

INFORME TÉCNICO: MINERÍA DE DATOS APLICADA AL SISTEMA SIO-CARNES

Análisis de Reglas de Asociación en Operaciones de Compra-Venta de Ganado Vacuno

Proyecto: Minería de Datos - Sistema SIO-Carnes

Institución: Universidad Tecnológica Nacional - Facultad Regional Paraná

Fecha: 21-06-2025

Alumno: Claudio Sebastián Castillo

PRESENTACIÓN

En el marco del creciente interés por aplicar técnicas de minería de datos al análisis de actividades agropecuarias en Argentina, este documento presenta un trabajo de análisis sobre la información generada por el sistema SIO-Carnes, una base de datos que registra las operaciones de compra-venta de ganado vacuno destinado a faena. Durante un período que abarca desde agosto de 2020 hasta agosto de 2021, este sistema registró y ofrece a través de una plataforma pública 1.7 millones de transacciones comerciales, ofreciendo una vista única de los patrones de comercialización del sector ganadero argentino.

El desafío principal de este proyecto consistió en transformar un dataset masivo y heterogéneo en información accionable mediante trabajos de limpieza, transformación y descubrimiento de reglas de asociación. Este proceso requirió actividades de preparación, que incluyeron limpieza de datos, eliminación de outliers, transformación de variables y discretización de atributos continuos. La complejidad de esta tarea se refleja en los números: de los 1.7 millones de registros iniciales, logramos obtener un dataset depurado de 709005 registros válidos, representando un aprovechamiento del 41.19% de los datos originales.

Los resultados obtenidos revelaron patrones interesantes del mercado ganadero argentino. Mediante la aplicación de los algoritmos Apriori y FP-Growth, identificamos múltiples itemsets frecuentes que dieron origen a reglas de asociación estadísticamente significativas. Estas reglas nos permitieron descubrir, por ejemplo, que las operaciones pequeñas (1-7.6 cabezas) dominan el mercado con un 70.9% de participación, o que existe una fuerte correlación geográfica entre ciertas razas bovinas y zonas específicas del país. En lo que sigue presentaremos nuestro trabajo.

1. INTRODUCCIÓN Y CONTEXTO

1.1 El Universo de Datos SIO-Carnes

El Sistema de Información de Operaciones sobre Carnes (SIO-Carnes) representa una fuente rica y completa sobre el mercado ganadero argentino. Esta base de datos gubernamental, alimentada por las liquidaciones electrónicas de AFIP y los Documentos de Tránsito Electrónicos de SENASA, captura el pulso diario de un sector económico fundamental para Argentina.

Durante el período analizado, que se extiende desde agosto de 2020 hasta agosto de 2021, el sistema registró un total de 388 días únicos de actividad comercial. Este período, particularmente relevante por coincidir con los efectos de la pandemia de COVID-19, nos ofrece una perspectiva única sobre el comportamiento del mercado ganadero en tiempos de incertidumbre económica.

El dataset que constituye el núcleo de nuestro análisis presenta dimensiones considerables, donde la información está estructurada en formato CSV con codificación UTF-8, conteniendo 1721218 registros distribuidos en 11 variables fundamentales. Esta información abarca la totalidad del territorio argentino, con representación de las 24 provincias del país y su clasificación en 10 zonas SENASA, proporcionando una cobertura geográfica completa del mercado nacional.

Característica	Valor
Período analizado	Agosto 2020 - Agosto 2021
Días únicos de operaciones	388
Registros totales	1,721,218
Variables	11
Tamaño del archivo	263 MB

Cobertura geográfica	24 provincias argentinas
Zonas de destino	10 zonas SENASA

1.2 La Búsqueda de Patrones Ocultos: Objetivos del Análisis

En este espacio de problema que codifica información del mercado ganadero argentino, nuestro objetivo consistió en identificar patrones de asociación entre variables categóricas mediante la aplicación de algoritmos especializados en reglas de asociación. Esta tarea, aparentemente simple, encierra una complejidad considerable cuando se trata de extraer conocimiento accionable de más de un millón de transacciones comerciales.

