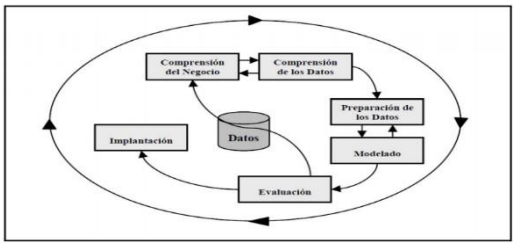


Figura 2. Valores Modelo de proceso CRISP–DM [19].

A continuación, se desarrolla una breve descripción de las fases utilizadas para la construcción del modelo.

2.2.1 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 minería de datos 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.

2.2.2 Comprensión de los datos

Esta fase comienza con la recolección de datos iniciales y continuó con las actividades que permitieron familiarizarse con los datos, 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.

2.2.3 Recopilación 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 Tabla 1.

Para optimizar los algoritmos supervisados propuestos para este proyecto, se normalizaron las variables de entrada al algoritmo. 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[19], que se define como:

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

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

Nomenclatura	Tipo de dato	Descripción
Peso Adulto	*Numérico*	Peso real del padre
PAN	*Numérico*	Peso al nacer
PAD	*Numérico*	Peso al destete
PAF	*Numérico*	Peso final
Circ Escrotal	*Numérico*	Circunferencia escrotal
FRAME	*Numérico*	Altura del animal
Certificado	*Numérico*	Edad promedio de los vientres primerizos
PN	*Numérico*	DEP del Toro Progenitor
PD	*Numérico*	
AM	*Numérico*	
CM	*Numérico*	
PF	*Numérico*	
CE	*Numérico*	
AOB	*Numérico*	
GD	*Numérico*	
MAR	*Numérico*	
Peso al nacer	*Numérico*	Datos de la vaca
Peso al destete	*Numérico*	
Cantidad nacimientos	*Numérico*	
Cantidad abortos	*Numérico*	
Baja antes del destete	*Numérico*	
Edad en meses	*Numérico*	
Certificado	*Numérico*	DEP del Toro Progenitor de la vaca
CE del padre	*Numérico*	
PN	*Numérico*	
PD	*Numérico*	
AM	*Numérico*	
CM	*Numérico*	
PF	*Numérico*	
CE	*Numérico*	
AOB	*Numérico*	
GD	*Numérico*	
MAR	*Numérico*	
Peso_Nac	*Texto*	Atributo Clase

2.2.4 Modelado

Se utilizó el software WEKA[9] (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 del tipo supervisado.

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

Las técnicas utilizadas para realizar la clasificación fueron del tipo Supervisadas o Predictivas, es decir, aquellas orientadas a predecir los valores de salida de una variable dependiente, en particular, para este estudio el PN. La clasificación de esta variable se realizó a partir de un conjunto de variables predictores, es decir que induce una relación entre un atributo de interés y las otras variables [17]. A continuación, se describen brevemente las técnicas que se compararon en este trabajo.

- Árboles de Decisión (AD): como su nombre lo indica es una estructura que se forma por las bifurcaciones en cada una de las decisiones, descubriendo reglas. En WEKA se lo conoce como algoritmo J48 [21], y se usaron para la clasificación del PN (alto / bajo). El mencionado algoritmo es una implementación libre en Java del algoritmo C4.5, que utiliza el concepto de entropía de la información para la selección de variables que mejor clasifiquen a la variable (clase) estudiada. En la Figura 3 se ve la representación gráfica que entrega WEKA, luego de que el algoritmo se ejecutó.

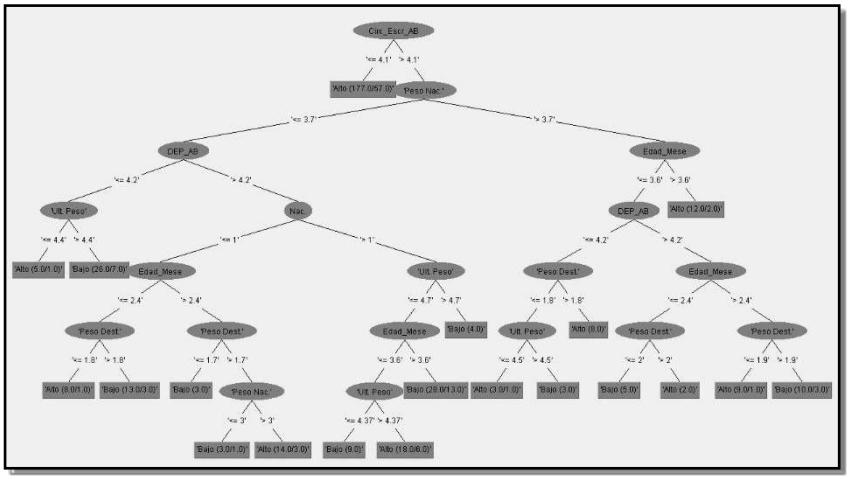


Figura 2. Visualización del Árbol de Decisión obtenido.

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

· Redes Neuronales Artificiales (RNA): Esta implementación imita el funcionamiento interno de las neuronas humanas [17]. En general, aunque pueden usarse muchos tipos de RNA para clasificación, en este caso se optó por una red multicapa feed-forward o perceptrones multicapa (MLPs) que son los clasificadores basados en redes neuronales más ampliamente estudiados y usados. [22], [23] y [24]. Este algoritmo es entrenado para realizar conexiones entre los valores de entrada y salida, aprendiendo de su error de pronóstico.

• Máquinas de Soporte Vectorial (MSV o SVM, del inglés Support Vector Machine), buscan el límite que separa las clases con el mayor margen posible [25], [26], como se observa en la Figura 3. Cuando no se pueden separar bien las dos clases, los algoritmos buscan el mejor límite que pueden. Las MSV hacen esto, sólo con una línea recta (usa un kernel lineal) y gracias a que hace esta aproximación lineal, se puede ejecutar con bastante rapidez. Ver Figura 4.

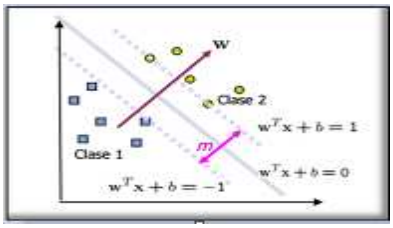


Figura 3. Hiperplano con separación óptima. Linealmente separable [26].

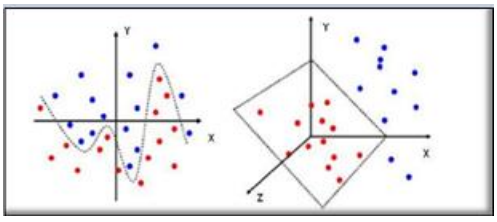


Figura 4. Transformación mediante la función Kernel para separabilidad no lineal. [25].

3. Resultados

Este apartado coincide con la fase Evaluación de la metodología CRISP-DM. En esta etapa, una vez construido el modelo, fue introducido en WEKA. Una vez que ingresa el archivo en formato .CSV (del inglés comma-separated values), que contiene los datos, 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 [21]. En la siguiente figura, se puede observar el diseño del mismo.

```
@relation 'INVCAMPO_ALTO-BAJO_OTRO TORO_2_NORMALIZANDO-csv'

@attribute PAN numeric @attribute PAD numeric @attribute PAF numeric @attribute PESO numeric @attribute C_Escro numeric
@attribute FRAME numeric @attribute Certiv numeric @attribute PN numeric @attribute PD numeric @attribute AM numeric
@attribute CM numeric @attribute PF numeric @attribute CE numeric @attribute AOB numeric @attribute GD numeric
@attribute MAR numeric @attribute Edad_Mese numeric @attribute 'Peso Nac.' numeric @attribute 'Peso Dest.' numeric
@attribute 'Ult. Peso' numeric @attribute Nac. numeric @attribute Abortos numeric @attribute Cesarias numeric
@attribute 'Mtes. Antes del destete' numeric @attribute Circ_Escr_AB numeric @attribute DEP_AB numeric @attribute Certiv_AB numeric
@attribute PN_AB numeric  @attribute PD_AB numeric  @attribute AM_AB numeric @attribute CM_AB numeric
@attribute PF_AB numeric @attribute CE_AB numeric @attribute AOB_AB numeric @attribute GD_AB numeric @attribute MAR_AB numeric
@attribute P_Nacer {Alto,Bajo}

@data
0,0,1,1,1,0,1,1,1,1,1,1,1,1,1,1,0.8,0.666667,0.916667,0.8,5,1,0,0,0,0,0,0.8,0,0.486486,0,0,0,0,0.115385,1,Alto
0,0,1,1,1,0,1,1,1,1,1,1,1,1,1,1,0.666667,0.916667,0.8,5,1,0,0,0,0,0,0.8,0,0.486486,0,0,0,0,0.115385,1,Alto
0,0,1,1,1,0,1,1,1,1,1,1,1,1,1,1,0.8,0.444444,0.5,0.933333,5,0,0,0,0,0,0,0.8,0,0.486486,0,0,0,0,0.115385,1,Bajo
```

Figura 5. Formato del archivo. arff empleado en el estudio.

Una vez cargados los datos, se procedió a entrenar los distintos algoritmos con el mismo lote de datos. En la pestaña Classify (Clasificación) (Figura 6), 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**.

En la misma ventana, podemos elegir las opciones de test, es decir, la manera de computar el porcentaje esperado de aciertos (en clasificación), o el error cuadrático medio (entre otros, en regresión). Estas opciones que dispone el producto son:

• Use training set: Es el usado para este trabajo. Se usa para hacer el test el mismo conjunto que el de entrenamiento. Este es el elegido para esta experimentación.
• Supplied test set: Si se tiene un fichero con datos de test distintos a los de entrenamiento, aquí es donde podemos seleccionarlo.
• Cross Validation: Se calcula el porcentaje de aciertos esperado haciendo una validación cruzada de k hojas (k por omisión es de = a 10).
• 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 el test.

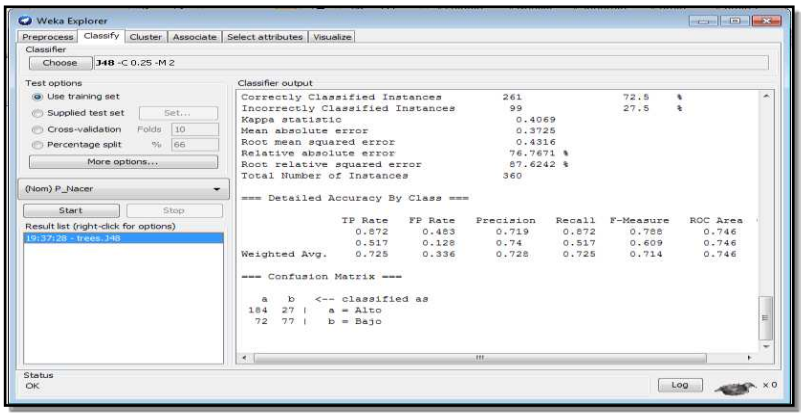


Figura 6. Clasificador obtenido por el algoritmo J48.

Para la evaluación de los clasificadores en este trabajo se utilizó la Matriz de Confusión (MC) y el análisis o curvas ROC (acrónimo de Receiver Operating Characteristic), que entrega WEKA para cada clasificador. En Hernández Orellano [16] se encuentran desarrollados cada uno de los conceptos con más detalle. A modo de resumen, una matriz de confusión muestra la clasificación de las instancias. Da una información muy útil porque no sólo refleja los errores producidos sino también informa del tipo de éstos. Dicha matriz tiene la siguiente estructura (Ver Figura 7).

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 7. Matriz de Confusión[10].

Donde:

- VP es la cantidad de positivos que fueron clasificados correctamente como positivos por el modelo.
- VN es la cantidad de negativos que fueron clasificados correctamente como negativos por el modelo.
- FN es la cantidad de positivos que fueron clasificados incorrectamente como negativos.
- FP es la cantidad de negativos que fueron clasificados incorrectamente como positivos.

De estos valores se definen dos métricas asociadas: sensibilidad y especificidad:

- La sensibilidad indica la capacidad de nuestro clasificador para dar como casos positivos los casos que realmente lo son; proporción de PN altos correctamente identificados.
- La especificidad indica la capacidad de nuestro estimador para dar como casos negativos los casos que realmente lo sean; en este caso, proporción de PN bajos correctamente identificados como tal.

Por otro lado, la curva ROC es una herramienta estadística utilizada en el análisis de los clasificadores, determinando la capacidad discriminante de una prueba diagnóstica dicotómica. La curva es el gráfico resultante de representar, para cada valor umbral, las medidas de sensibilidad y especificidad de la prueba diagnóstica.

En la Tabla2, se muestra el resumen del análisis del desempeño de los algoritmos AD, RNA y MSV. Esta tabla revela que la precisión de dichos algoritmos es diferente. De esta forma se establece que el algoritmo AD tiene mejor precisión, es decir, la mejor probabilidad de discriminar correctamente, debido a que el valor su media muestra es mayor: 72.5%. Cabe resaltar que el indicador precisión es el más importante para establecer el desempeño de un algoritmo de MD.

Tabla 2. Tabla de resultados de las pruebas de clasificación.

[10] https://rpubs.com/chzelada/275494

Porcentajes de Instancias Clasificadas

RNA	MVS	Árbol de Decisión
Correctas: 63.889 % Incorrectas: 36.111%	Correctas: 60.278 % Incorrectas: 39.722 %	Correctas: 72.5 % Incorrectas: 27.5 %

Matriz de Confusión

a	b	Clasificación
146	65	a-Alto
65	84	b-Bajo

a	b	Clasificación
173	38	a-Alto
103	44	b-Bajo

a	b	Clasificación
184	27	a- Alto
72	77	b- Bajo

Curva ROC

0.68	0.558	0.746

Por último, en la Figura 7, se muestra el área bajo la curva ROC (AUC). Esta área puede interpretarse como la probabilidad de que ante una instancia nueva de datos, la prueba los clasifique correctamente. Su rango de valores va desde 0, 5, siendo este valor el correspondiente a una prueba sin capacidad discriminante, hasta 1, que es cuando los dos grupos están perfectamente diferenciados por la prueba. Por tanto, se puede decir que cuanto mayor sea el AUC mejor será la prueba. El Algoritmo AD dio, también, el mejor resultado: 0.746.

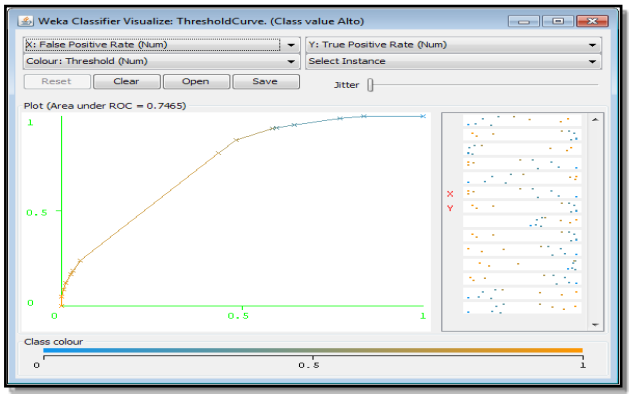


Figura 7. Visualización de la curva ROC del AD.

5. Conclusiones y discusión

Se encontró que el modelo propuesto por esta memoria, conserva una precisión (proporción de instancias clasificadas correctamente) aceptable en el caso del algoritmo AD, con un porcentaje levemente superior al 70%. Mientras que la RNA y la MVS, también con configuraciones por defecto, arrojaron porcentajes notoriamente inferiores, alrededor del 60%.

5.1 Discusión

En este primer trabajo que se llevó a cabo, los resultados globales marcaron una diferencia (en la capacidad de discriminación) entre los modelos propuestos, pero puede ser que, con diferentes configuraciones de sus parámetros, estos valores se reviertan. Por ejemplo, haber utilizado diferentes arquitecturas en las RNA o cambiar de Kernel en las MSV.

Por otro parte, la justificación del bajo porcentaje de instancias clasificadas correctamente puede argumentarse en las siguientes razones:

- No se contó con mayor cantidad de ejemplare. Se propondrá realizar esta investigación en otros rodeos.
- Se utilizaron pocos reproductores. Al tener mayor cantidad de ellos aumentara los valores de sus DEPs.
- Incrementar el número de variables, seguramente mejoraría el rendimiento de los clasificadores, pudiendo agregarse variables tales como: el tipo de alimentación de los animales, el factor climático, etc.
- Se podría experimentar con otros algoritmos Supervisados y hasta realizar pruebas con métodos del tipo No Supervisados.
- Por último, utilizar diferente herramienta computacional, por ejemplo, probar con software más sofisticados como Matlab[11], SPSS[12], etc.

