Análisis de Minería de Datos: Sistema SIO-Carnes

Operaciones de Compra-Venta de Ganado Vacuno con Destino a Faena

Análisis de Datos - UTN
Claudio Sebastian Castillo

2024

Introducción al Problema de Negocio

Contexto del Dataset

- **Fuente**: Sistema SIO-Carnes - datos.gob.ar
- **Período**: Agosto 2020 - Agosto 2021
- **Alcance**: 1,721,218 registros de operaciones de compra-venta
- **Cobertura**: 24 provincias argentinas, 10 zonas de destino

Hipótesis de Negocio

¿Existen patrones sistemáticos en las operaciones de compra-venta de ganado que permitan identificar asociaciones entre características geográficas, raciales y comerciales?

Objetivo del Análisis

Descubrir reglas de asociación que revelen: - Patrones geográficos de comercialización - Asociaciones entre razas y zonas de destino - Relaciones entre precios, cantidades y características del ganado

Estructura Inicial

Características del Dataset Original

Métrica	Valor
Registros totales	1,721,218
Variables	11
Tamaño	263 MB
Período temporal	12 meses

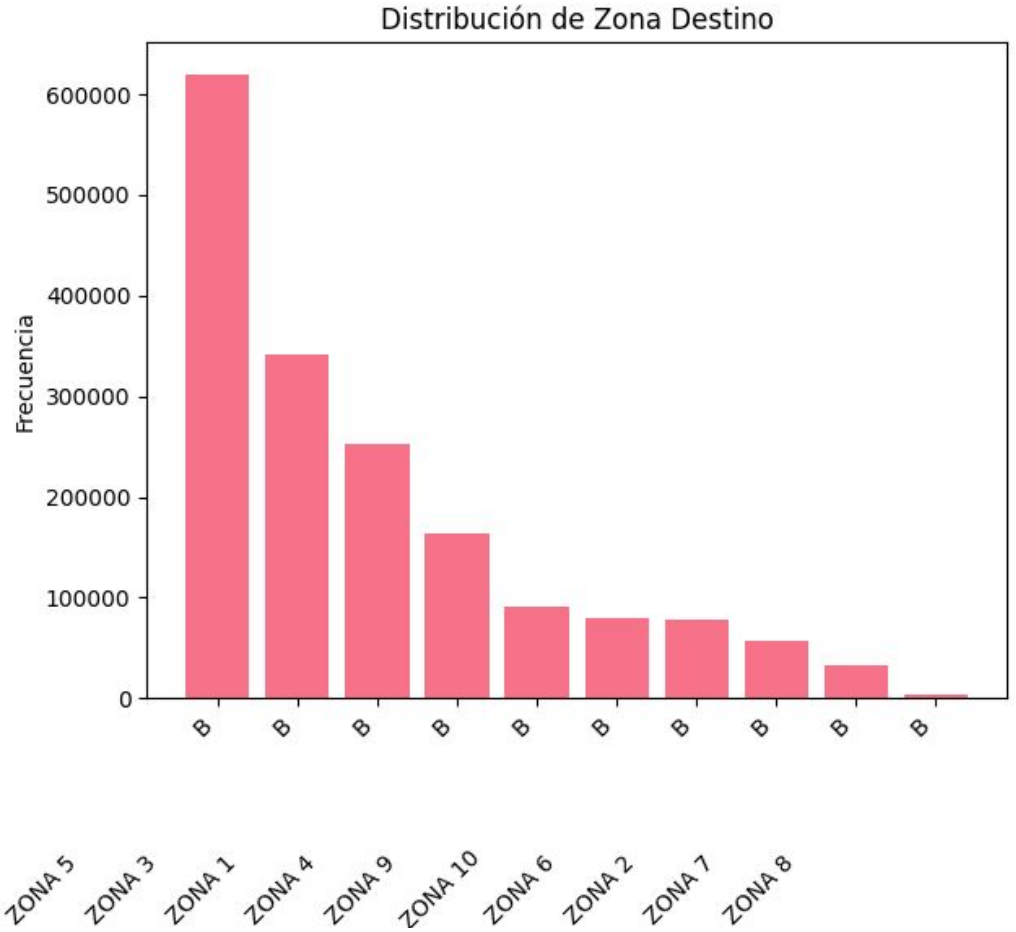
Variables Principales

- **Geográficas**: Provincia Origen, Partido Origen, Zona Destino
- **Características**: Raza, Categoría, Unidad de Medida
- **Comerciales**: Cabezas Comercializadas, Precio Kg, Cantidad Kg
- **Temporal**: Fecha Comprobante

Distribución Geográfica

- **Zona 5** (Entre Ríos, Santa Fe, Córdoba): 41.91% de operaciones
- **Zona 3** (Buenos Aires Centro): 20.21% de operaciones
- **Zona 1** (Buenos Aires y CABA): 13.13% de operaciones

Distribución Geográfica



Problemas de Calidad Identificados

Valores Faltantes Críticos

Variable	% Faltantes	Estado
Precio cabeza	94.58%	CRÍTICO
Precio Kg	5.42%	MODERADO
Cantidad de Kg	5.42%	MODERADO

Duplicación Masiva

- **799,789 registros duplicados** (46.47% del dataset)
- Posibles causas: errores de captura, re-envíos, múltiples sistemas