El proceso de descubrimiento de patrones requirió un enfoque metodológico riguroso que se estructuró en cinco pasos. Primero, emprendimos un análisis exploratorio para caracterizar la estructura subyacente de los datos, revelando las peculiaridades, inconsistencias y oportunidades que presentaba el dataset original. Posteriormente, desarrollamos un flujo de trabajo para la preparación de datos, esencial para garantizar la calidad de los análisis posteriores. La implementación de los algoritmos Apriori y FP-Growth constituyó el corazón técnico del proyecto, permitiéndonos descubrir itemsets frecuentes que revelan asociaciones significativas entre diferentes variables. A partir de estos itemsets, generamos y evaluamos reglas de asociación utilizando métricas estándar como soporte, confianza y lift. Finalmente, el verdadero valor del análisis emergió en la interpretación de estos resultados desde una perspectiva de negocio ganadero, traduciendo hallazgos estadísticos en conocimiento accionable.

2. METODOLOGÍA

2.1 Ambiente de trabajo

La implementación exitosa de técnicas de minería de datos sobre un dataset de la magnitud de SIO-Carnes requirió la construcción de un entorno de trabajo. El ecosistema de herramientas seleccionado refleja las mejores prácticas en ciencia de datos contemporánea. Pandas se convirtió en nuestra herramienta fundamental para la manipulación y análisis de datos estructurados, mientras que NumPy nos proporcionó la base para operaciones numéricas y matriciales de alto rendimiento. La integración de scikit-learn resultó esencial para las tareas de preprocesamiento y discretización, complementándose perfectamente con *mlxtend*, una librería

especializada que nos permitió implementar los algoritmos de reglas de asociación con precisión y eficiencia.

Para el análisis y presentación de resultados, combinamos Jupyter Notebook y VSCode como entornos de desarrollo, aprovechando la interactividad del primero para la exploración inicial y la robustez del segundo para el desarrollo de código de producción.

2.2 El Flujo de trabajo en Cinco Etapas

El diseño de nuestro flujo de trabajo de procesamiento siguió un patrón de transformación incremental, donde cada etapa construye sobre los resultados de la anterior, garantizando trazabilidad y reproducibilidad en todo el proceso. Esta arquitectura secuencial de cinco etapas se diseñó para maximizar la calidad de los datos finales mientras minimiza la pérdida de información valiosa.

La primera etapa, el análisis exploratorio de datos (EDA), funcionó como nuestro sistema de reconocimiento, revelando las características del dataset e identificando desafíos potenciales. La segunda etapa se concentró en la limpieza y filtrado, eliminando inconsistencias y datos corruptos que podrían comprometer los análisis posteriores.

La tercera etapa involucró transformaciones y la creación de variables derivadas que capturan relaciones más complejas en los datos. La cuarta etapa implementó técnicas de discretización para convertir variables continuas en categóricas, una transformación crítica para la aplicación de reglas de asociación. Finalmente, la quinta etapa aplicó los algoritmos de descubrimiento de patrones, coronando el proceso con la generación de insights accionables.

2.3 Marcos de Evaluación

La implementación de criterios de evaluación rigurosos constituyó un pilar fundamental de nuestra metodología. Así, procuramos ser rigurosos en la calidad de los datos para extracción de información, descartando observaciones que pudieran sesgar los resultados obtenidos (en especial aplicamos un criterio riguroso para eliminar valores duplicados).

En cuanto a las reglas de asociación, adoptamos las métricas estándar de la literatura académica. El **soporte** cuantifica la frecuencia relativa del itemset en el dataset, respondiendo a la pregunta: ¿qué tan común es esta combinación? La **confianza** mide la probabilidad condicional, indicando qué tan probable es que ocurra B dado que ocurrió A. El **lift** evalúa si la asociación es mejor que la casualidad, con valores superiores a 1 indicando asociaciones positivas. Finalmente, desarrollamos un **score combinado** que proporciona una medida integral de la calidad de cada regla.

3. ANÁLISIS EXPLORATORIO

Nuestra radiografía inicial del SIO-Carnes mostró una estructura compleja pero bien definida: 1,721,218 observaciones distribuidas en 11 variables. La composición del dataset revela una característica fundamental: el predominio de variables categóricas sobre numéricas. Con 9 variables de tipo y apenas 2 variables numéricas, nos encontramos ante un dataset naturalmente orientado hacia el análisis categórico.