Referencias

[11] https://matlabacademy.mathworks.com/es
[12] https://spss.softonic.com/

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. Disponible en: http://www.produccion-animal.com.ar/genetica_seleccion_cruzamientos/bovinos_en _general/24-Seleccion_genetica.pdf. Último acceso: 25/06/2019.
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: 25/06/2019.
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: 25/06/2019.
4. 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: 25/06/2019.
5. "Curso Teórico Práctico de Inseminación Artificial en Bovinos". Disponible en: https://www.engormix.com/ganaderia-carne/articulos/inseminacion-artificial-en-bovinos-t26957.htm. Último acceso: 25/06/2019.
6. 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: 25/06/2019.
7. Cavestany, D. y Méndez J. 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: 25/06/2019.
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: 25/06/2019.
9. "Historia y desarrollo". http://www.angus.org.ar/historia.php. Último acceso: 25/06/2019.
10. "Interacción entre genotipo y ambiente. Relación entre genotipo y entorno, correlación entre ambiente y genes. Disponible en: https://www.fenotipo.com/interaccin_entre_genotipo_y_ ambiente. Último acceso: 25/06/2019.
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: 25/06/2019.
12. "Cómo interpretar la evaluación genética". Anuario Las Lilas 2018-2019. Cabaña Las Lilas. Centro de Genética. Pág. 107. Disponible en: http://laslilas.com/pdf/Anuario-Genetica-2017.pdf. Último acceso: 25/06/2019.
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: 25/06/2019.
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: 25/06/2019.
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. "Programa Genético" Disponible en: http://www.estancialamariana.com/programa_ genetico.html. Último acceso: 25/06/2019.
17. Hernández Orallo, J. y otros. "Introducción a la minería de datos". Editorial: Pearson. Edición: I. Año 2004
18. 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: 25/06/2019
19. 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 Último acceso: 25/06/2019.
20. Gallardo Arancibia, J. Metodología para la definición de requisitos en proyectos de data mining. (2009). Tesis (Doctoral), Facultad de Informática (UPM). Disponible en: http://oa.upm.es/1946. Último acceso: 25/06/2019.
21. 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.
22. Bossero, Julio; Edwards, Diego J.; Pérez, Silvia N. Predicción del riesgo de abandono universitario utilizando métodos supervisados. Trabajo presentado en el Workshop de la V Jornadas Naciona-les y I Latinoamericanas de Ingreso y Permanencia en Carreras Científico – Tecnológicas. Facultad Regional Bahía Blanca. U. T. N. Bahía Blanca. Mayo de 2016. IPECyT 2016. Disponible en: http://www.edutecne.utn.edu.ar/ipecyt-2016/00-IPECy T_2016.pdf. Último acceso: 25/06/2019.
23. Spositto, O. y Bossero, J. Comparación de Algoritmos de Aprendizaje Supervisado para la obtención de perfiles de alumnos desertores. Trabajo presentado en el "Workshop del IV Congreso Nacional de In-geniería Ingeniería en Informática/Sistemas de Información. Publicación on line - ISSN 2347-0372. CONAIISI 2016. Disponible en: http://www.ucasal.edu.ar/ conaiisi2016/book/memorias.html. Último acceso: 25/06/2019.
24. Spositto, O., Blanco, G., Giuliano, M., Fernandez, L. y Bossero, J. Modelos de minería de datos para el diagnóstico de enfermedad de Parkinson mediante el análisis de voz. Trabajo presentado en el "Workshop del V Congreso Nacional de Ingeniería en Informática/Sistemas de Información. Publicación on line - ISSN. CONAIISI 2017. Santa Fe. Disponible en:. http://conaiisi2017.frsf.utn.edu.ar/index.php /memorias. Último acceso: 25/06/2019.
25. Bernal-de Lázaro,J., Prieto Moreno, A., Orestes Llanes, S. y García Moreno, S. Estudio comparativo de clasificadores empleados en el diagnóstico de fallos de sistemas industriales. (2011). Ing. Mecánica. Vol. 14. No. 2, mayo-agosto, 2011, p. 87-98 ISSN 1815-5944. Disponible en: http://scielo.sld.cu/pdf/im/v14n2/im01211.pdf. Último acceso: 25/06/2019.
26. Farías Concha, M. Máquinas Vectoriales híbridas para clasificar accidentes de tránsito en la región metropolitana. (2011). Pontificia Universidad Católica de Valparaíso. Disponible en: http://opac.pucv.cl/pucv_txt/txt-9500/UCF9980_01.pdf. Último acceso: 25/06/2019.

Certificado Autores y Publicación Artículo CONAIISI 2019, UNLaM, San Justo, Argentina

Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos No Supervisados_v9 - Final.docx

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]

[1] Universidad Nacional de La Matanza. Departamento de Ingeniería e Investigaciones Tecnológicas, spositto@unlam.edu.ar

[2] Universidad Nacional de La Matanza. Departamento de Ingeniería e Investigaciones Tecnológicas, g2blanco@unlam.edu.ar

[3] Universidad Nacional de La Matanza. Departamento de Ingeniería e Investigaciones Tecnológicas, mlevi@unlam.edu.ar

[4] Universidad Nacional de La Matanza. Departamento de Ingeniería e Investigaciones Tecnológicas, pmacias@unlam.edu.ar

[5] Universidad Nacional de La Matanza. Departamento de Ingeniería e Investigaciones Tecnológicas, 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[13] (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[14]:

- 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

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

temática llevados a cabo en países como Brasil, Colombia y España. Estos describen 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[15] ambas ubicadas en la localidad de Chascomús, en la provincia de Buenos Aires. Esta Cabaña perteneciente a la Asociación Argentina de Angus[16], 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.

[15] http://laslilas.com
[16] https://www.angus.org.ar/

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 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[17], 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.

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

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 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:

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

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[18] (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 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[19], 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.

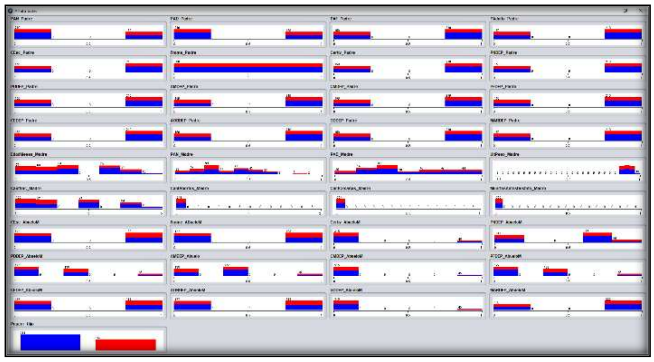


Figura 1. Opción Visualize All una vez cargado el archivo. arff empleado en el estudio.

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 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.

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

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

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.

5.2. 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:

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

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 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

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:

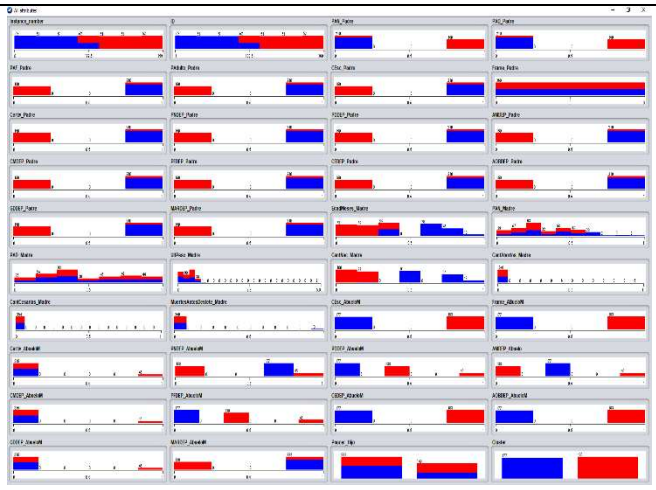


Figura 5. Opción Visualize All una vez cargado el archivo. arff que contenía la prueba [*3]: "Selección Prueba 3" (Classification Reglas PART)

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[20], SPSS[21], etc.

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

[20] https://matlabacademy.mathworks.com/es
[21] https://spss.softonic.com/

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

CONAIISI

VII Congreso Nacional de Ingeniería Informática - Sistemas de Información

2019

San Justo, 5 de diciembre de 2019

Se certifica que Osvaldo Spositto, Gabriel Blanco, Marcelo Levi, Patricio Macias Corral, Lorena Romina Matteo han participado como Autores del artículo 278 *"Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos No Supervisados"*, aceptado en el VII Congreso Nacional de Ingeniería Informática – Sistemas de Información, CONAIISI 2019, realizado los días 14 y 15 de noviembre en la Universidad Nacional de La Matanza.

Ing. Claudio D'Amico
Coord. Gral. CONAIISI

Dr. Carlos Neil
Coordinador RIISIC

Mg. Jorge Eterovic
Decano DIIT

UNLaM confedi DIIT RIISIC

B.4. Trabajos presentados a congresos y/o seminarios

Certificado Autores y Publicación Poster y Artículo WICC 2020

UNPA RedUNCI WICC 2020

Universidad Nacional de la Patagonia Austral
XXII WORKSHOP DE INVESTIGADORES EN CIENCIAS DE LA COMPUTACION 2020

Se certifica que **OSVALDO MARIO SPOSITTO (UNLAM)** ha participado en calidad de autor del artículo **CIENCIA DE DATOS APLICADA AL MEJORAMIENTO GENÉTICO DE LA RAZA ABERDEEN ANGUS (12780 - BDMD)** aceptado en el XXII WORKSHOP DE INVESTIGADORES EN CIENCIAS DE LA COMPUTACIÓN – WICC 2020, organizado por la Universidad Nacional de la Patagonia Austral - Junio 2020.

CERTIFICADO Nº 989 /2020 /UNPA

Lic. Patricia Pesado
Coordinadora
RedUNCI

Ing. Hugo Santos ROJAS
Rector
UNPA

UNPA RedUNCI WICC 2020

Universidad Nacional de la Patagonia Austral
XXII WORKSHOP DE INVESTIGADORES EN CIENCIAS DE LA COMPUTACION 2020

Se certifica que **GABRIEL ESTEBAN BLANCO (UNLAM)** ha participado en calidad de autor del artículo **CIENCIA DE DATOS APLICADA AL MEJORAMIENTO GENÉTICO DE LA RAZA ABERDEEN ANGUS (12780 - BDMD)** aceptado en el XXII WORKSHOP DE INVESTIGADORES EN CIENCIAS DE LA COMPUTACIÓN – WICC 2020, organizado por la Universidad Nacional de la Patagonia Austral - Junio 2020.

CERTIFICADO Nº 990 /2020 /UNPA

Lic. Patricia Pesado
Coordinadora
RedUNCI

Ing. Hugo Santos ROJAS
Rector
UNPA

UNPA RedUNCI WICC 2020

Universidad Nacional de la Patagonia Austral
XXII WORKSHOP DE INVESTIGADORES EN CIENCIAS DE LA COMPUTACION 2020

Se certifica que **JULIO C. BOSSERO (UNLAM)** ha participado en calidad de autor del artículo **CIENCIA DE DATOS APLICADA AL MEJORAMIENTO GENÉTICO DE LA RAZA ABERDEEN ANGUS (12780 - BDMD)** aceptado en el XXII WORKSHOP DE INVESTIGADORES EN CIENCIAS DE LA COMPUTACIÓN – WICC 2020, organizado por la Universidad Nacional de la Patagonia Austral - Junio 2020.

CERTIFICADO Nº 993 /2020 /UNPA

Lic. Patricia Pesado
Coordinadora
RedUNCI

Ing. Hugo Santos ROJAS
Rector
UNPA

UNPA RedUNCI WICC 2020

Universidad Nacional de la Patagonia Austral
XXII WORKSHOP DE INVESTIGADORES EN CIENCIAS DE LA COMPUTACION 2020

Se certifica que **LORENA MATTEO (UNLAM)** ha participado en calidad de autor del artículo **CIENCIA DE DATOS APLICADA AL MEJORAMIENTO GENÉTICO DE LA RAZA ABERDEEN ANGUS (12780 - BDMD)** aceptado en el XXII WORKSHOP DE INVESTIGADORES EN CIENCIAS DE LA COMPUTACIÓN – WICC 2020, organizado por la Universidad Nacional de la Patagonia Austral - Junio 2020.

Lic. Patricia Pesado
Coordinadora
RedUNCI

Ing. Hugo Santos ROJAS
Rector
UNPA

CERTIFICADO Nº 991 /2020 /UNPA

SEDICI

REPOSITORIO INSTITUCIONAL DE LA UNLP

To exit full screen, move mouse to top of screen or press F11

UNIVERSIDAD NACIONAL DE LA PLATA

Home Search Upload Material Institutional Frequentes questions Contact

Search resources

Search among more than 104600 resources available in the repository

Upload resources

Upload your works to SEDICI to increase its visibility and improve its impact

¿Cómo cargo mis trabajos?

Ayuda para Investigadores

Red de Universidades con Carreras en Informática (RedUNCI) → Eventos → WICC → WICC 2020

Login

Ciencia de datos aplicada al mejoramiento genético de la raza Aberdeen Angus

Authors: Sposito, Osvaldo Mario | Blanco, Gabriel Esteban | Matteo, Lorena | Bossero, Julio 2020

Document type: Objeto de conferencia

Like Share Twittear

Abstract

La Diferencia Esperada entre Progenies (DEP) es un indicador numérico que predice la calidad genética de las futuras crías de un toro o una vaca respecto de una base de comparación. Este es un valor genético que proporciona la mejor manera de comparar reproductores por la producción esperada en sus descendencias. En este trabajo estudiamos el uso de técnicas de Minería de Datos, del tipo Supervisadas y No Supervisadas, para 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. El objetivo es brindar una herramienta complementaria para que un criador ganadero pueda seleccionar mejor los reproductores que al ser apareados con sus vientres, produzcan progenies superiores. En estos primeros estudios emplearon datos provenientes de 360 animales proporcionados por un establecimiento ganadero de la provincia de Buenos Aires. Se espera que a partir de los modelos aprendidos por los algoritmos se pueda extraer información preliminar sobre el valor genético de un animal, que pueda resultar de gran utilidad en el sector ganadero en la toma de decisiones en un programa de mejoramiento genético.

Notes

Eje: Base de Datos y Minería de Datos.

General information

Exposure date: mayo 2020

Document language: Spanish

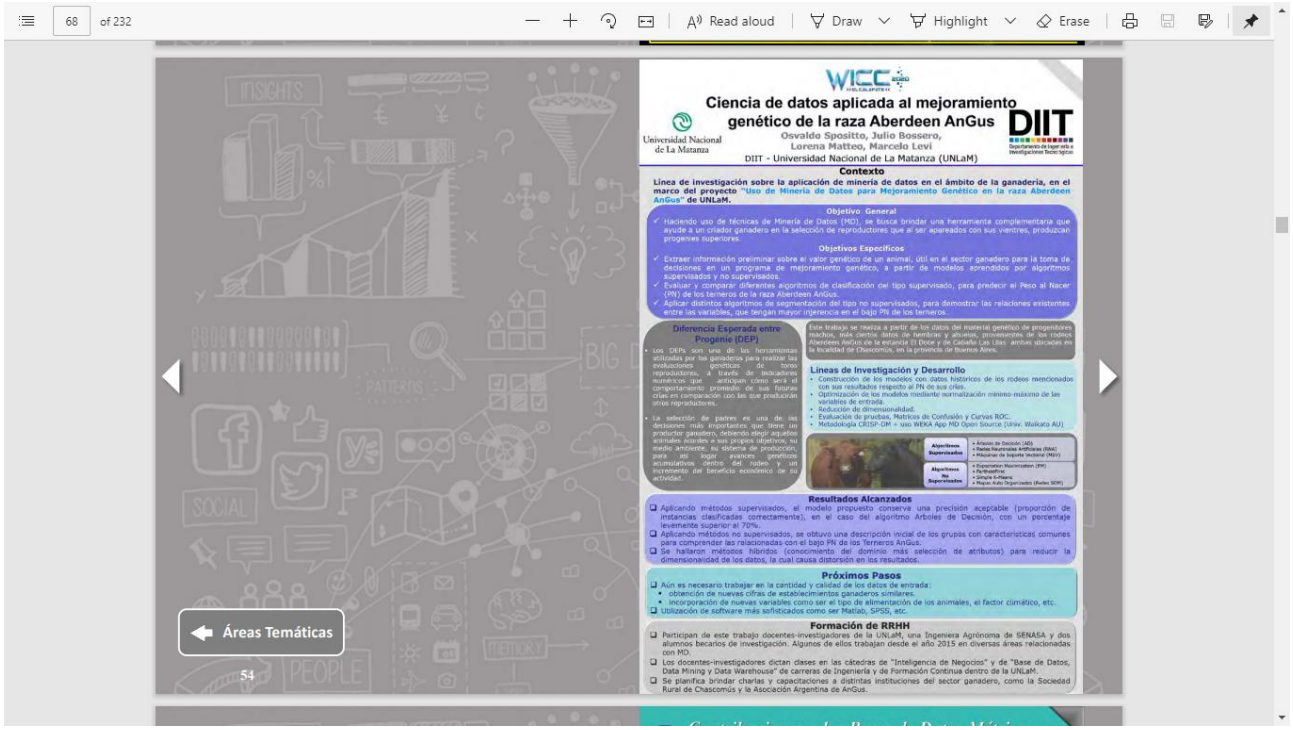
B.3. Capítulos de libros

Tapa Libro de Actas Poster WICC 2020

Índice Poster WICC 2020

	BASE DE DATOS Y MINERIA DE DATOS	
12930	Análisis cienciométrico de la producción en investigación científica y tecnológica	58
12774	Análisis de relaciones intra-institucionales e interdisciplinares de una universidad a partir de la producción registrada en Microsoft Academic: el caso de la Universidad Nacional de La Plata	59
12926	Aplicación de Técnicas Descriptivas de Minería de Textos sobre Contenido Digital	52
	Realizando Análisis Inteligente	
12767	Avances en el proyecto de Análisis y elaboración de datos para el desarrollo de un sistema de indicadores de ayuda social	46
12908	Bases de Datos Espaciales y Espacio Temporales	41
12712	Bioingeniería Informática Aplicada a la predicción de enfermedades cardiológicas y su implementación en el Hospital Delicia Concepción Masvernat de la Ciudad de Concordia, Provincia de Entre Ríos	45
12829	Búsqueda y Recopilación de Información sobre Legislación referida a Residuos Informáticos	50
12755	Búsquedas Selectivas sobre Flujos de Documentos	40
12780	Ciencia de datos aplicada al mejoramiento genético de la raza Aberdeen Angus	54
12900	Cluster para Aprendizaje y Práctica de BigData y Servicios de Learning Analytics	51
12907	Contribuciones a las Bases de Datos Métricas	55
12914	Estrategias de Desambiguación de perfiles y similitud temática para un Metabuscador de las Ciencias de la Computación	56
12763	Guía de recomendaciones para el tratamiento del Big Data como evidencia digital	42