Outliers Significativos

- **Cabezas Comercializadas**: 8.94% outliers
- **Precio Kg**: 1.26% outliers
- **Cantidad de Kg**: 5.80% outliers

Transformaciones Aplicadas

Pipeline de Limpieza

- **Eliminación de columnas inutilizables**: Precio cabeza (94.58% faltantes)
- **Tratamiento de valores faltantes**: 93,286 registros eliminados
- **Eliminación de duplicados**: 750,033 registros eliminados
- **Tratamiento de outliers**: Método IQR, 168,894 registros eliminados

Variables Derivadas Creadas

- **Total Vendido** = Precio Kg × Cantidad Kg
- **Variables temporales**: Mes, Trimestre, Semestre
- **Variables discretizadas**: CabezasDisc, PrecioKgDisc, CantidadKgDisc

Resultado Final

- **Dataset limpio**: 709,005 registros (41.19% retención)
- **Variables finales**: 17 columnas
- **Calidad**: 100% datos válidos para análisis

Estrategia de Discretización

Método: KBinsDiscretizer con 5 bins uniformes

Distribución de Cabezas Comercializadas

- **Rango 1 (1.0-7.6)**: 502,602 registros (70.9%)
- **Rango 2 (7.6-14.2)**: 130,724 registros (18.4%)
- **Rango 3 (14.2-20.8)**: 52,197 registros (7.4%)
- **Rango 4 (20.8-27.4)**: 19,225 registros (2.7%)
- **Rango 5 (27.4-34.0)**: 4,257 registros (0.6%)

Distribución de Precios por Kg

- **Rango 3 (91.6-127.4)**: 217,335 registros (41.0%) - **Más frecuente**
- **Rango 2 (55.8-91.6)**: 164,287 registros (31.0%)
- **Rango 4 (127.4-163.2)**: 101,977 registros (19.2%)

Metodología Utilizada

Algoritmos Implementados

- **Apriori**: Búsqueda de itemsets frecuentes
- **FP-Growth**: Algoritmo alternativo para comparación
- **Association Rules**: Generación de reglas con métricas

Métricas Evaluadas

- **Soporte**: Frecuencia relativa del itemset
- **Confianza**: Probabilidad condicional de la regla
- **Lift**: Medida de dependencia entre antecedente y consecuente
- **Score Combinado**: Métrica balanceada (40% confianza + 40% lift + 20% soporte)

Resultados del Análisis de Reglas de Asociación

Resumen de Resultados

Métrica	Valor
Itemsets frecuentes	144
Reglas generadas	376
Items únicos	82
Soporte mínimo	0.05

Items Más Frecuentes

1. **CabezasDisc_1 (1.0-7.6)**: 70.89% soporte
2. **CantidadKgDisc_1 (1.0-2016.6)**: 58.75% soporte
3. **Zona 5**: 41.91% soporte
4. **Bovino Criollo**: 34.53% soporte
5. **Aberdeen Angus**: 30.82% soporte

Comparación de Algoritmos

- **Apriori**: 144 itemsets frecuentes
- **FP-Growth**: 144 itemsets frecuentes
- **Concordancia**: 100% - ambos algoritmos generaron resultados idénticos

Reglas de Mayor Impacto Comercial

Regla 1: Asociación Cantidad-Cabezas (Lift: 3.56)

`CantidadKg (4032-6048) => Cabezas (7.6-14.2)` - **Interpretación**: Operaciones medianas tienden a comercializar ganado en lotes específicos - **Aplicación**: Optimización de logística para cargas medianas

Regla 2: Patrón Geográfico-Racial (Lift: 2.95)

`Zona 1 + CantidadKg (1-2016) => Bovino Criollo + Cabezas (1-7.6)` - **Interpretación**: Buenos Aires concentra operaciones pequeñas de Bovino Criollo - **Aplicación**: Estrategias de mercado regional específicas

Regla 3: Especialización Lechera (Lift: 2.40)

`Holando Argentino => Zona 5 + Cabezas (1-7.6)` - **Interpretación**: Región Centro especializada en ganado lechero - **Aplicación**: Desarrollo de cadenas de valor lácteas regionales

Desafíos y Limitaciones Encontradas

Desafíos Técnicos

- **Alta duplicación** (46.47%): Requirió investigación de causas
- **Valores faltantes críticos**: Pérdida de variable “Precio cabeza”
- **Outliers significativos**: Necesidad de tratamiento cuidadoso
- **Tamaño del dataset**: Optimización para procesamiento eficiente

Limitaciones del Análisis

- **Pérdida de datos**: 58.81% del dataset original eliminado
- **Sesgo temporal**: Un solo año de datos
- **Granularidad**: Variables discretizadas pueden perder matices
- **Contexto externo**: Sin variables macroeconómicas o climáticas

Conclusiones y Recomendaciones Estratégicas

1. **Concentración geográfica**: Zona 5 domina el 42% de operaciones
2. **Especialización racial**: Patrones claros de distribución geográfica por raza
3. **Segmentación comercial**: Operaciones pequeñas (1-7.6 cabezas) representan 71% del mercado
4. **Predictibilidad**: Reglas de alta confianza (>99%) para ciertos patrones