Un aspecto particularmente relevante es la densidad de datos del 89.58%, calculada considerando valores faltantes. Este porcentaje, aunque respetable, anticipa desafíos en el proceso de limpieza que deberán abordarse cuidadosamente para evitar pérdidas excesivas de información.

El análisis de la cardinalidad de variables categóricas reveló un espectro fascinante de diversidad que va desde categorías altamente específicas hasta clasificaciones más generales. En el extremo de alta cardinalidad encontramos variables **Partido Origen** con 350 valores únicos, evidenciando la granularidad geográfica de los datos.

Particularmente interesante resulta el caso de **Precio Cabeza** y **Precio Kg**, que con ~15 mil y ~78 mil valores únicos respectivamente, muestran la enorme variabilidad en las condiciones comerciales del mercado ganadero. Esta alta cardinalidad, sin embargo, plantea desafíos para el análisis de reglas de asociación, sugiriendo la necesidad de técnicas de discretización sofisticadas.

En el rango de cardinalidad moderada, variables como **Provincia Origen** (24 valores), **Raza** (23 valores) y **Categoría** (33 valores) ofrecen un equilibrio ideal para el análisis de patrones. Estas variables proporcionan suficiente diversidad para generar insights significativos sin crear una complejidad combinatoria que complique la interpretación de resultados.

Las variables de baja cardinalidad, como **Zona Destino** (10 valores) y **Unidad de Medida** (3 valores), proporcionan marcos de referencia claros y estables para la interpretación de patrones más complejos.

El análisis de las variables numéricas reveló un panorama de contrastes extremos que caracteriza la realidad del mercado ganadero argentino. La variable **Cabezas Comercializadas** presenta quizás el caso de mayor variabilidad con un rango que va desde una sola cabeza hasta valores extremos, evidenciando la coexistencia de operaciones familiares microscópicas junto a transacciones corporativas masivas.

La diferencia entre la media (11.47 cabezas) y la mediana (4.00 cabezas) nos cuenta una historia reveladora sobre la estructura del mercado. Esta disparidad sugiere que la mayoría de las operaciones son pequeñas, pero las grandes transacciones ejercen una influencia desproporcionada en los promedios. La variable Cantidad de Kg presenta un patrón similar pero menos extremo.

El análisis de calidad de datos reveló tres categorías principales de problemas que requerirían atención prioritaria en las etapas posteriores del flujo de trabajo. El primer y más crítico desafío lo constituyen los valores faltantes, donde Precio cabeza emerge como la variable más problemática con un 94.58% de valores ausentes. Esta tasa de incompletitud, prácticamente absoluta, sugiere problemas sistemáticos en la recolección de datos o diferencias en los métodos de registro entre diferentes fuentes.