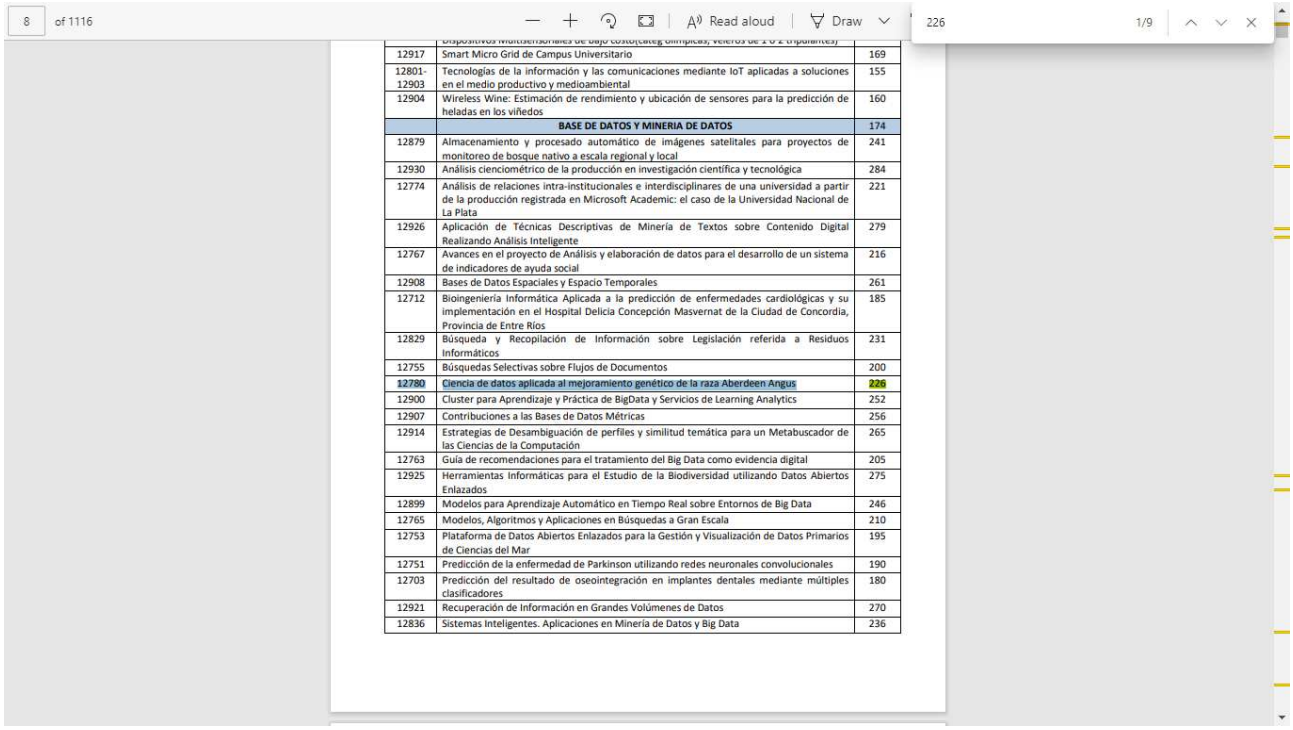
Poster WICC 2020



B.3. Capítulos de libros

Tapa Libro de Actas Artículo WICC 2020

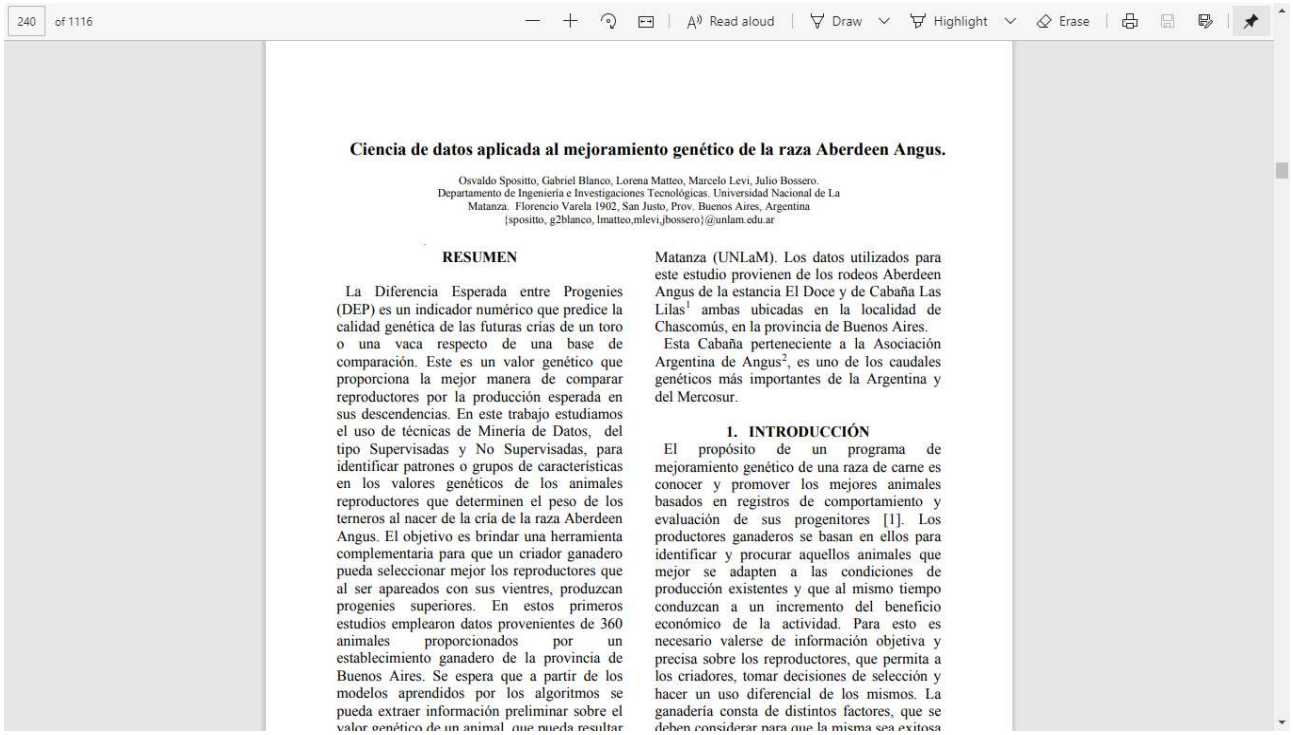
Índice Libro de Actas Artículo WICC 2020



12917	Smart Micro Grid de Campus Universitario	169
12801-12903	Tecnologías de la información y las comunicaciones mediante IoT aplicadas a soluciones en el medio productivo y medioambiental	155
12904	Wireless Wine: Estimación de rendimiento y ubicación de sensores para la predicción de heladas en los viñedos	160
	BASE DE DATOS Y MINERIA DE DATOS	174
12879	Almacenamiento y procesado automático de imágenes satelitales para proyectos de monitoreo de bosque nativo a escala regional y local	241
12930	Análisis cienciométrico de la producción en investigación científica y tecnológica	284
12774	Análisis de relaciones intra-institucionales e interdisciplinares de una universidad a partir de la producción registrada en Microsoft Academic: el caso de la Universidad Nacional de La Plata	221
12926	Aplicación de Técnicas Descriptivas de Minería de Textos sobre Contenido Digital Realizando Análisis Inteligente	279
12767	Avances en el proyecto de Análisis y elaboración de datos para el desarrollo de un sistema de indicadores de ayuda social	216
12908	Bases de Datos Espaciales y Espacio Temporales	261
12712	Bioingeniería Informática Aplicada a la predicción de enfermedades cardiológicas y su implementación en el Hospital Delicia Concepción Masvernat de la Ciudad de Concordia, Provincia de Entre Ríos	185
12829	Búsqueda y Recopilación de Información sobre Legislación referida a Residuos Informáticos	231
12755	Búsquedas Selectivas sobre Flujos de Documentos	200
12780	Ciencia de datos aplicada al mejoramiento genético de la raza Aberdeen Angus	226
12900	Cluster para Aprendizaje y Práctica de BigData y Servicios de Learning Analytics	252
12907	Contribuciones a las Bases de Datos Métricas	256
12914	Estrategias de Desambiguación de perfiles y similitud temática para un Metabuscador de las Ciencias de la Computación	265
12763	Guía de recomendaciones para el tratamiento del Big Data como evidencia digital	205
12925	Herramientas Informáticas para el Estudio de la Biodiversidad utilizando Datos Abiertos Enlazados	275
12899	Modelos para Aprendizaje Automático en Tiempo Real sobre Entornos de Big Data	246
12765	Modelos, Algoritmos y Aplicaciones en Búsquedas a Gran Escala	210
12753	Plataforma de Datos Abiertos Enlazados para la Gestión y Visualización de Datos Primarios de Ciencias del Mar	195
12751	Predicción de la enfermedad de Parkinson utilizando redes neuronales convolucionales	190
12703	Predicción del resultado de oseointegración en implantes dentales mediante múltiples clasificadores	180
12921	Recuperación de Información en Grandes Volúmenes de Datos	270
12836	Sistemas Inteligentes. Aplicaciones en Minería de Datos y Big Data	236

B.4. Trabajos presentados a congresos y/o seminarios

Artículo WICC 2020



Ciencia de datos aplicada al mejoramiento genético de la raza Aberdeen Angus.

Osvaldo Spositto, Gabriel Blanco, Lorena Matteo, Marcelo Levi, Julio Bossero.
Departamento de Ingeniería e Investigaciones Tecnológicas. Universidad Nacional de La Matanza. Florencio Varela 1902, San Justo, Prov. Buenos Aires, Argentina
{spositto, g2blanco, lmatteo,mlevi,jbossero}@unlam.edu.ar

RESUMEN

La Diferencia Esperada entre Progenies (DEP) es un indicador numérico que predice la calidad genética de las futuras crías de un toro o una vaca respecto de una base de comparación. Este es un valor genético que proporciona la mejor manera de comparar reproductores por la producción esperada en sus descendencias. En este trabajo estudiamos el uso de técnicas de Minería de Datos, del tipo Supervisadas y No Supervisadas, para 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. El objetivo es brindar una herramienta complementaria para que un criador ganadero pueda seleccionar mejor los reproductores que al ser apareados con sus vientres, produzcan progenies superiores. En estos primeros estudios emplearon datos provenientes de 360 animales proporcionados por un establecimiento ganadero de la provincia de Buenos Aires. Se espera que a partir de los modelos aprendidos por los algoritmos se pueda extraer información preliminar sobre el valor genético de un animal, que pueda resultar

Matanza (UNLaM). Los datos utilizados para este estudio provienen de los rodeos Aberdeen Angus de la estancia El Doce y de Cabaña Las Lilas[1] ambas ubicadas en la localidad de Chascomús, en la provincia de Buenos Aires.

Esta Cabaña perteneciente a la Asociación Argentina de Angus[2], es uno de los caudales genéticos más importantes de la Argentina y del Mercosur.

1. 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. La ganadería consta de distintos factores, que se deben considerar para que la misma sea exitosa

B.4. Trabajos presentados a congresos y/o seminarios

Artículo Autores WICC 2020 -- Vaquitas wicc 2020 vfinal.docx

Ciencia de datos aplicada al mejoramiento genético de la raza Aberdeen Angus.

Osvaldo Spositto, Gabriel Blanco, Lorena Matteo, Marcelo Levi, Julio Bossero.

Departamento de Ingeniería e Investigaciones Tecnológicas. Universidad Nacional de La Matanza. Florencio Varela 1902, San Justo, Prov. Buenos Aires, Argentina

{spositto, g2blanco, lmatteo,mlevi,jbossero}@unlam.edu.ar

RESUMEN

La Diferencia Esperada entre Progenies (DEP) es un indicador numérico que predice la calidad genética de las futuras crías de un toro o una vaca respecto de una base de comparación. Este es un valor genético que proporciona la mejor manera de comparar reproductores por la producción esperada en sus descendencias. En este trabajo estudiamos el uso de técnicas de Minería de Datos, del tipo Supervisadas y No Supervisadas, para 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. El objetivo es brindar una herramienta complementaria para que un criador ganadero pueda seleccionar mejor los reproductores que al ser apareados con sus vientres, produzcan progenies superiores. En estos primeros estudios emplearon datos provenientes de 360 animales proporcionados por un establecimiento ganadero de la provincia de Buenos Aires. Se espera que a partir de los modelos aprendidos por los algoritmos se pueda extraer información preliminar sobre el valor genético de un animal, que pueda resultar de gran utilidad en el sector ganadero en la toma de decisiones en un programa de mejoramiento genético.

Palabras clave: Diferencia Esperada entre Progenies, Minería de Datos, Algoritmos Supervisados y No Supervisados, Mejoramiento Genético.

CONTEXTO

La línea de investigación aquí presentada está enmarcada dentro del Programa de Incentivos para Docentes Investigadores de la Secretaría de Políticas Universitarias (PROINCE) 2019-2020. El mismo lleva el título: *Uso de Minería de Datos para Mejoramiento Genético en la raza Aberdeen Angus.* Este proyecto es financiado por la Universidad Nacional de La Matanza (UNLaM). Los datos utilizados para este estudio provienen de los rodeos Aberdeen Angus de la estancia El Doce y de Cabaña Las Lilas[22] ambas ubicadas en la localidad de Chascomús, en la provincia de Buenos Aires.

Esta Cabaña perteneciente a la Asociación Argentina de Angus[23], es uno de los caudales genéticos más importantes de la Argentina y del Mercosur.

1. 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. 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, entre otros.

La reproducción de bovinos mediante la Inseminación Artificial (IA) [2,3] es bastante sencilla y tiene muchas ventajas. Esta técnica se está aplicando desde hace bastante tiempo en el país. Hoy la Inseminación Artificial a Tiempo Fijo (IATF) es una técnica que, mediante la utilización de hormonas, permite sincronizar los celos y ovulaciones. Gracias a esto, es posible, inseminar una gran cantidad de animales en un corto período de tiempo [4].

[22] http://laslilas.com
[23] https://www.angus.org.ar/

Una de las herramientas utilizadas, por los ganaderos, para realizar las evaluaciones genéticas de los toros reproductores es la *Diferencia Esperada entre Progenie* (DEP), en base a estos valores, los productores pueden tomar decisiones de selección en base a información objetiva [5]. 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.

Seguidamente se detallan brevemente las descripciones de las siglas que componen los DEP's. Estas fueron extraídas del Anuario Las Lilas 2017 [6]:

- PN (Peso al nacer): Expresado en kilos, indica las diferencias genéticas para el PN de las crías de un padre determinado.
- PD (Peso al destete): Expresado en kilos. 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.
- CE (Circunferencia escrotal): Expresada en centímetros y ajustada por edad de vida, es un indicador indirecto de la fertilidad de los rodeos.
- 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): Predictor de la producción lechera y aptitud materna que transmite un toro a sus hijas.
- AOB (Área del Ojo de Bife): Es la altura de la 12a costilla. Es 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): Es un indicador del porcentaje de grasa intramuscular del músculo dorsal largo.

Este trabajo se realiza bajo la hipótesis que si se aplica una metodología para realizar Minería de Datos (MD) a partir de los datos del material genético de los progenitores machos, más ciertos datos de las hembras, como la edad, el historial de partos, etc., se puede construir un modelo predictivo que mejor determine la etiqueta Peso al Nacer (Ver Tabla 1). Por otro lado, mediante los algoritmos No Supervisados, se intenta demostrar las relaciones existentes entre las variables, que también tengan mayor injerencia en el PN de los terneros.

Tabla 3. Descripción de las variables del conjunto de datos.

Nomenclatura	Tipo de dato	Descripción
Peso Adulto	*Numérico*	Peso real del padre
PAN	*Numérico*	Peso al nacer
PAD	*Numérico*	Peso al destete
PAF	*Numérico*	Peso final
Circ. Escrotal	*Numérico*	Circunferencia escrotal
FRAME	*Numérico*	Altura del animal
Certificado	*Numérico*	Edad promedio de los vientres primerizos
PN	*Numérico*	DEP del Toro Progenitor
PD	*Numérico*	
AM	*Numérico*	
CM	*Numérico*	
PF	*Numérico*	
CE	*Numérico*	
AOB	*Numérico*	
GD	*Numérico*	
MAR	*Numérico*	
Peso al nacer	*Numérico*	Datos de la vaca
Peso al destete	*Numérico*	
Cantidad nacimientos	*Numérico*	
Cantidad abortos	*Numérico*	
Baja antes del destete	*Numérico*	
Edad en meses	*Numérico*	
Certificado	*Numérico*	DEP del Toro Progenitor de la vaca
CE del padre	*Numérico*	
PN	*Numérico*	
PD	*Numérico*	
AM	*Numérico*	
CM	*Numérico*	
PF	*Numérico*	
CE	*Numérico*	
AOB	*Numérico*	
GD	*Numérico*	
MAR	*Numérico*	
Peso_Nac	*Texto*	Atributo Clase

En el primer año de la investigación se seleccionaron diferentes técnicas del tipo Supervisadas y No Supervisadas [7,8].

Las primeras son aquellas orientadas a predecir el valor de un atributo (etiqueta o clase) de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de datos, cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Las segundas, también conocidas como técnicas de Clustering, agrupan 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.

Existen en la actualidad distintas metodologías para llevar a cabo un proceso de MD, en este trabajo se optó, siguiendo la literatura consultada, por Cross Industry Standard Process for Data Mining (CRISP-DM) [8,9]. Esta tecnología interrelaciona diferentes fases del proceso entre sí, de tal manera que se consolida un proceso iterativo y recíproco.

Para la ejecución del modelo construido y las pruebas para realizar una comparación entre las diferentes técnicas, se utilizó el software WEKA[24] (acrónimo de Waikato Environment for Knowledge Analysis, en español «entorno para análisis del conocimiento de la Universidad de Waikato»). Para realizar las mismas, fue necesario la construcción de un modelo con los datos históricos de anteriores rodeos y sus resultados respecto al peso al nacer de sus crías. Para optimizar el modelo fue necesario normalizar las variables de entrada. Normalizar significa, en este caso, comprimir o extender los valores de la variable para que estén en un rango definido. Se empleó la fórmula de Normalización mínimo-máximo, la cual 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 [10], que se define como:

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

Para la evaluación de los clasificadores se utilizó una Matriz de Confusión (MC) y el análisis o curvas ROC (acrónimo de Receiver Operating Characteristic) [8], estos se encuentran dentro del software WEKA para cada clasificador. Con estas herramientas se evalúa entre otras cosas:

- La sensibilidad indica la capacidad de nuestro clasificador para dar como casos positivos los casos que realmente lo son; proporción de PN altos correctamente identificados.
- La especificidad indica la capacidad de nuestro estimador para dar como casos negativos los casos que realmente lo

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

sean; por ejemplo: proporción de PN bajos correctamente identificados como tal.

- Este trabajo tiene como objetivo principal generar conocimiento especializado en el área de Minería de Datos en lo referente a un programa de mejoramiento genético. Específicamente, esta línea se centra principalmente en el estudio de dos ejes: los algoritmos Supervisados y No Supervisados.
- Además, como ya se mencionó, los datos de entrada, para realizar las pruebas contaron con los valores genéticos de sus padres/abuelos. El mayor obstáculo que se presentó es la cantidad y calidad de los mismos, si bien es 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.
- En algunos casos, incrementar el número de variables, mejora el rendimiento de los clasificadores, se pretende agregar variables tales como: el tipo de alimentación de los animales, el factor climático, etc.
- Se estudiarán nuevos algoritmos y con varias configuraciones para comprobar cual tiene la mayor capacidad de exactitud en su predicción.
- Por último, utilizar diferente herramienta computacional, por ejemplo probar con software más sofisticados como Matlab[1], SPSS[2], etc.

2. LÍNEAS DE INVESTIGACIÓN y DESARROLLO

Este trabajo tiene como objetivo principal generar conocimiento especializado en el área de Minería de Datos en lo referente a un programa de mejoramiento genético. Específicamente, esta línea se centra principalmente en el estudio de dos ejes: los algoritmos Supervisados y No Supervisados.

Además, como ya se mencionó, los datos de entrada, para realizar las pruebas contaron con los valores genéticos de sus padres/abuelos. El mayor obstáculo que se presentó es la cantidad y calidad de los mismos, si bien es 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.

En algunos casos, incrementar el número de variables, mejora el rendimiento de los clasificadores, se pretende agregar variables tales como: el tipo de alimentación de los animales, el factor climático, etc.

Se estudiarán nuevos algoritmos y con varias configuraciones para comprobar cual tiene la mayor capacidad de exactitud en su predicción.

Por último, utilizar diferente herramienta computacional, por ejemplo probar con software más sofisticados como Matlab[1], SPSS[2], etc.

3. RESULTADOS OBTENIDOS Y ESPERADOS

Este grupo de investigación viene trabajando en proyectos PROINCE en años anteriores, también asociados a la misma temática:

- PROINCE C176 (2015-2016). "*Análisis Comparativo de Modelos de Clasificación de Minería de Datos (Data Mining). Su Aplicación en la Predicción de Perfiles de Alumnos en Riesgo de Deserción*".
- PROINCE C199 (2017-2018). "*Modelos de Minería de Datos para el Diagnóstico Pre_ coz de Enfermedades Neurodegenerativas*".
- PROINCE C205 (2017-2018). *"Uso de Minería de Datos para Acelerar la Recuperación de Documentos"*.

Los resultados de estas investigaciones dieron lugar a varias publicaciones [11-15].

Tal como quedó expuesto, se utilizaron dos tipos distintos de técnicas de MD. Se presentaron 2 trabajos en el año 2019. En uno se realizó una comparación en cuanto al desempeño de tres algoritmos del tipo Supervisado [16]:

- Árboles de Decisión (AD).
- Redes Neuronales Artificiales (RNA).
- Máquinas de Soporte Vectorial (MSV o SVM, del inglés Support Vector Machine)

En el otro trabajo presentado, se compararon cuatro algoritmos No Supervisados [17]:

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

En el primero, se encontró que el modelo propuesto conserva una precisión (proporción de instancias clasificadas correctamente) aceptable en el caso del algoritmo AD, con un porcentaje levemente superior al 70%.

Para los algoritmos No supervisados, fue necesario un subconjunto de atributos menor del conjunto total inicial, que incluya aquellos relevantes para la tarea de agrupamiento. Se llegó a la conclusión que los datos relevantes para agrupar a los terneros según su PN involucra principalmente las características de su Madre y del Abuelo Materno.

En el segundo año se espera que, los criterios mencionados anteriormente, logrados de forma automática por los algoritmos, se puedan comparar con la estimación de los expertos en la realidad de los terneros a nacer este año. Es importante notar que la clasificación puede diferir. Sobre todo, teniendo en cuenta, la dificultad de calcular el peso del recién nacido.

Por último, se están realizando con la Sociedad Rural de Chascomús y la Asociación Argentina de Angus, esto posibilitaría la obtención de datos de mejor calidad y de nuevos establecimientos ganaderos.

4. FORMACIÓN DE RECURSOS HUMANOS

Parte del grupo de desarrollo del proyecto trabaja desde el año 2015 en diversas áreas relacionadas con la Minería de Datos.

Actualmente forman parte del equipo, además de docentes de la UNLaM, una Ingeniera Agrónoma que trabaja en el SENASA y dos alumnos becarios de investigación.

Por otra parte, los docentes-investigadores que integran el proyecto dictan clases en la cátedra de Inteligencia de Negocio de la carrera Licenciatura en Gestión de Tecnología y en la cátedra Base de Datos y Data Mining y Data Warehouse de la carrera de Ingeniería Informática. Se prevé, además, la capacitación y formación de recursos humanos, a través de cursos de actualización y posgrado en el área de estudio; la transferencia de conocimiento y resultados; y la posibilidad de brindar charlas informativas del desarrollo e implementación del proyecto a distintas instituciones del sector ganadero, como la Sociedad Rural de Chascomús y la Asociación Argentina de Angus.

5. BIBLIOGRAFÍA

1. Firpo Brenta, L. y otros. (2012). Selección genética y mejoramiento animal. Disponible en: http://www.produccion-animal.com.ar /genética_seleccion_cruzamientos/bovinos_en_general/24-Seleccion_genetica.pdf. Último acceso: 06/09/2019.
2. Agrocor. (2011). Inseminación artificial en bovinos Curso Teórico Práctico de Inseminación Artificial en Bovinos. Disponible en: https://www.engormix.com /ganaderia-carne/articulos/inseminacion-artificial-en-bovinos-t26957.htm. Último acceso: 06/09/2019.
3. Díaz, P. Fonseca, V. Martínez P. y Rey A. (2003). Inseminación Artificial en bovinos. Biblioteca Digita, U. de Chile. Disponible en: www.biblioteca.org.ar/libros/8913.pdf. Último acceso: 06/02/2020.
4. Sommantico, S. (2018). Inseminación Arti_ ficial a Tiempo Fijo: la tecnología de la que se habla mucho y se usa poco. Disponible en: https://www.infocampo.com.ar/inseminacion- artificial-a-tiempo-fijo-la-tecnologia-de-la-que-se-habla-mucho-y-se-usa-poco/ Último acceso: 26/02/2020.
5. Guitou H. y Monti A. (1998). Interpretación y uso correcto de los DEPs como herramienta de selección. INTA Castelar. Disponible en: https://es.scribd.com/document/337947981/20-Interpretacion-Deps. Último acceso: 26/02/2020.
6. Cómo interpretar la evaluación genética. Anuario Las Lilas 2018-2019. Cabaña Las Lilas. Centro de Genética. Pág. 107. Disponible en: http://laslilas.com/pdf/Anuario -Genetica-2017.pdf. Último acceso: 26/02/2020.
7. Perez Lopez, C. y otros. (2007). Minería de datos. Técnicas y herramientas ISBN: 9788497324922. Ed. Paraninfo Cengage L. Madrid. España.
8. Hernández Orallo, J. y otros. (2004). Introdu_ cción a la minería de datos". Pearson. Edición: I.
9. Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. (2005). U. de Alcalá, Madrid Disponible en: http://www.sc.ehu.es /jiwdocoj/remis/docs/GarreAdis05.pdf. Ultimo acceso: 26/02/2020.
10. Han Jiawei. Data Mining: Concepts and Techniques. 3ra. Edición. (2011). ISBN 978-0-12-381479-1.
11. Predicción del riesgo de abandono universitario utilizando métodos supervisados. (2016) IPECyT 2016. F. R. B. Blanca.
12. "Comparación de Algoritmos de Aprendizaje Supervisado para la obtención de perfiles de alumnos desertores". (2016). el CONAIISI 2016. Salta. Argentina.
13. "Modelos de minería de datos para el diagnóstico de enfermedad de Parkinson mediante el análisis de voz". Presentado CONAIISI 2017. Santa Fe. Argentina.
14. "Aceleración en la Recuperación de Información utilizando Algoritmos de Minería de Datos de R". Presentado en CACIC 2018. U. N. del Centro, Tandil.
15. "Selection of voice parameters for Parkinson´s disease prediction from collected mobile data". (2019) XXII Symposium on Image, Signal Processing and Art. Vision. Bucaramanga, Colombia.
16 *"Clasificación del Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos Supervisados"*. Trabajo presentado y no publicado en JAIIO 2019. UNSa.
17 *"Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos No Supervisados"*. Trabajo presentado en CONAIISI 2019. UNLaM. Buenos Aires. Argentina.

B.4. Trabajos presentados a congresos y/o seminarios

Certificado Autores CACIC 2020

Artículo Autores CACIC 2020 -- Algoritmos de balanceo de clases Cacic 2020-v4.doc

SMOTE, Algoritmo para balanceo de clases en un estudio aplicado a la ganadería.

Abstract. En el estudio de los algoritmos de Minería de Datos del tipo supervisados surge el problema del desbalance de clases, que implica que la información no se encuentre distribuida equitativamente entre todas las clases que la componen, por lo que se generan efectos no deseados en el proceso de clasificación. Este trabajo considera el caso de conjuntos de datos que solamente tiene dos clases y una de ellas cuenta con una mayor cantidad de ejemplos que la otra. El interés principal del trabajo es la aplicación de la técnica de balanceo de clases SMOTE (Synthetic Minority Oversampling Technique), que con algoritmos de interpolación incrementa en forma "sintética" los ejemplos de la clase minoritaria. Los resultados experimentales muestran que algunas técnicas, en el proceso de entrenamiento, obtienen mejores porcentajes de clasificación, cuando se usan estos datos artificiales. El dataset utilizado registra la Diferencia Esperada entre Progenie de animales de la raza Aberdeen Angus.

Keywords: Desbalance de clases, SMOTE, Algoritmos Supervisados, Weka, DEP

1 Introducción

Se ha observado que algunas de las técnicas de MD del tipo supervisadas, si utilizan un conjunto de datos desequilibrado para la clasificación, presentan un rendimiento de generalización deficiente, debido a un fuerte sesgo hacia las clases mayoritarias [1]. Las clases con el mayor número de instancias se denominan clases mayoritarias y las clases con el menor número de instancias son referido como las clases minoritarias. Intuitivamente, dado que hay una gran cantidad de ejemplos de clases mayoritarias, un modelo de clasificación tiende a favorecer las clases mayoritarias mientras que clasifica incorrectamente los ejemplos de las clases minoritarias [2].

Las técnicas supervisadas, se expresan mediante algoritmos capaces de tratar y analizar datos de forma automática, con el objeto de extraer cualquier tipo de información subyacente en dichos datos. Como se sabe, en el aprendizaje supervisado, los algoritmos trabajan con datos "*etiquetados*", intentado encontrar una función que, dadas las variables de entrada, les asigne la etiqueta de salida adecuada. El algoritmo se entrena con un "histórico" de datos y así "aprende" al asignar la etiqueta de salida a un nuevo valor, es decir, predice el valor de salida [1].

El problema del desbalanceo de las clases, entonces, consiste en la predominancia de ciertos valores en los datos de entrenamiento y la escasez de otros.

Este trabajo, es parte de un proyecto de investigación que se lleva adelante en la Universidad Nacional de La Matanza, denominado "*Uso de Minería de Datos para Mejoramiento Genético en la Raza Aberdeen Angus*", donde se estudia la aplicación de distintas técnicas de MD con el objeto de encontrar, a partir de los valores genéticos de animales de la raza Aberdeen Angus, patrones o grupos de características que puedan determinar a priori, el peso de los terneros al nacer. En esta investigación se estudiaron tanto algoritmos del tipo No Supervisados como Supervisados. En la aplicación de los primeros, se utilizaron las siguientes técnicas: EM (Expectation Maximization), FarthestFirst, Simple K-Means y Mapas AutoOrganizados (Redes SOM). Ese trabajo titulado "*Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos No Supervisados*", fue publicado en CoNaIISI 2019 [3]. En este trabajo se puede leer el procedimiento realizado para obtener los datos usados para confeccionar la vista minable [4]. Estos mismos datos fueron utilizados en los experimentos con métodos supervisados y en particular en el estudio que es motivo de esta presentación.

En cuanto al estudio, sobre el empleo de técnicas Supervisadas, se realizó una comparación entre tres clasificadores: Árbol de Decisión (AD), Red Neuronal Artificial (RNA), del tipo Perceptron Multicapa y una Máquina de Soporte Vectorial (MSV). Este trabajo, se tituló "*Clasificación del Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos Supervisados*", fue presentado y aprobado en JAIOO 2019, pero no se publicó.

En ambos trabajos se utilizó un modelo de datos o vista minable, compuesto con datos provenientes de las evaluaciones genéticas de los toros: conocidos como la Diferencia Esperada entre Progenie (DEP) [5], que permite a los productores tomar decisiones de selección en base a información objetiva. 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 genéticos de las hembras, y con datos históricos de los progenitores, se alcanzaron valores aceptables de predicción, respecto de la variable a clasificar: Peso al nacer (PN), valor que de acuerdo a los veterinarios es determinante del potencial del desarrollo futuro del animal. Dicha variable se estableció, para los trabajos mencionados anteriormente, en "*Alta*" y "*Baja*". Un PN promedio es de 38 kilogramos, por tal motivo, a los pesos mayores o iguales a 38 kg., se los clasificó como Alto y al resto como Bajo.

Se concluyó que, con los algoritmos utilizados, en este último estudio, se obtuvieron valores aceptables (porcentaje de aciertos superior al 60%), en referencia a las métricas: *precisión, sensibilidad y especificidad*. Y que, para estos datos y con la configuración propuesta, para cada técnica, por el software WEKA[25] (Waikato Environment for Knowledge Analysis, en español "*entorno para análisis del conocimiento de la Universidad de Waikato"*) [7], el algoritmo Árbol de Decisión tuvo la mejor precisión, es decir, la mejor probabilidad de discriminar correctamente, debido a que el valor de su media muestral es mayor: 72.5%. Cabe resaltar que el indicador precisión es una de las medidas principales, para establecer el desempeño de un algoritmo de clasificación en el área de la MD [1].

El principal interés en este trabajo es mejorar la clasificación obtenida en el trabajo expuesto en el párrafo anterior. Para ello, se modificaron, en la vista minable, los objetos de la clase minoritaria, sin eliminar objetos de la clase mayoritaria, lo cual, según la literatura actual, puede producir pérdida de información importante [2][8], mediante la aplicación del filtro SMOTE (cuya traducción al español es "*técnica de sobre muestreo de minorías sintéticas*") [9]. Luego, se realizó una nueva comparación, utilizando ese nuevo set de datos, a través del software WEKA, usando siempre la configuración por defecto. A continuación, se hace una reseña de algunos trabajos relacionados con el uso del algoritmo SMOTE. No se han encontrado trabajos en donde se aplique la técnica SMOTE relacionado con la ganadería.

Antecedentes y Trabajos Relacionados

El problema del desbalanceo de clases es una cuestión que se está abordando en la actualidad de forma activa y son muchos los investigadores que estudian y proponen nuevas técnicas para poder hacerle frente a este problema. La mayoría de ellos solo se han concentrado en resolver situaciones como la nuestra de clasificación que tienen que ver con dos clases. En [8], se encuentra un detalle de varios trabajos relacionados con el desbalanceo de clases ordenados cronológicamente. Algunos otros trabajos encontrados respecto al uso de SMOTE son los siguientes:

- De Jesús, Juan. (2016). *Técnicas de muestreo para mejorar el rendimiento del algoritmo back-propagation en problemas de desbalance de clases: Un estudio empírico sobre la clasificación en imágenes de percepción remota*.
  Resumen: En este trabajo se analizaron las diferentes técnicas de re muestreo para tratar el desbalance de clases en dominios de dos clases con ayuda de datos de imágenes de percepción, para determinar que técnica arroja un mejor clasificador. Se utilizó una red neuronal del tipo perceptron multicapa como clasificador. Los resultados mostraron aquellos algoritmos que mejor funcionaron al momento de clasificar, de acuerdo a un análisis estadístico aplicado a los algoritmos.
- J. Monroy de Jesús y otros. (2018). *Algoritmo de aprendizaje eficiente para tratar el problema del desbalance de múltiples clases*.
  Resumen: Los resultados mostrados demostraron que la diversidad de versiones de algoritmos y cuan competitivos resultaron en el desempeño de la clasificación con respecto a los métodos de sobre-muestreo y sub-muestreo (ROS, SMOTE y RUS).
- David Municio Duran. (2019) *Técnicas de oversampling aplicadas al análisis de imágenes hiperespectrales.*
  Resumen: En este trabajo se ha estudiado en detalle el impacto de diferentes algoritmos de oversampling en el proceso de clasificación de imágenes hiperespectrales. El objetivo del trabajo fue mejorar los resultados de clasificación en imágenes con alta dimensionalidad y gran desbalanceo entre clases. Los resultados obtenidos evidenciaron las bondades del proceso de balanceo de datos.
- Rosa María Valdovinos Rosas. (2006). *Técnicas de Submuestreo, Toma de Decisiones y Análisis de Diversidad en Aprendizaje Supervisado con Sistemas Múltiples de Clasificación.*
  Resumen: Este trabajo se encarga del estudio de Sistemas Múltiples de Clasificación (SMC), para el reconocimiento de patrones. Específicamente, este trabajo se centró en la limpieza de un conjunto de datos, se emplearon para la reducción del tamaño del conjunto de entrenamiento, un algoritmo de subconjunto selectivo modificado, y para la generación de patrones sintéticos, se utilizó el algoritmo SMOTE.

[25] Es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Weka es software libre distribuido bajo la licencia GNU-GPL. www.cs.waikato.ac.nz/~ml/weka/

2. Algoritmo de balanceo de clases

Para este trabajo se hace uso del algoritmo SMOTE como parte de los métodos basados en muestreo (sampling), para balanceo de clases [2]. Se ha elegido SMOTE debido a que es uno de los algoritmos más utilizados para tratar el problema de desbalanceo de clases [8][10].
SMOTE es una técnica basada en sobremuestreo (oversampling), que genera instancias "*sintéticas*" o artificiales en el espacio de atributos con el objetivo de equilibrar la muestra de datos basado en la regla del vecino más cercano.
Para comprender como funciona SMOTE, es necesario conocer el algoritmo de la regla del Vecino más cercano [1], que consiste en suponer que instancias próximas entre sí tienen mayor probabilidad de pertenencia a la misma clase. La generación se realiza interpolando nuevas instancias en lugar de duplicarlas como hacen los algoritmos del tipo re-muestreo (resampling) [1]. Para cada una de las instancias minoritarias se buscan las instancias minoritarias vecinas (más cercanas). Se crean N o α (alfa) instancias sobre el segmento que une la instancia original y cada una de las vecinas. Se encontró un estudio [8], que hace un trabajo comparativo entre distintos métodos basados en el algoritmo SMOTE. Asimismo, en el trabajo de Rodríguez Torres [11], se encuentra una descripción algorítmica de esta técnica, que puede ser reescrita en estos tres pasos

1) Se determina la cantidad promedio de interpolaciones que debe aportar cada elemento minoritaria: α (ya sea calculado o elegido).

2) Se asigna a cada muestra minoritario el número $\lfloor\alpha$ o $\lceil\alpha$[26], cantidad de interpolaciones en las cuales intervendrá ya sea con ayuda de azar o determinístico de modo tal de lograr la cantidad de muestras artificiales deseada.