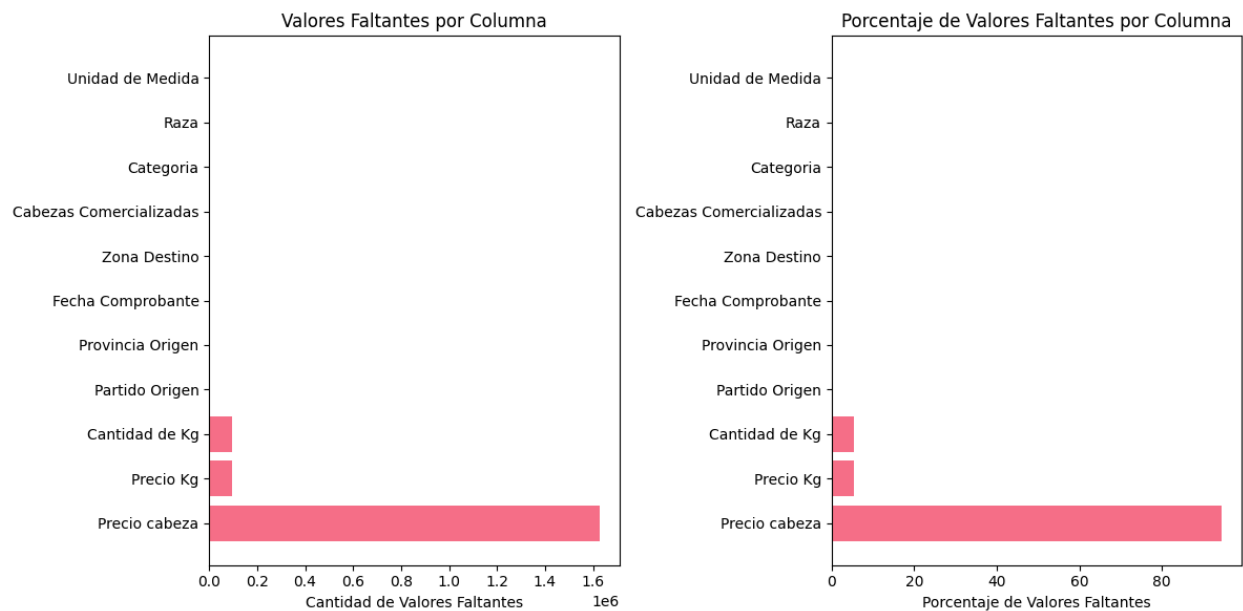


Gráfico 1 - *Valores Faltantes por columna*

Las variables Precio Kg y Cantidad de Kg muestran un comportamiento relacionado con 93,286 valores faltantes cada una (5.42%), sugiriendo que estos registros corresponden a transacciones registradas bajo un sistema diferente de medición, probablemente el sistema de precio por cabeza que predominaba en el dataset.

El segundo gran desafío lo constituye la duplicación masiva de registros. Con 799789 registros duplicados (46.47% del dataset), nos enfrentamos a un problema de calidad que podría sesgar significativamente los análisis si no se aborda adecuadamente. La detección mediante el algoritmo `pandas.DataFrame.duplicated()` con criterio de coincidencia exacta en todas las columnas reveló la magnitud de esta problemática.

Finalmente, la presencia de outliers significativos en todas las variables numéricas, detectados mediante el método de rango intercuartílico con factor 1.5, completa el trío de desafíos principales. La identificación de 124,732 outliers en **Cabezas Comercializadas** y 150,386 en **Cantidad de Kg** plantea dilemas metodológicos importantes: ¿representan estos valores extremos errores de medición o reflejan genuinamente la heterogeneidad del mercado ganadero?

Problema de Calidad	Magnitud	Impacto Potencial
Valores faltantes en Precio cabeza	94.58%	Eliminación de variable
Valores faltantes en Precio/Cantidad Kg	5.42%	Pérdida moderada de registros
Registros duplicados	46.47%	Sesgo en frecuencias
Outliers en variables numéricas	7-9%	Distorsión de distribuciones

4. PROCESO DE TRANSFORMACIÓN Y LIMPIEZA

Enfrentados a la realidad de un dataset con problemas de calidad significativos, adoptamos un enfoque riguroso que priorizó la preservación de información valiosa sobre la retención de datos problemáticos. La decisión crítica, pero también la más clara, fue la eliminación de la variable **Precio cabeza**. Con un 94.58% de valores faltantes, esta variable no aportaba nada de información para nuestro trabajo.

Reconocemos que el precio por cabeza constituye una métrica comercial importante en el sector ganadero, especialmente para ciertas categorías de ganado donde el peso individual puede variar significativamente. Sin embargo, la densidad de datos era tan baja que cualquier análisis basado en esta variable habría sido estadísticamente poco confiable y potencialmente engañoso.

Entonces, la decisión se fundamentó en un criterio técnico riguroso: variables con más del 90% de valores faltantes fueron consideradas inutilizables para el análisis.

4.1 Valores Faltantes

Una vez eliminada la variable más problemática, nos enfrentamos al dilema de qué hacer con los registros que presentaban valores faltantes en Precio Kg y Cantidad de Kg. Estos registros, que representaban el 5.42% del dataset total, requerían una decisión cuidadosa.

Sabemos que existen múltiples enfoques para sortear esta dificultad: imputación por media, imputación por mediana, imputación por regresión, o la eliminación completa de registros. La estrategia implementada fue la eliminación de registros, una decisión que aunque redujo el tamaño del dataset, garantizó la integridad de los análisis posteriores. Esta aproximación conservadora nos permitió trabajar con datos completos y confiables, evitando la introducción de sesgos que podrían haber resultado de técnicas de imputación inadecuadas.