3) Luego se procesa cada elemento de la muestra calculando según el número asignado la cantidad de vecinos cercanos interpolando por azar un punto sobre el segmento que los une.

Entonces, para cada muestra minoritaria se efectúan el el $\lfloor\alpha$ o $\lceil\alpha$. La ubicación dentro del segmento que los une también se determina por azar. Entonces, se calcula la cantidad promedio de interpolaciones que debe aportar cada elemento minoritario: α (da lo mismo calculado o elegido). Se asigna a cada muestra minoritario el número $\lfloor\alpha$ o $\lceil\alpha$ cantidad de interpolaciones en las cuales intervendrá ya sea con ayuda de azar o determinístico de modo tal de lograr la cantidad de muestras artificiales deseada. Luego se procesa cada elemento de la muestra calculando según el número asignado la cantidad de vecinos cercanos interpolando por azar un punto sobre el segmento que los une.

La Figura 1 muestra un ejemplo de procedimiento SMOTE:

A) Para cada ejemplo de la clase minoritaria ***k***, calcula el vecino más cercano (***i, j, l, n. m***).
B) Se elige aleatoriamente un ejemplo de 5 puntos más cercanos.
C) Se genera sintéticamente el evento k_1, de modo que k_1 se encuentra entre k e i.
D) Se ven el conjunto de datos después de aplicar SMOTE 3 veces.

Estos ejemplos sintéticos ayudan a equilibrar la distribución original de la clase, que generalmente mejora significativamente el aprendizaje. Sin embargo, el algoritmo SMOTE también tiene su desventaja, como por ejemplo generalización del espacio de clase minoritaria [12].

[26] En matemática, según la forma de considerar el número entero más próximo a un número real dado, se pueden considerar: Al entero inferior se lo llamar piso ($\lfloor$) y al entero superior techo ($\lceil$).

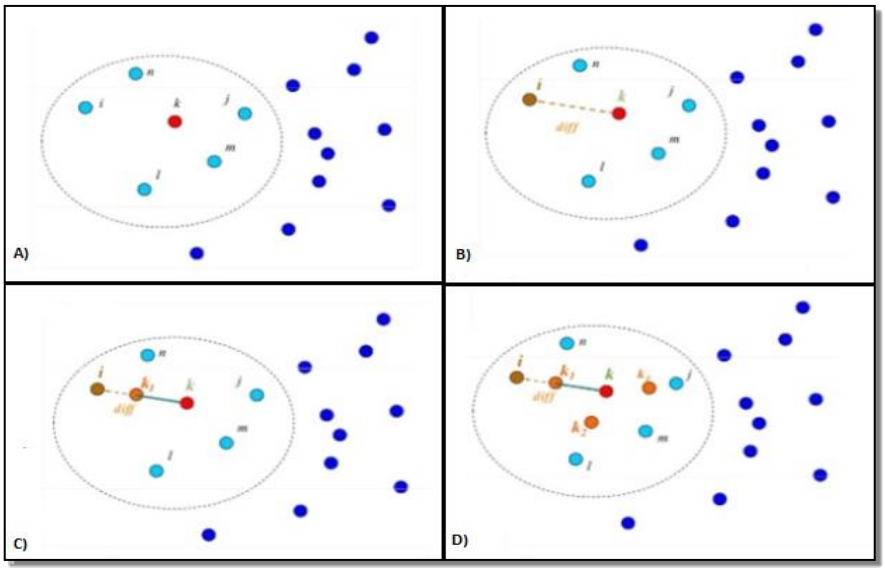


Fig. 1. Ejemplo del algoritmo SMOTE [16].

A partir del algoritmo SMOTE original, se han desarrollado muchos otros algoritmos basados en SMOTE a lo largo de los años y algunos de ellos mejoran efectivamente el rendimiento en el aprendizaje desequilibrado [2][10]. El principal inconveniente del algoritmo es el alto coste computacional que tiene.

3 Método utilizado y resultados

Como ya se mencionó, este trabajo representa una continuación del trabajo "*Clasificación del Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos Supervisados*". En esta oportunidad se realizó una comparación entre los resultados con los datos originales, contra resultados obtenidos al ejecutar los mismos algoritmos con datos sintéticos, estos creados con el filtro SMOTE, a través del software WEKA. El modelo original, descriptos en párrafos anteriores, se detallan en [3]. Estos datos correspondientes a los años 2017 y 2018, contiene un total de 360 ejemplares hembras, las cuales fueron inseminadas por dos reproductores de la cabaña Las Lilas. La nómina de variables utilizas se muestra en la Tabla 1.

Tabla 4. 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.

Se describen brevemente las técnicas que se compararon. Para más información, sobre estos algoritmos se pueden consultar en [1][7][13].

- Árboles de Decisión (AD): como su nombre lo indica es una estructura que se forma por las bifurcaciones en cada una de las decisiones, descubriendo reglas. En WEKA se lo conoce como algoritmo J48, que es una implementación libre en java del algoritmo C4.5, que utiliza el concepto de entropía de la información para la selección de variables que mejor clasifiquen a la variable PN (clase) estudiada.
- Red Neuronal Artificial (RNA): Esta implementación imita el funcionamiento interno de las neuronas humanas [1]. En general, aunque pueden usarse muchos tipos de RNA para clasificación, se han usado las redes multicapa feedforward o perceptron multicapa (MLP) que son los clasificadores basados en redes neuronales más ampliamente estudiados y utilizados. Este algoritmo es entrenado para realizar conexiones entre los valores de entrada y salida, aprendiendo de su error de pronóstico.
- Máquinas de Soporte Vectorial, buscan el límite que separa las clases con el mayor margen posible [1][14], Una de las características de esta técnica es que cuando no se pueden separar correctamente las dos clases, el algoritmo buscan el mejor límite posible. Las MVS efectúan esto, sólo con una línea recta (usa un kernel lineal) y gracias a esta aproximación lineal, se puede ejecutar con bastante rapidez.

Como se hizo anteriormente, para la evaluación de los clasificadores se empleó una Matriz de Confusión (MC) [1][13] y el análisis o curvas ROC (acrónimo de Receiver Operating Characteristic) [1][15], que entrega WEKA, como resultado luego de testear cada uno de los clasificadores. A modo de resumen, una matriz de confusión muestra la clasificación de las instancias. Brinda información muy útil porque no sólo refleja los errores producidos sino también informa del tipo de éstos. Donde:

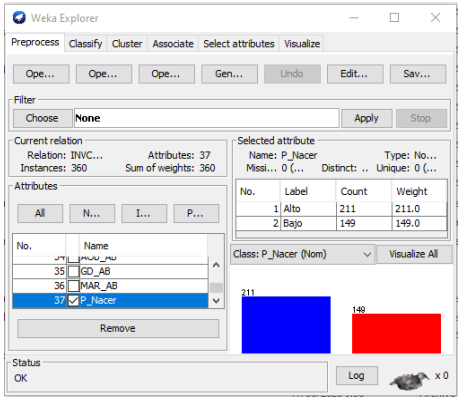
- VP es la cantidad de positivos que fueron clasificados correctamente como positivos por el modelo.
- VN es la cantidad de negativos que fueron clasificados correctamente como negativos por el modelo.
- FN es la cantidad de positivos que fueron clasificados incorrectamente como negativos.
- FP es la cantidad de negativos que fueron clasificados incorrectamente como positivos.

De estos valores se definen dos métricas asociadas importantes: Sensibilidad y especificidad:

- La sensibilidad nos indica la capacidad de nuestro estimador para dar como casos positivos los casos realmente lo son; proporción de pesos altos correctamente identificados.
- La especificidad nos indica la capacidad de nuestro estimador para dar como casos negativos los casos realmente lo sean; en nuestro caso, proporción de pesos bajos correctamente identificados.

Por último, la curva ROC, es una herramienta estadística utilizada en el análisis de los clasificadores, determinando la capacidad discriminante de una prueba.

Para realizar la comparación se tomaron los 360 datos originales y luego de cargarlos en WEKA, se le aplicó el filtro SMOTE. Weka permite configurar el valor del vecino más cercano (y por el porcentaje de instancias que se necesita crear para que las clases se balanceen, en este caso no se usó el valor por defecto del programa (100%), sino que se ajustó al 40%, para igualar las clases. En la figura 3 se puede observar la distribución de la variable PN, en los datos originales y en la figura 4 como cambió la distribución de valores luego de aplicar el filtro. Como se observa los 360 datos originales, se distribuían: 211 instancias de PN Altas y 149 Bajas, luego de aplicar el filtro, las instancias se convirtieron en 419, es decir, 59 instancias más en el peso bajo. Dejando al PN Alta, con la misma cantidad.



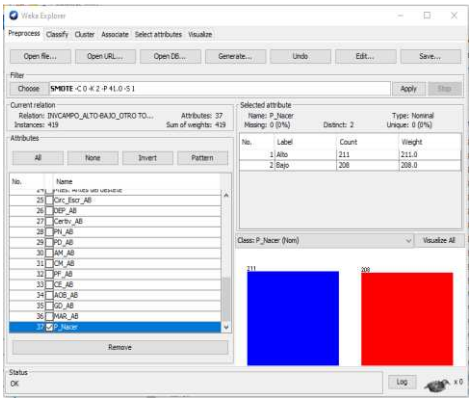


Fig. 2. Distribución con datos originales **Fig. 3. Distribución con datos sintético**

Luego de ejecutar cada uno de los algoritmos con el nuevo set de datos, en la siguiente Tabla es posible comparar los resultados obtenidos luego de ejecutar en WEKA cada uno de los algoritmos, siempre usando la opción ***Use training set.***

Tabla 5. Comparación de los resultados obtenidos de las pruebas de clasificación con datos originales y con datos sintéticos.

Porcentaje de Instancias Clasificadas Datos Originales			Porcentaje de Instancias Clasificadas Datos con instancias artificiales		
RNA	**MVS**	**Árbol de Decisión**	**RNA**	**MVS**	**Árbol de Decisión**
Correctas: 63.9 % Incorrectas: 36.1 %	Correctas: 60.3 % Incorrectas: 39.7 %	Correctas: 72.5 % Incorrectas: 27.5 %	Correctas: 66.8 % Incorrectas: 33.2 %	Correctas: 52.9 % Incorrectas: 47.1 %	Correctas: 65.9 % Incorrectas: 34.1 %
Matriz de Confusión a-Alto b-Bajo			**Matriz de Confusión a-Alto b-Bajo**		
a b 146 65 65 84	a b 173 38 105 44	a b 184 27 72 77	a b 186 25 114 94	a b 194 17 180 28	a b 150 61 82 126
Curva Roc			**Curva Roc**		
0.68	0.558	0.746	0.749	0.527	0.688

Se observa que solo la RNA mejoró su precisión, en cambio los otros clasificadores mermaron sus porcentajes. Respecto a las áreas bajo la curva (AUC) en las Figuras 5 y 6, está la gráfica de ambas curvas. AUC puede interpretarse como la probabilidad de que, ante una instancia nueva de datos, la prueba los clasifique correctamente. Su rango de valores va desde 0, siendo este valor el correspondiente a una prueba sin capacidad discriminante, hasta 1, que es cuando los dos grupos están perfectamente diferenciados por la prueba. Por tanto, podemos decir que cuanto mayor sea el AUC mejor será la prueba. También la RNA, tuvo el mejor valor.

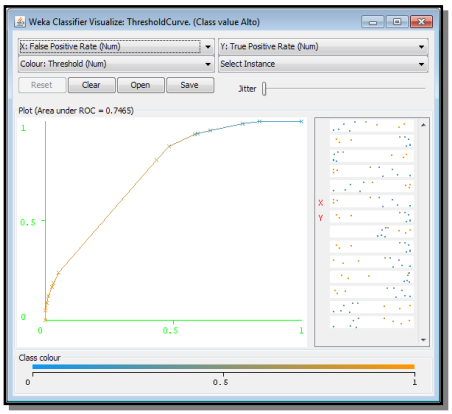


Fig. 1. Curva Roc del AD con los datos originales.

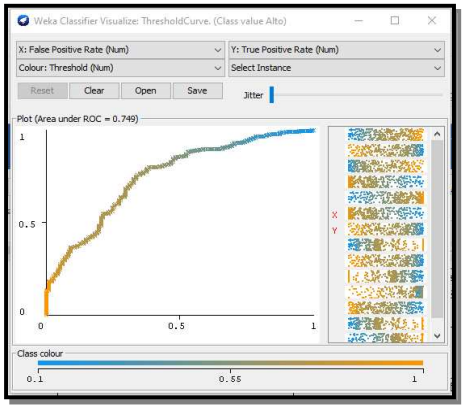


Fig. 2. Curva Roc de la RNA con datos formados por el filtro SMOTE.

4 Conclusiones y Trabajo a Futuro

Este estudio tuvo como objetivo verificar el comportamiento de tres algoritmos supervisados, con un dataset que tiene una porción de los datos construidos en forma artificial, mediante la aplicación del algoritmo de sobre-muestreo SMOTE, provisto por el software WEKA.

Después de haber efectuado todas las pruebas pertinentes sobre los modelos de clasificación propuestos usando estos datos sintéticos, es posible elaborar una serie de conclusiones:

- Es el primer trabajo sobre desbalanceo de clases en el área de la ganadería con estos tipos de datos.
- Dependiendo del tipo de algoritmo, se demuestra que con algunas técnicas se puede mejorar los porcentajes de instancias bien clasificadas. En este caso, las MSV no tuvo mejoras.
- Se usaron distintos tipos de valores alfa, siempre se observó que el algoritmo, siempre duplicó la clase minoritaria.
- Se debería hacer un estudio sobre las instancias generadas artificialmente para ver el grado de similitud que tienen con la clase minoritaria.

Para futuras investigaciones se prevé seguir con las siguientes líneas:

- Realizar una comparación con los algoritmos No supervisados, empleados también en la investigación anterior [3].
- Probar con vistas minables de mayor cantidad de muestras y con un mayor porcentaje de diferencia entre las clases mayoritarias y minoritarias.
- Probar con vistas minables que posean más de dos clases.

- Probar con otras variantes del algoritmo SMOTE [8] y con otros algoritmos que tratan el desbalanceo de clase. Por ejemplo: el método de Sobremuestreo por agrupaciones o Clustered Based Oversampling (CBOS).
- Incursionar en otros programas para el aprendizaje automático y la minería de datos.

Referencias

1 Hernández Orallo, Introducción a la minería de datos, Pearson, 2004.

2 Mera, C. & Arrieta Ramos, J.M., «Estudio Comparativo de Técnicas de Balanceo de Datos en el Aprendizaje de Múltiples Instancias», 2015. [En línea]: https://www.researchgate.net /publication/283642919_Estudio_Comparativo_de_Tecnicas_de_Balanceo_de_Datos_en_el_Aprendizaje_de_Multiples_Instancias/link/5642151a08aec448fa621f60/download. [Último acceso: 01/07/2020].

3 Spositto, O., «Peso al Nacer de Terneros Aberdeen Angus mediante Algoritmos No Supervisados,» 2019. [En línea]: https://www.researchgate.net/profile/Lorena_Matteo/ publication/337445353_Peso_al_Nacer_de_Terneros_Aberdeen_Angus_mediante_Algoritmos_No_Supervisados/links/5dd7f187458515dc2f439029/Peso-al-Nacer-de-Terneros-Aberdeen-Angus-mediante-Algoritmos-No-Supervisad. [Último acceso: 01/07/2020].

4 Quinteros, O. y otros., «Construcción de una vista minable para aplicar minería de datos secuenciales temporales», 2016. [En línea]: http://sedici.unlp.edu.ar/handle/10915/56747. [Último acceso: 01/07/2020].

5 Monti, A, «Interpretación y uso correcto de los DEPs como herramienta de selección», 1998. [En línea]: https://es.scribd.com/document/337947981/20-Interpretacion-Deps. [Último acceso: 01/07/2020].

6 Simeone, O., «A Very Brief Introduction to Machine Learning», 2018. [En línea]: https://arxiv.org/pdf/1808.02342.pdf. [Último acceso: 01/07/2020].

7 Witen I., Data Mining. Practical Machine Learning Tools and Techniques., Morgan Kaufmann. ISBN: 0-12-088407-0., 2005.

8 Castro Pérez, N. «Preprocesamiento de datos termográficos por medio de técnicas de balanceo de clases y análisis de cúmulos (Clustering)», 2013. [En línea]: http://docplayer.es/ 17472207-Preprocesamiento-de-datos-termograficos-por-medio-de-tecnicas-de-balanceo-de-clases-y-analisis-de-cumulos-clustering.html. [Último acceso: 2020 07 01].

9 Chawla, N.V., «SMOTE: Synthetic Minority Oversampling Technique», 2002. [En línea]: https://www.jair.org/index.php/jair/_article/view/10302. [Último acceso: 01/07/2020].

10 Moreno, J. «SMOTE-I: mejora del algoritmo SMOTE para balanceo de clases minoritarias», 2009. [En línea]: https://www.researchgate.net/publication/229045207_SMOTE-I_mejora_ del_algoritmo_SMOTE_para_balanceo_de_clases_minoritarias/_citation/download. [Último acceso: 01/07/2020].

11 Rodríguez Torres, F., «SMOTE-D, una versión determinista de smote», 2017. [En línea]: https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/335/1/RodriguezTF.pdf. [Último acceso: 01/07/2020].

12 Huang, P. «Classication of Imbalanced Data Using Synthetic Oversampling Techniques», 2015. [En línea]: https://escholarship.org/content/qt72w743h7/ qt72w743h7.pdf. [Último acceso: 01/07/2020].

13 Jiawei H. y otros, Data Mining: Concepts and Techniques, 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: 3ra. Edición. (2011). ISBN 978-0-12-381479-1.

14 Farías Concha, M., «Máquinas Vectoriales híbridas para clasificar accidentes de tránsito en la región metropolitana», 2011. [En línea]: http://opac.pucv.cl/pucv_txt/txt-9500/ UCF9980_01.pdf. [Último acceso: 01/07/2020].

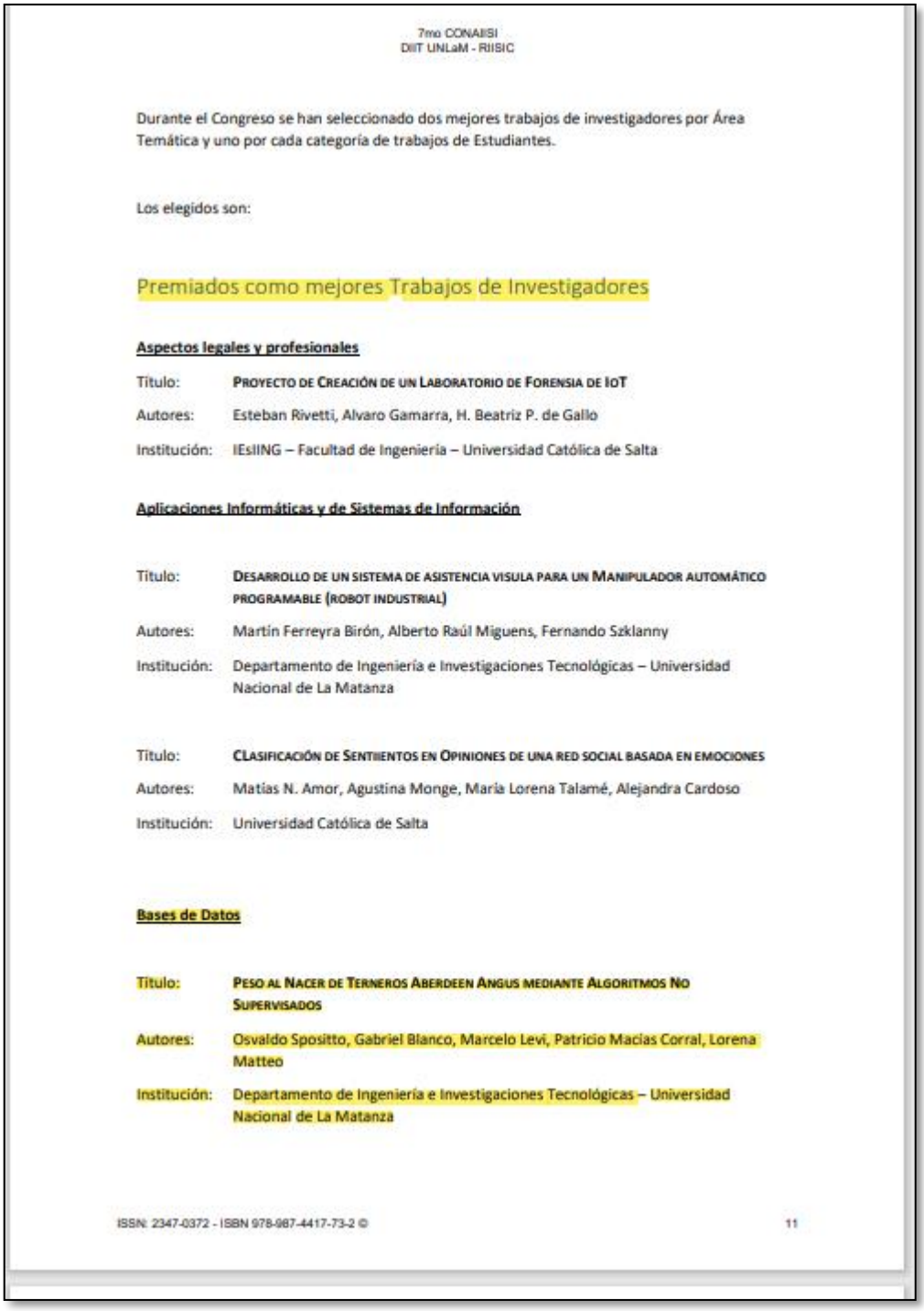
15 Benavides, Ana, «Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones», 2017. [En línea]: https://idus.us.es/bitstream/handle/11441/63201/Valle%20Benavides %20 Ana%20Roc%C3%ADo%20del%20TFG.pdf?sequence=1. [Último acceso: 01/07/2020].

16 Dal Pozzo, A. y otros. «Racing for unbalanced methods selection», 2013. [En línea]: https://www.slideshare.net/dalpozz/racing-for-unbalanced-methods-selection. [Último acceso: 01/07/2020].

B.3. Capítulos de libros

Tapa Libro de Actas Artículo CONAIISI 2019

Artículo CONAIISI 2019 con Medalla de Mención en Área Temática Base de Datos



7mo CONAIISI
DIIT UNLaM - RIISIC

Durante el Congreso se han seleccionado dos mejores trabajos de investigadores por Área Temática y uno por cada categoría de trabajos de Estudiantes.

Los elegidos son:

Premiados como mejores Trabajos de Investigadores

Aspectos legales y profesionales

Título: **PROYECTO DE CREACIÓN DE UN LABORATORIO DE FORENSIA DE IOT**

Autores: Esteban Rivetti, Alvaro Gamarra, H. Beatriz P. de Gallo

Institución: IEsIING – Facultad de Ingeniería – Universidad Católica de Salta

Aplicaciones Informáticas y de Sistemas de Información

Título: **DESARROLLO DE UN SISTEMA DE ASISTENCIA VISUAL PARA UN MANIPULADOR AUTOMÁTICO PROGRAMABLE (ROBOT INDUSTRIAL)**

Autores: Martín Ferreyra Birón, Alberto Raúl Miguens, Fernando Szklanny

Institución: Departamento de Ingeniería e Investigaciones Tecnológicas – Universidad Nacional de La Matanza

Título: **CLASIFICACIÓN DE SENTIIENTOS EN OPINIONES DE UNA RED SOCIAL BASADA EN EMOCIONES**

Autores: Matías N. Amor, Agustina Monge, María Lorena Talamé, Alejandra Cardoso

Institución: Universidad Católica de Salta

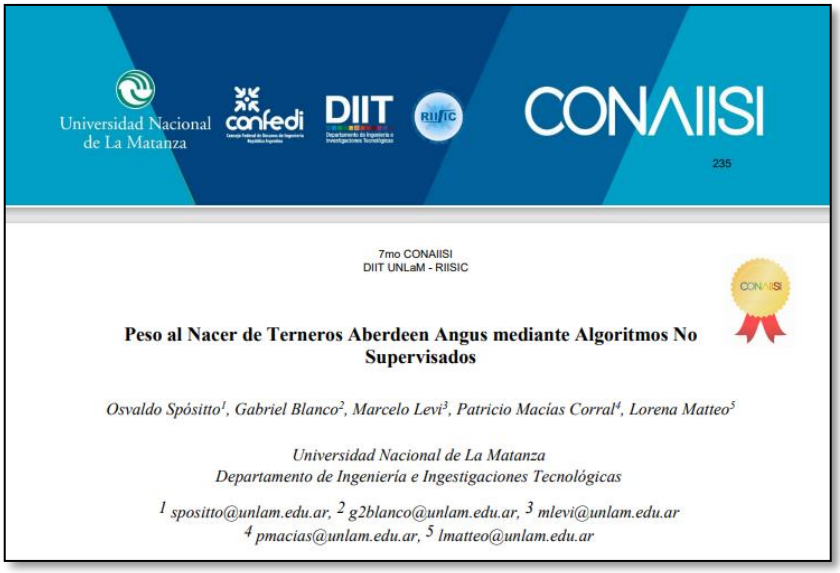
Bases de Datos

Título: **PESO AL NACER DE TERNEROS ABERDEEN ANGUS MEDIANTE ALGORITMOS NO SUPERVISADOS**

Autores: Osvaldo Spositto, Gabriel Blanco, Marcelo Levi, Patricio Macías Corral, Lorena Matteo

Institución: Departamento de Ingeniería e Investigaciones Tecnológicas – Universidad Nacional de La Matanza

ISSN: 2347-0372 - ISBN 978-987-4417-73-2 © 11



Universidad Nacional de La Matanza · confedi · DIIT · RIISIC · CONAIISI

235

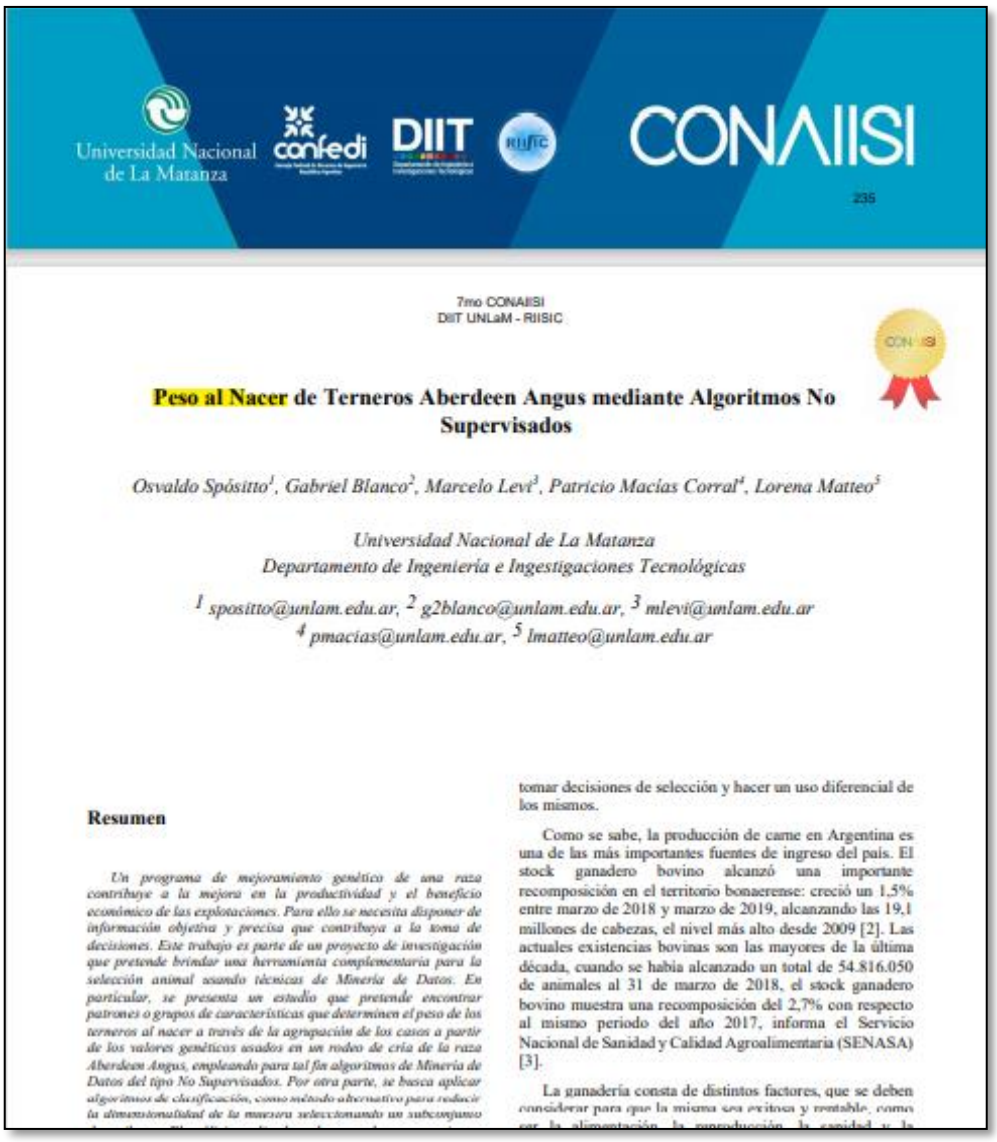
7mo CONAIISI
DIIT UNLaM - RIISIC

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



Universidad Nacional de La Matanza · confedi · DIIT · RIISIC · CONAIISI

235

7mo CONAIISI
DIIT UNLaM - RIISIC

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] spositto@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

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

Comparativa de programas informáticos de Minería de datos

<table>
<tr><td></td><td>rapidminer</td></tr>
<tr><td>Desarrollador</td><td>RapidMiner, Alemania</td></tr>
<tr><td>Tipo de Programa</td><td>Software Libre</td></tr>
<tr><td>Sistema Operativo</td><td>Multiplataforma
Windows
Mac OSX
Linux</td></tr>
<tr><td>Lenguaje de Programación</td><td>Java</td></tr>
<tr><td>Algoritmos</td><td>Algoritmos de aprendizaje automático como: Algoritmo de generación de Reglas, algoritmos probabilísticos</td></tr>
<tr><td>Tipo de Modelo</td><td>Predictivo</td></tr>
<tr><td>Características</td><td>Representación interna de los procesos de análisis de datos en ficheros XML.
Permite el desarrollo de programas a través de un lenguaje de script.
Puede usarse de diversas maneras:
A través de un GUI.
En línea de comandos.
En batch (lotes).
Desde otros programas a través de llamadas a sus bibliotecas.
Extensible.
Incluye gráficos y herramientas de visualización de datos.
Dispone de módulos de integración con R y Python.</td></tr>
<tr><td>Licencia</td><td>
<table>
<tr><td>Tipo / Características</td><td>Educational Program</td><td>RapidMiner Go</td><td colspan="3">RapidMiner Studio</td></tr>
<tr><td></td><td>ML para fines académicos como estudiante o profesor</td><td>ML en una interfaz web guiada y totalmente automatizada</td><td colspan="3">Cree flujos de trabajo de AA en una plataforma de ciencia de datos integral</td></tr>
<tr><td></td><td></td><td></td><td>Gratis</td><td>Profesional</td><td>Empresa</td></tr>
<tr><td>Caducidad de la Licencia</td><td>Gratis con renovación a los 12 meses</td><td>$ 10 al mes</td><td>Gratis Prueba de 30 días Enterprise</td><td>$7.500 por usuario por año</td><td>$15.000 por usuario por año</td></tr>
<tr><td>Filas de datos</td><td>Ilimitado</td><td></td><td>10.000</td><td>100.000</td><td>Ilimitado</td></tr>
<tr><td>Procesador lógico</td><td>Ilimitados</td><td></td><td>1</td><td>2</td><td>Ilimitados</td></tr>
<tr><td>Soporte</td><td>Comunitario</td><td></td><td>Comunitario</td><td>Empresarial</td><td>Empresarial</td></tr>
<tr><td>Preparación turbo</td><td>SI</td><td></td><td>NO</td><td>SI</td><td>SI</td></tr>
<tr><td>Modelo automático</td><td>SI</td><td></td><td>NO</td><td>SI</td><td>SI</td></tr>
<tr><td>Operaciones de modelos automatizadas</td><td>SI</td><td></td><td>NO</td><td>NO</td><td>SI</td></tr>
</table>
</td></tr>
</table>

	rapidminer					
	Ejecución del proceso en segundo plano	SI		NO	NO	SI
Versión Actual	9.4 26 de septiembre de 2019					
Sitio Web	https://rapidminer.com/get-started/					

	WEKA The University of Waikato
Desarrollador	Universidad De Waikato, Nueva Zelanda
Tipo de Programa	Software Libre
Sistema Operativo	Multiplataforma • Windows • Mac OSX • Linux
Lenguaje de Programación	Java
Algoritmos	• Classify: Algoritmos de clasificación, distribuidos por paquetes, como por ejemplo ID3 o C4.5 • Cluster: Diferentes algoritmos de segmentación como el simple k-means. • Associate: Algoritmos para encontrar relaciones de asociación entre variables (Apriori entre otros)
Tipo de Modelo	Predictivo

Características	• Está disponible libremente bajo la licencia pública general de GNU.
	• Es muy portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma.
	• Contiene una extensa colección de técnicas para preprocesamiento de datos y modelado.
	• Puede usarse de diversas maneras:
	  ○ A través de un GUI.
	  ○ En línea de comandos.
	Weka soporta varias tareas estándar de minería de datos, especialmente, preprocesamiento de datos, clustering, clasificación, regresión, visualización, y selección. Todas las técnicas de Weka se fundamentan en la asunción de que los datos están disponibles en un fichero plano (*flat file*) o una relación, en la que cada registro de datos está descrito por un número fijo de atributos (normalmente numéricos o nominales, aunque también se soportan otros tipos). Weka también proporciona acceso a bases de datos vía SQL gracias a la conexión JDBC (*Java Database Connectivity*) y puede procesar el resultado devuelto por una consulta hecha a la base de datos. No puede realizar minería de datos multi-relacional, pero existen aplicaciones que pueden convertir una colección de tablas relacionadas de una base de datos en una única tabla que ya puede ser procesada con Weka.

	
	Un área importante que actualmente no cubren los algoritmos incluidos en Weka es el modelado de secuencias.
Licencia	open source, GNU GPL 3
Versión Actual	3.6.15 (book), 3.8.3 (stable), 3.9 (development)
Sitio Web	https://www.cs.waikato.ac.nz/~ml/weka/

	
Desarrollador	Universidad de Ljubljana, Eslovenia
Tipo de Programa	Software Libre
Sistema Operativo	Multiplataforma • Windows • Mac OS • Linux
Lenguaje de Programación	Python Cython C C++
Algoritmos	k-means, hierarchichal, consensus, Naive Bayes classifier, k-nearest neighbors, classification tres, support vector machines
Tipo de Modelo	Predictivo
Características	Se pueden desarrollar widgets, los cuales pueden ser extendidos como complementos, permitiendo que los componentes y códigos puedan ser reutilizados. Sus funcionalidades: • Visualización interactiva de datos. • Programación visual (interfaz gráfica de usuario). • Adiciones disponibles para minar datos de fuentes de datos externas, minería de texto, análisis de redes. • Recuerda las selecciones y sugiere las combinaciones más utilizadas. • Existen más de 100 widgets que cubren las tareas estandarizadas del análisis de datos
Licencia	open source, GNU GPL 3
Versión Actual	3.26
Sitio Web	https://orange.biolab.si/

	Open for Innovation KNIME
Desarrollador	Fue desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania, bajo la supervisión del profesor Michael Berthold. En la actualidad, la empresa KNIME.com GmbH, radicada en Zúrich, Suiza
Tipo de Programa	Software Libre
Sistema Operativo	Multiplataforma • Linux • Windows • Mac OSX • Cualquier S.O. que soporte máquina virtual Java Windows, Linux, Mac Os
Lenguaje de Programación	Java
Algoritmos	Algoritmo de árboles, algoritmos de MD, algoritmos de aprendizaje automático
Tipo de Modelo	Predictivo
Características	Está desarrollado sobre la plataforma Eclipse y programado, esencialmente, en java. Está concebido como una herramienta gráfica y dispone de una serie de nodos (que encapsulan distintos tipos de algoritmos) y flechas (que representan flujos de datos) que se despliegan y combinan de manera gráfica e interactiva. Los nodos implementan distintos tipos de acciones que pueden ejecutarse sobre una tabla de datos: • Manipulación de filas, columnas, etc., como muestreos, transformaciones, agrupaciones, etc. • Visualización (histogramas, etc.). • Creación de modelos estadísticos y de minería de datos, como árboles de decisión, máquinas de vector soporte, regresiones, etc. • Validación de modelos, como curvas ROC, etc. • *Scoring* o aplicación de dichos modelos sobre conjuntos a nuevos de datos. • Creación de informes a medida gracias a su integración con BIRT. El carácter abierto de la herramienta hace posible su extensión mediante la creación de nuevos nodos que implementen algoritmos a la medida del usuario. Además, existe la posibilidad de utilizar de llamar directa y transparentemente a Weka y o de incorporar de manera sencilla código desarrollado en R o python/jython. KNIME integra diversos componentes para aprendizaje automático y minería de datos a través de su concepto de fraccionamiento de datos (*data pipelining*) modular. La interfaz gráfica de usuario permite el montaje fácil y rápido de nodos para preprocesamiento de datos (ETL: extracción, transformación, carga), para el análisis de datos y modelado y visualización. KNIME es desde 2006 utilizado en la investigación farmacéutica,[1] pero también se utiliza en otras áreas, como: análisis de datos de cliente de CRM, inteligencia de negocio y análisis de datos financieros.