El impacto cuantitativo de esta decisión fue significativo pero manejable: eliminamos 93286 registros, manteniendo ~1.6 millones de observaciones válidas.

4.2 Duplicados

El descubrimiento de que ~750 mil registros estaban duplicados (46.47% del dataset post-limpieza inicial) representó un hallazgo importante de nuestra fase de exploración. Esta magnitud de duplicación no sólo amenazaba la validez estadística de nuestros análisis, sino que también planteaba interrogantes sobre los procesos de recolección y almacenamiento de datos del sistema SIO-Carnes.

Para la estrategia de eliminación de duplicados utilizamos el algoritmo `drop_duplicates` de pandas con la configuración `keep='first'`, lo que significa que para cada grupo de registros idénticos, conservamos el primer registro encontrado y eliminamos todas las copias subsecuentes.

La magnitud del resultado de 'deduplicación' fue considerable: de ~1.6 millones de registros, se eliminaron ~750 mil, dejando ~877 mil registros únicos. Esta severa reducción significó que nuestro dataset final tenía aproximadamente la mitad del tamaño inicial.

4.4 Outliers

El tratamiento de outliers se resolvió mediante el método de rango intercuartílico (IQR) con factor 1.5, un estándar establecido en la literatura estadística. En **Cabezas Comercializadas**, identificamos 78501 outliers (8.94% del dataset), estableciendo límites válidos entre -17.50 y 34.50 cabezas. Aunque el límite inferior negativo carece de sentido práctico, el límite superior de 34.5 cabezas sugiere que operaciones superiores a este valor representan casos excepcionales.

La variable **Precio Kg** mostró un comportamiento más controlado, con apenas 10109 outliers (1.26%) fuera del rango de 19.50 a 199.50 pesos por kilogramo. Este rango parece razonable para el mercado ganadero argentino durante el período analizado, sugiriendo que los outliers en esta variable probablemente representan errores de registro o condiciones comerciales excepcionales.

Cantidad de Kg presentó 45784 outliers (5.80%), con límites válidos entre -4892 y 10079 kilogramos. Nuevamente, aunque el límite inferior negativo es matemáticamente imposible, el límite superior de aproximadamente 10 toneladas por operación parece razonable para el mercado ganadero.

Entonces se tomó la decisión de eliminar 168894 registros por outliers. Reconocemos que algunos de estos valores extremos podrían representar operaciones comerciales legítimas, pero su inclusión habría distorsionado significativamente los análisis estadísticos posteriores.

4.5 Creación de Variables

La transformación de datos no se limitó a la limpieza y eliminación de problemas; también involucró la creación de nuevas variables que capturan relaciones y patrones que no eran evidentes en los datos originales. Esta fase de ingeniería de características representa uno de los aspectos más creativos y estratégicos del proceso de minería de datos.

La primera variable derivada, **Total Vendido**, surgió de la multiplicación simple pero poderosa entre **Precio Kg** y **Cantidad de Kg**. Esta variable captura el valor económico total de cada transacción, proporcionando una métrica unificada para evaluar la magnitud comercial de las operaciones.

La dimensión temporal del análisis se enriqueció mediante la extracción de componentes específicos de la **Fecha Comprobante**. La creación de las variables **Mes**, **Trimestre** y **Semestre** nos permitiría analizar patrones estacionales y tendencias temporales que son fundamentales para entender la dinámica del mercado ganadero.

Finalmente, preparando el terreno para la aplicación de algoritmos de reglas de asociación, creamos versiones discretizadas de las variables numéricas principales: `CabezasDisc`, `PrecioKgDisc` y `CantidadKgDisc`. Esta transformación, esencial para la aplicación de técnicas de análisis categórico, requirió un proceso de *binning* que preservara la información estadística relevante mientras simplificara la complejidad computacional.

4.6 Discretización

La discretización de variables continuas representa uno de los procesos más delicados en la preparación de datos para reglas de asociación. Utilizamos `sklearn.preprocessing.KBinsDiscretizer` con una configuración cuidadosamente calibrada: cinco intervalos por variable (`n_bins=5`), y codificación ordinal numérica (`encode='ordinal'`). La elección de cinco bins nos parece que mantiene un balance óptimo entre granularidad informativa y manejabilidad computacional.