Licencia		Plataforma de análisis KNIME	Servidor pequeño	Servidor mediano	Servidor grande
		Código abierto para crear ciencia de datos	Equipos pequeños intercambian y ejecutan flujos de trabajo	Utilice flujos de trabajo a través del navegador.	Grandes equipos, múltiples instalaciones,

	Open for Innovation KNIME				
			de forma re-mota.	Acceso a la API REST	colaboración global.
	Suscripción anual (basada en 5 usuarios y 8 núcleos)	Gratis	Servidor KNIME para Azure A partir de 2,07 US $ / hora Servidor KNIME para AWS 2,07 US $ / hora	25.000 EUR 29.000 USD Anual	45.500 EUR 52.000 USD Anual
Versión Actual	Plataforma de análisis KNIME 4.2.0 Servidor pequeño KNIME Server Small 4.10 para Azure				
Sitio Web	https://www.knime.com/				

	SAS THE POWER TO KNOW.		
Desarrollador	**SAS** Institute		
Tipo de Programa	Software no Libre		
Sistema Operativo	• Windows • mainframe de IBM • Unix/Linux • OpenVMS Alpha		
Lenguaje de Programación	C		
Algoritmos	Modelado descriptivo y predictivo: k-means, diagramas de dispersión y estadísticas de resumen detalladas.		
Tipo de Modelo	Predictivo, descriptivo		
Características	• Modelado descriptivo y predictivo: Utiliza técnicas de aprendizaje automático para crear modelos predictivos desde una interfaz visual o de programación. • Desarrollo de modelos abiertos basados en código: Se accede desde otros lenguajes (Python, R, Lua, Java) o se utiliza API REST públicas para agregar SAS Analytics a las aplicaciones existentes. • Capacidades de alto rendimiento. • Una manera rápida, fácil y autosuficiente para que los usuarios de negocios generen modelos. • Procesamiento escalable • Preparación, resumen y exploración de datos sofisticados.		
Licencia		Educación	Empresa
	Suscripción anual	Gratis	Con costo Sin información
Versión Actual	5		
Sitio Web	https://www.sas.com/en_us/home.html		

	
Desarrollador	Desarrollador(es) R Development Core Team
Tipo de Programa	Software Libre
Sistema Operativo	• Windows • Macintosh • Unix • GNU/Linux
Lenguaje de Programación	C, C++, Fortran, R
Algoritmos	
Tipo de Modelo	R proporciona un amplio abanico de herramientas estadísticas (modelos lineales y no lineales, tests estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento, etc.) y gráficas una Herramienta Útil.
Características	R hereda de S su orientación a objetos. La tarea de extender R se ve facilitada por su permisiva política de lexical scoping. Además, R puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python. Otra de las características de R es su capacidad gráfica, que permite generar gráficos con alta calidad. R posee su propio formato para la documentación basado en LaTeX. R también puede usarse como herramienta de cálculo numérico, campo en el que puede ser tan eficaz como otras herramientas específicas tales como GNU Octave y su equivalente privativo: MATLAB. Se ha desarrollado una interfaz, RWeka para interactuar con Weka que permite leer y escribir ficheros en el formato arff y enriquecer R con los algoritmos de minería de datos de dicha plataforma.
Licencia	Software libre, GNU GPL 2+
Versión Actual	R-4.0.2
Sitio Web	https://www.r-project.org/

	
Desarrollador	Artelnics
Tipo de Programa	Software no Libre
Sistema Operativo	• Microsoft Windows • OS X • Linux
Lenguaje de Programación	C++
Algoritmos	

	neuraldesigner
Tipo de Modelo	
Características	Es una herramienta de aprendizaje automático de propósito general. Permite arquitecturas con múltiples capas de procesamiento no lineal y contiene utilidades para resolver problemas de regresión de funciones, reconocimiento de patrones, series temporales y auto-asociación. La entrada al programa es un conjunto de datos, y la salida es su correspondiente modelo predictivo. El software permite exportar la expresión matemática de la red neuronal con el fin de ser usada en cualquier lenguaje de programación o sistema informático.

Licencia

	Académicos			Estándar			
	Investigar	Estudiante	Universidad	Gratis	Pequeña	Medio	Grande
Filas de datos	Ilimitado	100.000	100.000	10	100.000	1.000.000	Ilimitado
Escritorio	Si	Si	Si	Si	Si	Si	Si
Nube	AWS (BYOL)	AWS (BYOL)	AWS (BYOL)	No	AWS (BYOL)	AWS (BYOL)	AWS (BYOL)
Soporte	Estándar	Estándar	Estándar	Básico	Estándar	Empresa	Premium
Precio	Anual: $ 995 De por vida: $ 2,495	Contactarse	Contactarse	$ 0	Anual: $ 995 De por vida: $ 2,495	Anual: $ 2,495 Vida útil: $ 6,245	Anual: $ 4,995 Vida útil: $ 12,495

Versión Actual	Windows: 4.2.0 macOS: 4.1.0 Ubuntu: 2.9.9
Sitio Web	https://www.neuraldesigner.com/

	SPSS IBM
Desarrollador	IBM (International Business Machines Corporation)
Tipo de Programa	Software Comercial
Sistema Operativo	Multiplataforma • Windows. • Mac OS. • Linux. • Unix.
Lenguaje de Programación	Java
Algoritmos	Ecuaciones estructurales
Tipo de Modelo	Predictivo

	
Características	• Soporta varias fuentes de datos como es leer datos de archivos planos, hojas de cálculo y principales bases de datos relacionales. • Posee un interfaz gráfica fácil de usar e intuitiva para visualizar proceso de minería de datos. • Otorga modelos predictivos más precisos. Modelo automatizado. • Variedad de métodos algorítmicos. • Excelente análisis de texto.
Licencia	Prueba gratuita de 14 días y Licencias Base, Standard, Professional, Premium par costos hay que contactarse
Versión Actual	26.0
Sitio Web	https://www.ibm.com/products/spss-statistics

Aplicaciones ganaderas de Argentina y en el exterior

Tras una búsqueda de software tanto de gestión ganadera como de mejoras en la calidad de los progenitores encontramos varios que se diferencian entre si más que nada en el modularidad y la cohesión, ninguno por ser cerrados menciona si utiliza IA para obtener indicadores. Se encontraron algunos originarios de la provincia de Santa Fe, otros de Córdoba y finalmente otro que se utiliza en el nordeste de Brasil.
En primera instancia y con mayor cantidad de módulos de gestión siendo el módulo que nos interesa **CRIADORES REGISTRADOS**, el mismo podemos encontrar detalles en http://www.softhuella.com.ar/especificaciones/?opcion=productores y aparentemente condice con ciertas características buscadas en la selección para reproducción.
También contamos en Argentina con la herramienta **Visual Tambo**, que no es más que un módulo de Visual productos http://visualsystemas.com.ar/visual-tambo.php y está orientado a productores, médicos veterinarios que asesoran a productores y agrupaciones de productores o cooperativas. Contempla selección de progenitores y momento de preñez.
Tambero, con información relevante en https://www.tambero.com/es indicando que se usa en gran variedad de paises. Otra para llevar el **Control Ganadero** de Colombia, https://www.controlganadero.co/
Una guía denominada **"Cuaderno de campo para vacunos"** para orientar el registro de la actividad detallado en el siguiente documento del INTA, https://inta.gob.ar/sites/default/files/script-tmp-cuadernovacunos.pdf
A su vez se encontró una **aplicación del RFID** para aumentar el control y mejorar la gestión en la producción ganadera cuyo detalle puede verse en el siguiente paper:
https://ri.itba.edu.ar/bitstream/handle/123456789/959/K26%20-%20Aplicaci%C3%B3n%20del%20RFID%20para%20aumentar%20el%20control%20y%20mejorar%20la%20gesti%C3%B3n%20en%20la%20producci%C3%B3n%20ganadera.pdf?sequence=1&isAllowed=y
Por último, la detección más importante es del software **Bertha**, de origen americano USA, pero muy utilizado también en Brasil, siendo el más completo, tanto en versión web como aplicación móvil, como se aprecia en su página, https://gobertha.com/ dando un completísimo tablero de control y seguimiento.

Certificado de Cursos y Seminarios:

- **Certificados de Escritura Científica I y II**

Certificado. Curso de Escritura Científica I 1

Marca para seguimiento.

cyt
Vie 31/7/2020 16:00
Para: LORENA MATTEO

Matteo, Lorena Romina. Certi...
421 KB

Buenas tardes.
Le hacemos llegar por este medio el certificado correspondiente al Curso de Escritura Científica I.

Muchas gracias.
Saludos cordiales,
Damián

Secretaría de Ciencia y Tecnología
Universidad Nacional de La Matanza
Florencio Varela 1903, San Justo, La Matanza
T.E.: 4480-8900 int. 8871/ 8872
Sitio web: http://cyt.unlam.edu.ar
Facebook: Secretaría de Ciencia y Tecnología UNLaM

Curso de Escritura Científica I

Se certifica que

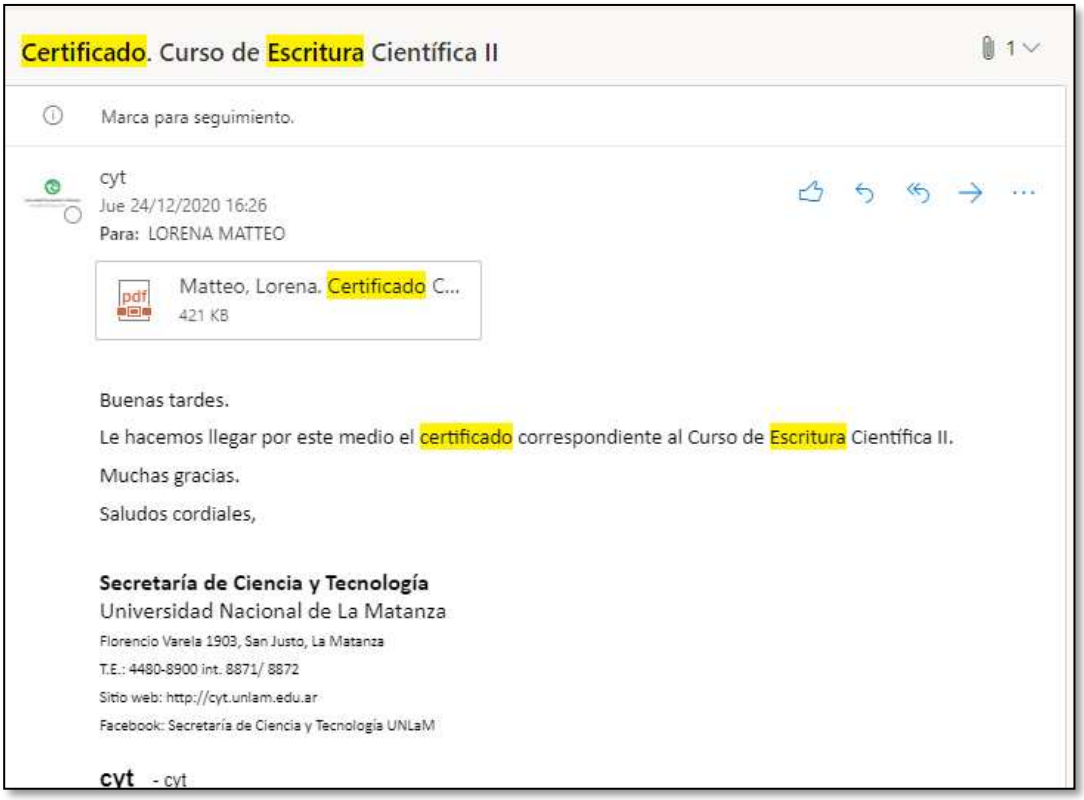
Matteo, Lorena Romina

DNI: 23.701.391

aprobó el *Curso de Escritura Científica I*, organizado por la Secretaría de Ciencia y Tecnología de la Universidad Nacional de La Matanza, con una duración total de 30 horas.

San Justo, jueves 30 de julio de 2020.

Mg. Ana Bidiña
Secretaria de Ciencia y Tecnología
UNLaM



Certificado. Curso de Escritura Científica II

1

Marca para seguimiento.

cyt
Jue 24/12/2020 16:26
Para: LORENA MATTEO

Matteo, Lorena. Certificado C...
421 KB

Buenas tardes.

Le hacemos llegar por este medio el certificado correspondiente al Curso de Escritura Científica II.

Muchas gracias.

Saludos cordiales,

Secretaría de Ciencia y Tecnología
Universidad Nacional de La Matanza
Florencio Varela 1903, San Justo, La Matanza
T.E.: 4480-8900 int. 8871/ 8872
Sitio web: http://cyt.unlam.edu.ar
Facebook: Secretaría de Ciencia y Tecnología UNLaM

cyt - cyt

Curso de Escritura Científica II

Se certifica que

Matteo, Lorena

DNI: 23.701.391

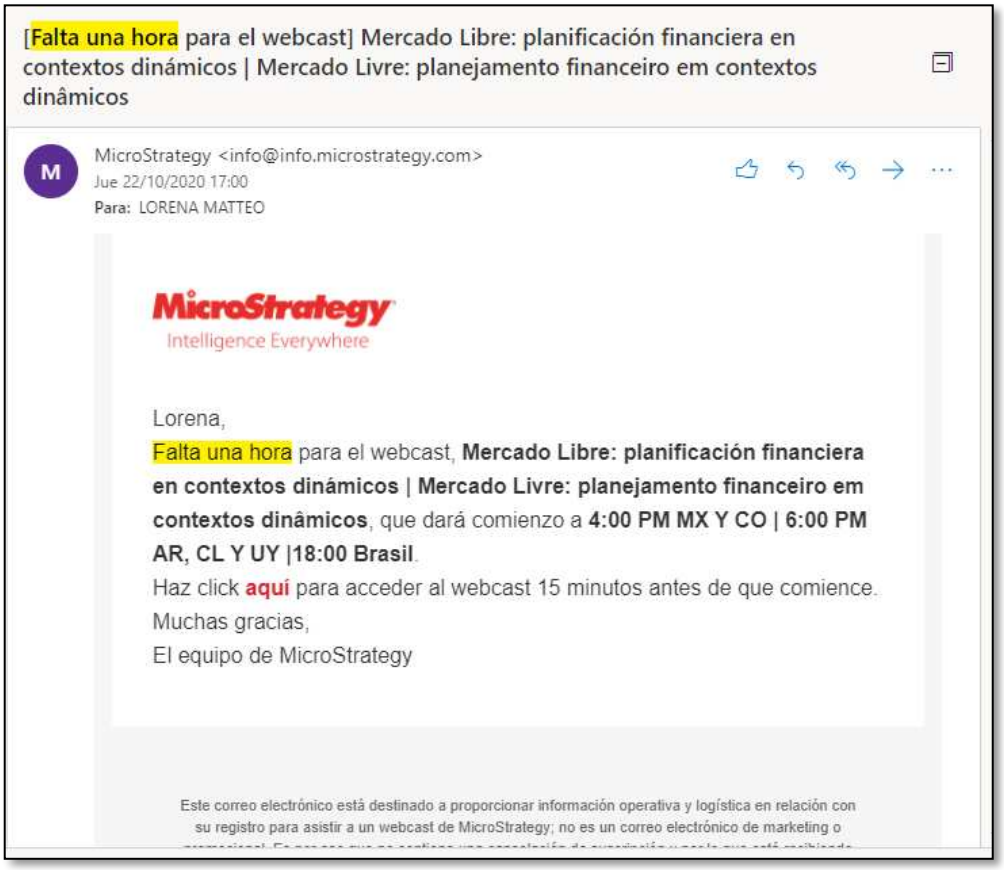
aprobó el *Curso de Escritura Científica II*, organizado por la Secretaría de Ciencia y Tecnología de la Universidad Nacional de La Matanza, con una duración total de 20 horas.

San Justo, martes 22 de diciembre de 2020.

Mg. Ana Bidiña
Secretaria de Ciencia y Tecnología
UNLaM

- **Amazon Web Services (AWS) – Inteligencia Artificial para Creadores**

- **Webinars Microstrategy 2020**



[Falta una hora para el webcast] Cómo usar analítica avanzada para predecir la deserción de estudiantes

MicroStrategy
Jue 1/10/2020 15:00
Para: LORENA MATTEO

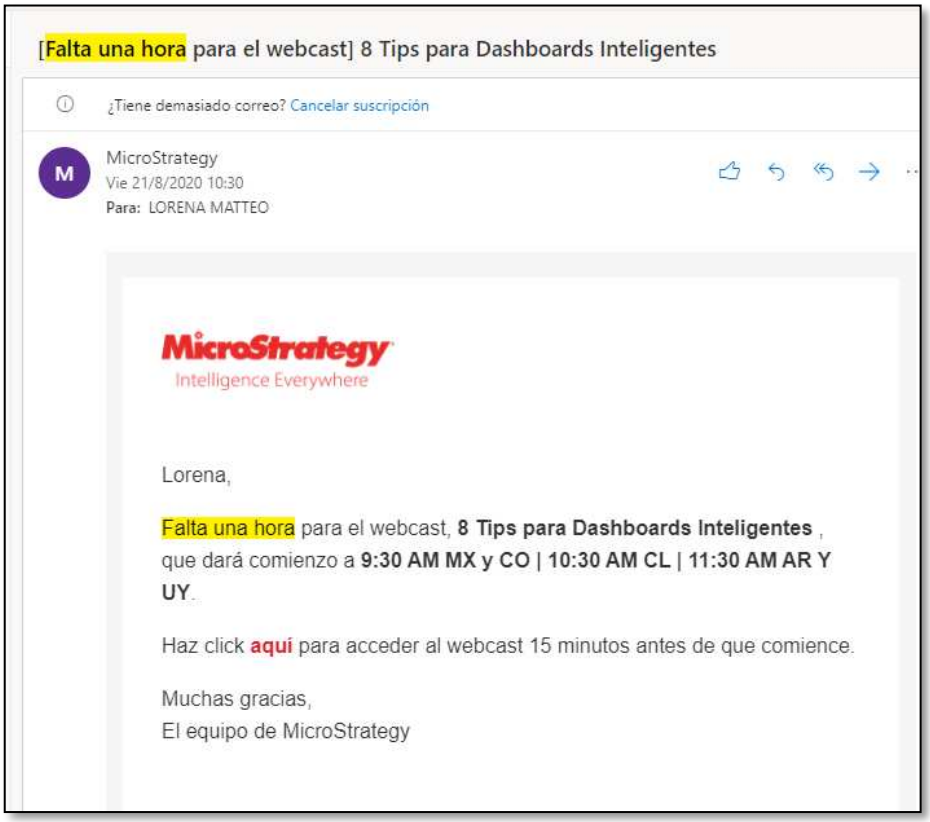
MicroStrategy
Intelligence Everywhere

Lorena,

Falta una hora para el webcast **Cómo usar analítica avanzada para predecir la deserción de estudiantes**, que comienza a las **4 PM AR Y CL**.

Haz click **aquí** para acceder al webcast 15 minutos antes de que comience.

Muchas gracias,
El equipo de MicroStrategy



[Falta una hora para el webcast] 8 Tips para Dashboards Inteligentes

¿Tiene demasiado correo? Cancelar suscripción

MicroStrategy
Vie 21/8/2020 10:30
Para: LORENA MATTEO

MicroStrategy
Intelligence Everywhere

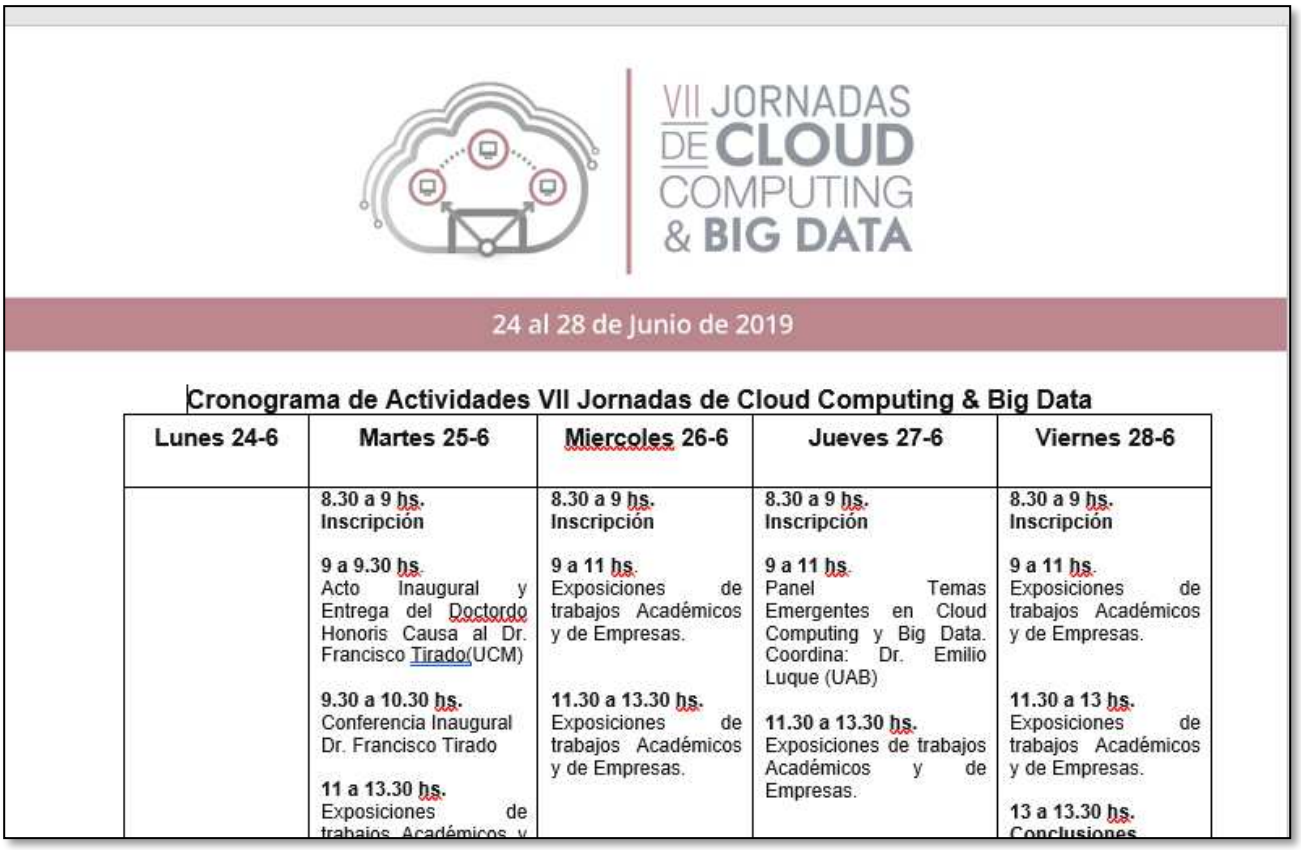
Lorena,

Falta una hora para el webcast, **8 Tips para Dashboards Inteligentes** , que dará comienzo a **9:30 AM MX y CO | 10:30 AM CL | 11:30 AM AR Y UY**.

Haz click **aquí** para acceder al webcast 15 minutos antes de que comience.

Muchas gracias,
El equipo de MicroStrategy

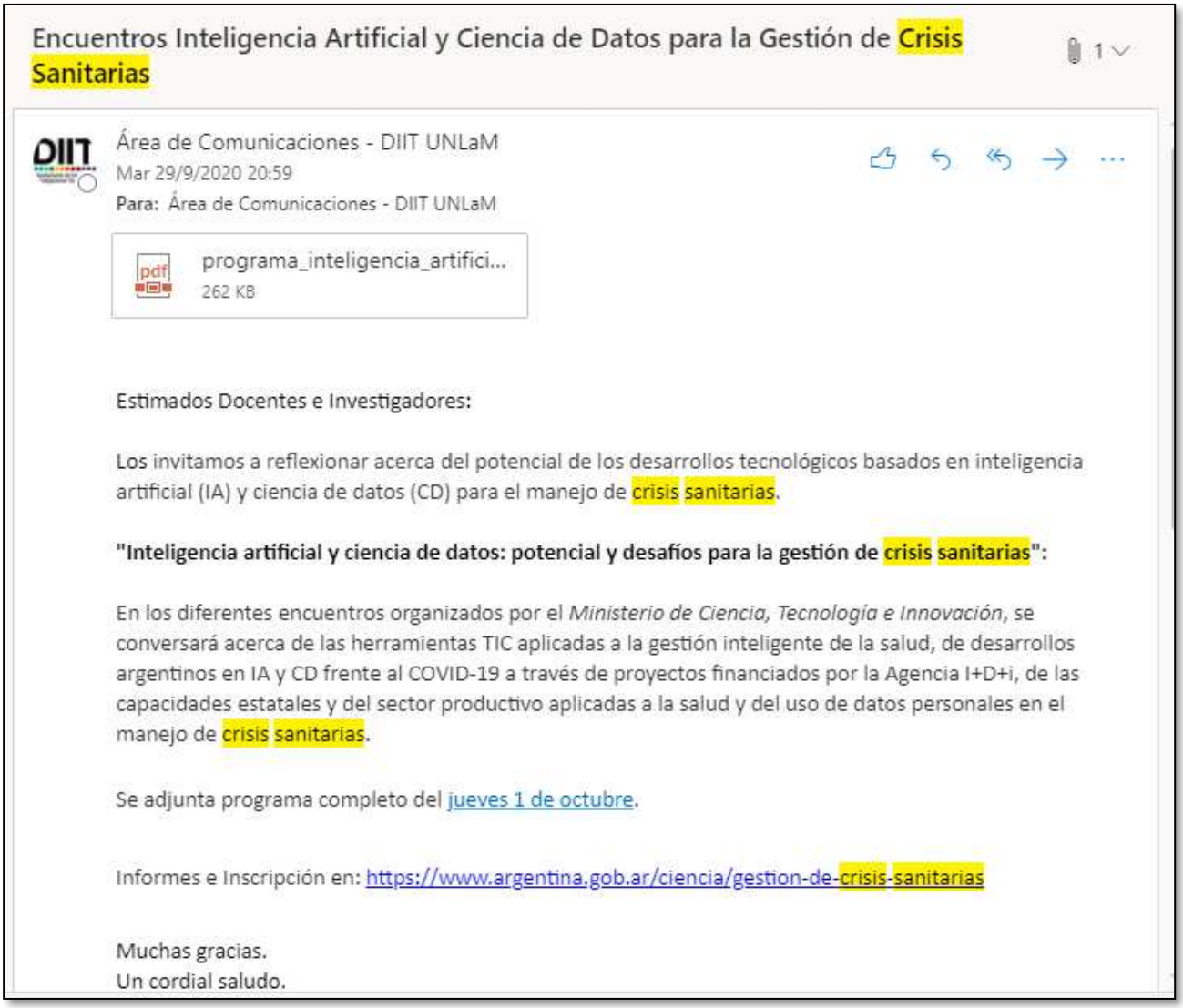
- **Cronograma de Actividades VII Jornadas de Cloud Computing & Big Data**



Cronograma de Actividades VII Jornadas de Cloud Computing & Big Data

Lunes 24-6	Martes 25-6	Miercoles 26-6	Jueves 27-6	Viernes 28-6
	8.30 a 9 hs. **Inscripción** **9 a 9.30 hs.** Acto Inaugural y Entrega del Doctordo Honoris Causa al Dr. Francisco Tirado(UCM) **9.30 a 10.30 hs.** Conferencia Inaugural Dr. Francisco Tirado **11 a 13.30 hs.** Exposiciones de trabajos Académicos y	**8.30 a 9 hs.** **Inscripción** **9 a 11 hs.** Exposiciones de trabajos Académicos y de Empresas. **11.30 a 13.30 hs.** Exposiciones de trabajos Académicos y de Empresas.	**8.30 a 9 hs.** **Inscripción** **9 a 11 hs.** Panel Temas Emergentes en Cloud Computing y Big Data. Coordina: Dr. Emilio Luque (UAB) **11.30 a 13.30 hs.** Exposiciones de trabajos Académicos y de Empresas.	**8.30 a 9 hs.** **Inscripción** **9 a 11 hs.** Exposiciones de trabajos Académicos y de Empresas. **11.30 a 13 hs.** Exposiciones de trabajos Académicos y de Empresas. **13 a 13.30 hs.** **Conclusiones**

- **Seminario "Inteligencia artificial y ciencia de datos: potencial y desafíos para la gestión de crisis sanitarias"**
    - Ciclo de encuentros: Inteligencia artificial y ciencia de datos, potencial y desafíos para la gestión de crisis sanitarias | Argentina.gob.ar vía Zoom, jueves 1, 8, 15, 22 y 29 de octubre de 2020, de 16 a 18 horas.



Encuentros Inteligencia Artificial y Ciencia de Datos para la Gestión de Crisis Sanitarias

1

Área de Comunicaciones - DIIT UNLaM
Mar 29/9/2020 20:59
Para: Área de Comunicaciones - DIIT UNLaM

programa_inteligencia_artifici...
262 KB

Estimados Docentes e Investigadores:

Los invitamos a reflexionar acerca del potencial de los desarrollos tecnológicos basados en inteligencia artificial (IA) y ciencia de datos (CD) para el manejo de crisis sanitarias.

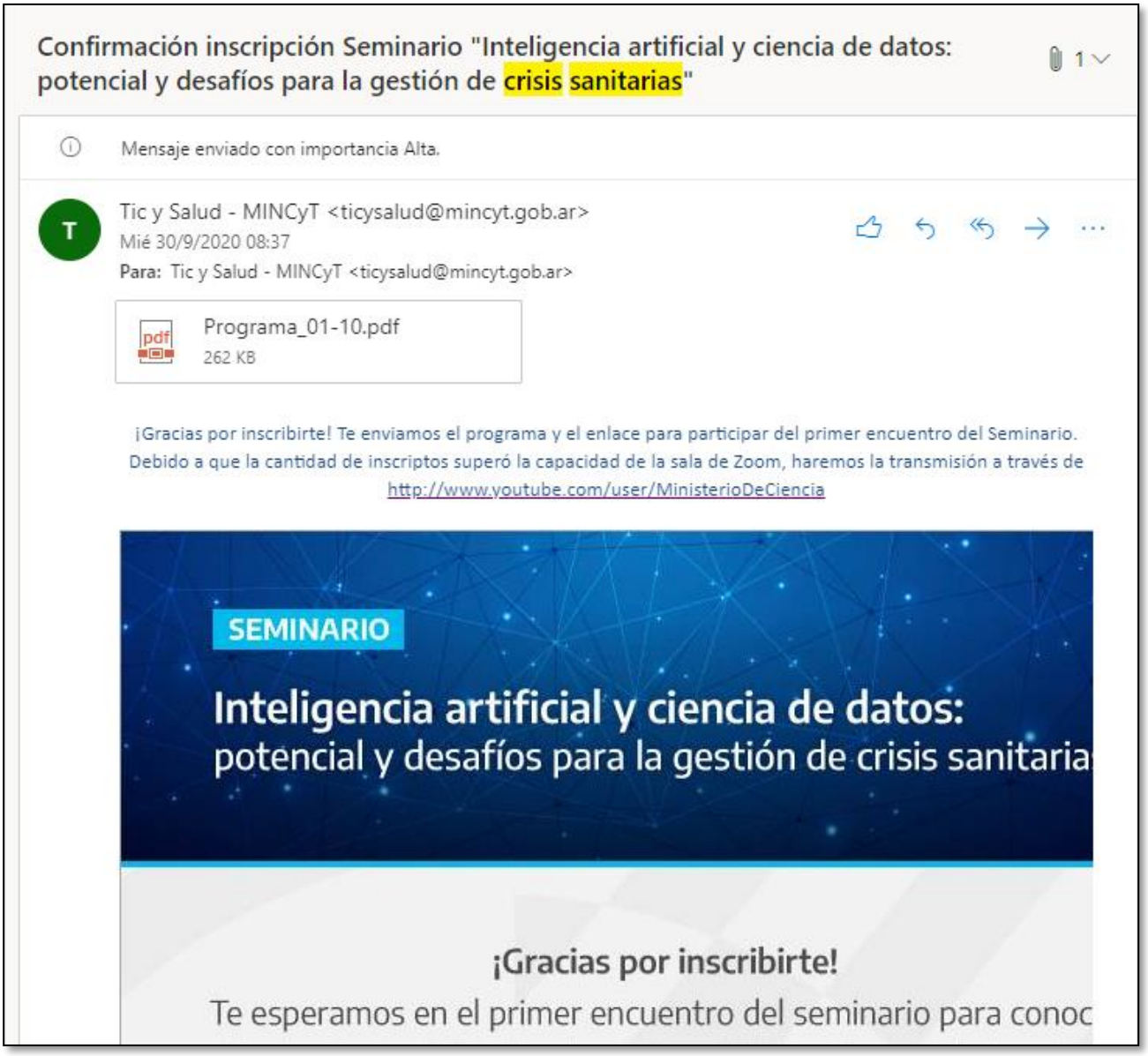
"Inteligencia artificial y ciencia de datos: potencial y desafíos para la gestión de crisis sanitarias":

En los diferentes encuentros organizados por el *Ministerio de Ciencia, Tecnología e Innovación*, se conversará acerca de las herramientas TIC aplicadas a la gestión inteligente de la salud, de desarrollos argentinos en IA y CD frente al COVID-19 a través de proyectos financiados por la Agencia I+D+i, de las capacidades estatales y del sector productivo aplicadas a la salud y del uso de datos personales en el manejo de crisis sanitarias.

Se adjunta programa completo del jueves 1 de octubre.

Informes e Inscripción en: https://www.argentina.gob.ar/ciencia/gestion-de-crisis-sanitarias

Muchas gracias.
Un cordial saludo.


Confirmación inscripción Seminario "Inteligencia artificial y ciencia de datos: potencial y desafíos para la gestión de crisis sanitarias"
1
Mensaje enviado con importancia Alta.
T
Tic y Salud - MINCyT <ticysalud@mincyt.gob.ar>
Mié 30/9/2020 08:37
Para: Tic y Salud - MINCyT <ticysalud@mincyt.gob.ar>
Programa_01-10.pdf
262 KB
¡Gracias por inscribirte! Te enviamos el programa y el enlace para participar del primer encuentro del Seminario. Debido a que la cantidad de inscriptos superó la capacidad de la sala de Zoom, haremos la transmisión a través de http://www.youtube.com/user/MinisterioDeCiencia
SEMINARIO
Inteligencia artificial y ciencia de datos:
potencial y desafíos para la gestión de crisis sanitaria
¡Gracias por inscribirte!
Te esperamos en el primer encuentro del seminario para conoc


CICLO DE ENCUENTROS
Inteligencia artificial y ciencia de datos:
potencial y desafíos para la gestión de crisis sanitarias.
Esta semana te esperamos para dialogar sobre las capacidades estatales en inteligencia artificial y ciencia de datos para la gestión de crisis sanitarias.
Jueves 22 de octubre de 16.00 a 18.00 h
LINK A YouTube
Fundación SADOSKY
Agencia I+D+i
Ministerio de Salud
Ministerio de Ciencia, Tecnología e Innovación
Argentina

MINISTERIO DE EDUCACIÓN
Secretaría de Políticas Universitarias
Programa de Incentivos a Docentes - Investigadores

PRESENTACION DEL INFORME DE PROYECTOS DE INVESTIGACION
Final PROINCE 2019

CONVOCATORIA: **Final PROINCE 2019**

APELLIDO Y NOMBRES: **GIULIANO , MONICA**

TIPO Y NRO DE DOCUMENTO: **DNI 17770124**

COMISIÓN EVALUADORA: **Ingeniería**

....

Lugar y Fecha **Firma del titular del Proyecto**

10020200400007LM