Los resultados de la discretización revelaron distribuciones interesantes que reflejan la estructura subyacente del mercado ganadero. Para `Cabezas Comercializadas`, el Bin 1 (1.0-7.6 cabezas) concentra el 70.9% de las operaciones, confirmando que el mercado está dominado por transacciones pequeñas. Esta concentración extrema en el primer bin sugiere un mercado altamente fragmentado, compuesto principalmente por productores pequeños y medianos.

La distribución de `Precio por Kg` muestra un patrón más equilibrado, con el bin 3 (91.6-127.4 pesos) concentrando el 41.0% de las operaciones. Esta distribución aproximadamente normal sugiere un mercado con precios relativamente estables, donde la mayoría de las transacciones ocurren en un rango de precios medio-central.

`Cantidad de Kg` presenta una distribución altamente sesgada hacia volúmenes pequeños, con el 58.8% de las operaciones en el bin 1 (1.0-2,016.6 kg). Esta concentración, similar a la observada en cantidad de cabezas, refuerza la caracterización del mercado como dominado por operaciones de pequeña escala.

5. IMPLEMENTACIÓN DE ALGORITMOS

5.1 Transformando el Dataset

La transición de un dataset tabular tradicional a un formato adecuado para reglas de asociación requirió una transformación fundamental en la estructura de los datos. Este proceso, conocido como transformación a formato transaccional, convirtió nuestras 709005 observaciones limpias en una matriz binaria de 82 items únicos, donde cada fila representa una "canasta" de características que ocurren simultáneamente en una transacción específica.

La implementación de one-hot encoding para variables categóricas creó una representación binaria donde cada valor único de cada variable categórica se convierte en una columna independiente. Por ejemplo, la variable **Zona Destino** con sus 10 valores únicos se expandió en 10 columnas booleanas, donde solo una columna presenta valor verdadero (1) por cada registro, indicando la zona específica de esa transacción.

El proceso de normalización de valores categóricos precedió a la transformación binaria y resultó crucial para garantizar consistencia. La aplicación de `.upper()` estandarizó todas las cadenas de texto, mientras que la eliminación de espacios en blanco redundantes evitó la creación de items duplicados con diferencias tipográficas menores. La codificación cuidadosa de caracteres especiales aseguró que nombres de provincias con tildes o símbolos especiales fueran tratados consistentemente.

El resultado final fue una matriz de dimensiones 709005×82 , donde cada celda contiene un valor booleano indicando la presencia (1) o ausencia (0) de un item específico en cada transacción. Esta representación proporciona la base perfecta para la aplicación de algoritmos de reglas de asociación.

5.2 Calibrando los Algoritmos

La configuración de parámetros para los algoritmos de reglas de asociación representa un ejercicio de equilibrio entre exhaustividad y practicidad. El parámetro fundamental, `min_support = 0.05`, establece que sólo consideraremos itemsets que aparezcan en al menos el 5% de las transacciones. Esta elección no fue arbitraria: umbrales más bajos habrían generado miles de reglas de baja significancia estadística, mientras que umbrales más altos habrían eliminado patrones potencialmente valiosos.

La decisión de establecer `use_colnames = True` reflejó nuestra prioridad por la interpretabilidad de resultados. Esta configuración preserva los nombres originales de las columnas en lugar de utilizar índices numéricos, facilitando la comprensión de las reglas generadas.

Para la generación de reglas, seleccionamos confianza como métrica de ordenamiento principal, estableciendo un umbral mínimo del 50% (`min_threshold = 0.5`). Esta configuración garantiza que solo consideremos reglas donde la presencia del antecedente implique el consecuente en al menos la mitad de los casos observados, proporcionando un nivel razonable de predictibilidad.

5.3 Algoritmo Apriori: Implementación y Resultados

Itemsets frecuentes generados:

- Itemsets de tamaño 1: 26
- Itemsets de tamaño 2: 79
- Itemsets de tamaño 3: 35
- Itemsets de tamaño 4: 4
- **Total:** 144 itemsets frecuentes

5.4 Algoritmo FP-Growth: Implementación Comparativa

Configuración del FP-Tree:

- Soporte mínimo: 0.05
- Construcción bottom-up del árbol
- Poda de nodos infrecuentes

Resultados comparativos:

- Itemsets frecuentes FP-Growth: 144
- Concordancia con Apriori: 100%

Reglas generadas totales: 376

6. RESULTADOS OBTENIDOS

6.1 Los Itemsets Dominantes

El análisis de itemsets frecuentes reveló una jerarquía de patrones que refleja la estructura fundamental del mercado ganadero argentino. Así, primera en ese orden de jerarquía encontramos que las operaciones pequeñas, representadas por **CabezasDisc_1_(1.0-7.6)**, dominan absolutamente el mercado con un soporte del 70.89%. Este hallazgo confirma nuestra intuición del análisis exploratorio: el mercado ganadero argentino está fundamentalmente constituido por pequeños productores que comercializan cantidades modestas de ganado.

La segunda posición la ocupa **CantidadKgDisc_1_(1.0-2016.6)** con un soporte del 58.75%, reforzando la narrativa de las operaciones pequeñas pero desde la perspectiva del peso comercializado. Esta relación entre cantidad de cabezas y volumen en kilogramos sugiere una coherencia estructural en el mercado: las operaciones pequeñas en número de animales corresponden consistentemente a volúmenes bajos en peso.

El aspecto geográfico emerge con fuerza en el tercer lugar, dónde Zona Destino_ZONA 5 B alcanza un soporte del 41.91%. Esta zona, que abarca Entre Ríos, Santa Fe y Córdoba, se confirma como el centro de la actividad ganadera argentina. Su predominancia no es sorprendente dado que esta región presenta las mejores condiciones ganaderas del país y cuenta con una infraestructura desarrollada para la comercialización de ganado.

La dimensión racial del mercado se revela en las posiciones cuarta y quinta, con BOVINO CRIOLLO (34.53%) y ABERDEEN ANGUS (30.82%) emergiendo como las razas dominantes. La supremacía del Bovino Criollo refleja la tradición ganadera argentina y su adaptación al clima y condiciones locales, mientras que la fuerte presencia del Aberdeen Angus evidencia la adopción de genética especializada en carne de alta calidad.

Ranking	Item	Soporte	Interpretación Comercial
1	CabezasDisc_1_(1.0-7.6)	70.89%	Mercado dominado por pequeños productores
2	CantidadKgDisc_1_(1.0-2016.6)	58.75%	Volúmenes bajos confirman operaciones pequeñas
3	Zona Destino_ZONA 5 B	41.91%	Región Centro como epicentro ganadero
4	Raza_BOVINO CRIOLLO	34.53%	Tradición ganadera argentina predominante
5	Raza_ABERDEEN ANGUS	30.82%	Genética especializada bien establecida
6	PrecioKgDisc_3_(91.6-127.4)	30.65%	Rango de precios modal del mercado

7	PrecioKgDisc_2_(55.8-91.6)	23.17%	Segmento de precios medio-bajos activo
8	CantidadKgDisc_2_(2016.6-403 2.2)	21.79%	Operaciones medianas representativas
9	Zona Destino_ZONA 3 B	20.21%	Buenos Aires Centro como hub comercial

6.2 Reglas de Alta Confianza

Top 5 reglas con mayor confianza:

1. **Regla 1:** [CantidadKgDisc_1, Zona_1] → [CabezasDisc_1]
 - Soporte: 0.0657
 - Confianza: 1.0000
 - Lift: 1.4106
 - **Interpretación:** Operaciones pequeñas en Buenos Aires siempre involucran pocas cabezas
2. **Regla 2:** [Zona_1, CantidadKgDisc_1, Bovino_Criollo] → [CabezasDisc_1]
 - Soporte: 0.0502
 - Confianza: 1.0000
 - Lift: 1.4106
 - **Interpretación:** Patrón específico de Buenos Aires con ganado criollo
3. **Regla 3:** [Vaca_Regular, CantidadKgDisc_1, Bovino_Criollo] → [CabezasDisc_1]
 - Soporte: 0.0662
 - Confianza: 0.9999
 - Lift: 1.4105
 - **Interpretación:** Categoría específica asociada a operaciones pequeñas

6.3 Reglas de Alto Lift

Top 5 reglas con mayor lift:

1. **Regla 1:** [CantidadKgDisc_3] → [CabezasDisc_2]
 - Soporte: 0.0733
 - Confianza: 0.6560
 - Lift: 3.5578

- **Interpretación:** Fuerte asociación entre volúmenes y cantidad de cabezas medianas
- 2. **Regla 2:** [CabezasDisc_2] → [CantidadKgDisc_3]
 - Soporte: 0.0733
 - Confianza: 0.3974
 - Lift: 3.5578
 - **Interpretación:** Asociación bidireccional confirmada
- 3. **Regla 3:** [CabezasDisc_1, Bovino_Criollo] → [CantidadKgDisc_1, Zona_1]
 - Soporte: 0.0502
 - Confianza: 0.1941
 - Lift: 2.9532
 - **Interpretación:** Patrón geográfico-racial específico

6.4 Reglas de Alto Soporte

Top 5 reglas con mayor soporte:

1. **Regla 1:** [CabezasDisc_1] → [CantidadKgDisc_1]
 - Soporte: 0.5860
 - Confianza: 0.8266
 - Lift: 1.4069
 - **Interpretación:** Regla fundamental del mercado: pocas cabezas = poco volumen
2. **Regla 2:** [CantidadKgDisc_1] → [CabezasDisc_1]
 - Soporte: 0.5860
 - Confianza: 0.9973
 - Lift: 1.4069
 - **Interpretación:** Asociación bidireccional muy fuerte

7. INTERPRETACIÓN DESDE PERSPECTIVA DE NEGOCIO

Patrones Identificados

Concentración regional:

- Zona 5 (Entre Ríos, Santa Fe, Córdoba): 41.91% de operaciones
- Especialización en ganado Aberdeen Angus y Holando Argentino
- Operaciones de mayor escala promedio

Patrones de Buenos Aires:

- Zona 1 y 3: 33.34% de operaciones combinadas
- Predominio de Bovino Criollo
- Operaciones de menor escala

Distribución racial:

- Bovino Criollo: 34.53% (raza más frecuente)
- Aberdeen Angus: 30.82% (raza premium)
- Holando Argentino: 7.41% (especializado lechero)

Categorías dominantes:

- Vaca Regular (6+ dientes): Mayor representación
- Asociación fuerte con operaciones pequeñas
- Correlación geográfica específica

Estructura de operaciones:

- 70.9% son operaciones pequeñas (1-7.6 cabezas)
- 58.8% involucran volúmenes bajos (1-2,016 kg)
- Mercado altamente fragmentado

Rangos de precios:

- Modo en rango 91.6-127.4 \$/kg (41.0% de operaciones)
- Distribución concentrada en precios medios
- Pocos outliers en precios extremos

10. CONCLUSIONES

La importante transformación del dataset original, que redujo ~1.7 millones de registros a ~709mil observaciones válidas, podría parecer una pérdida demasiado onerosa de información. Sin embargo, entendemos que fue una decisión fundada y valiosa, que nos permitió eliminar ruido, duplicaciones e inconsistencias, resultando en un dataset de calidad mejorada, libre de los sesgos que habrían comprometido la validez de nuestros hallazgos.

La validación cruzada de resultados mediante la aplicación simultánea de los algoritmos Apriori y FP-Growth constituyó también una fortaleza de nuestro trabajo. La concordancia de ambas metodologías no solo valida la robustez de nuestros hallazgos, sino que también proporciona

confianza en la reproducibilidad de los patrones identificados. Esta convergencia de resultados actúa como una doble verificación que fortalece las conclusiones extraídas.

Las reglas de asociación generadas, superando los umbrales de confianza establecidos del 50%, representan más que simples estadísticas: constituyen patrones genuinos del comportamiento comercial ganadero argentino. Quizás el hallazgo más revelador sea la identificación de una estructura de mercado claramente segmentada. El mercado ganadero argentino no es un sistema homogéneo, sino un ecosistema complejo con segmentación geográfica bien definida y preferencias raciales específicas que siguen patrones predecibles y coherentes con las características regionales y tradiciones ganaderas del país.