**Logotipo

Descripción generada automáticamente**

**APLICACIÓN DE UN MÉTODO DE CLASIFICACIÓN SUPERVISADO A UN PROBLEMA DE SELECCIÓN DE CLIENTES MANEJANTES DE COMPETENCIA PARA BAVARIA S.A.**

Mateo Eduardo Cabarcas Támara

Cristian David Bolaños Gonzales

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
Nombres completos, Título académico más alto

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2024

|  |  |
| --- | --- |
| **Cita** | (Cabarcas Támara & Bolaños Gonzales, 2024) |
| **Referencia**  **Estilo APA 7 (2020)** | Cabarcas Támara, M., & Bolaños Gonzales, C. (2024). *Aplicación de un método de clasificación supervisado a un problema de selección de clientes manejantes de competencia para Bavaria S.A.* Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteVIII.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Dedicatoria**

Texto de dedicatoria centrado.

**Agradecimientos**

Texto de agradecimientos centrado.

**Tabla de contenido**

[Resumen 9](#_Toc183116022)

[Abstract 10](#_Toc183116023)

[1. Descripción del problema 11](#_Toc183116024)

[1.1. Problema de negocio 11](#_Toc183116025)

[1.2. Aproximación desde la analítica de datos 12](#_Toc183116026)

[1.3. Origen de los datos 13](#_Toc183116027)

[1.4. Métricas de desempeño 13](#_Toc183116028)

[2. Objetivos 16](#_Toc183116029)

[2.1. Objetivo general 16](#_Toc183116030)

[2.2. Objetivos específicos 16](#_Toc183116031)

[3. Datos 17](#_Toc183116032)

[3.1. Datos originales 17](#_Toc183116033)

[3.2. Datasets 17](#_Toc183116034)

[3.3. Analítica descriptiva 20](#_Toc183116035)

[4. Proceso de analítica 25](#_Toc183116036)

[4.1. Pipeline principal 25](#_Toc183116037)

[4.2. Preprocesamiento 25](#_Toc183116038)

[4.3. Modelos 25](#_Toc183116039)

[4.4. Métricas 25](#_Toc183116040)

[5. Metodología 26](#_Toc183116041)

[5.1. Baseline 26](#_Toc183116042)

[5.2. Validación 26](#_Toc183116043)

[5.3. Iteraciones y evolución 26](#_Toc183116044)

[5.4 Herramientas 26](#_Toc183116045)

[6. Resultados y discusión 27](#_Toc183116046)

[6.1. Métricas 27](#_Toc183116047)

[6.2. Evaluación cualitativa 27](#_Toc183116048)

[6.3. Consideraciones de producción 27](#_Toc183116049)

[7. Conclusiones 28](#_Toc183116050)

[8. Recomendaciones 29](#_Toc183116051)

[Referencias 30](#_Toc183116052)

[Anexos 31](#_Toc183116053)

[Anexo 1. Tabla de explicación de las variables 31](#_Toc183116054)

**Lista de tablas**

[Tabla 1. Resumen de variables finales 31](#_Toc183116235)

**Lista de figuras**

[Figura 1. Listado de variables antes de la operación entre columnas 18](#_Toc183116388)

[Figura 2. Listado de variables luego de obtener la variable "Var\_vol" 18](#_Toc183116389)

[Figura 3. Distribución de variables categóricas 20](#_Toc183116390)

[Figura 4. Distribución de variables categóricas (Continuación) 21](#_Toc183116391)

[Figura 5. Boxplots de variables numéricas 22](#_Toc183116392)

[Figura 6. Histogramas de frecuencias de variables numéricas 23](#_Toc183116393)

[Figura 7. Matriz de correlaciones de variables numéricas 23](#_Toc183116394)

[Figura 8. Resultados de estadísticos principales para variables numéricas 24](#_Toc183116395)

[Figura 9. Asimetría, moda y curtosis de variables numéricas 24](#_Toc183116396)

**Siglas, acrónimos y abreviaturas**

**UdeA** Universidad de Antioquia

**SKU** Stock Keeping Unit, traducido como “número de referencia único”

**AUC** Área bajo la curva

**ROC** Receiver operating characteristic curve

**ROS** Return on Sales

# Resumen

La precisa clasificación de clientes a partir del análisis de su comportamiento de compra es una cuestión cada vez más relevante para las compañías de consumo masivo. Este proyecto surge bajo la oportunidad de seleccionar correctamente los clientes para ejecutar la estrategia de participación de mercado de Bavaria. El objetivo principal es diseñar e implementar el mejor algoritmo de clasificación que permita a través de históricos de los clientes clasificarlos en una base de competencia y poder disminuir la incertidumbre con la que se toma esta decisión en la actualidad. La metodología empleada aborda limpieza y tratamiento de datos, análisis y extracción de correlaciones, así como el modelado de diversos algoritmos de aprendizaje automático, seguido del ajuste de hiperparámetros propios de cada modelo. La evaluación del rendimiento se llevó a cabo mediante métricas como, accuracy, precisión, recall, F1 score y área bajo la curva ROC. Se obtiene un rendimiento del 75% con el modelo Random Forest mediante la técnica de validación cruzada.

<https://github.com/MateoCa0397/Monografia-udea>

*Palabras clave*: Clasificación de clientes, competencia, Machine Learning, modelado predictivo.

# Abstract

Accurate customer classification based on purchasing behavior is an increasingly relevant issue for consumer goods companies. This project arises from the opportunity to correctly select customers to execute Bavaria's market share strategy. The main objective is to design and implement the best classification algorithm that allows, through customer histories, classify them on a competence base and reduce the uncertainty with which this decision is currently made. The methodology used addresses data cleaning and processing, analysis and extraction of correlations, as well as the modeling of various machine learning algorithms, followed by the adjustment of specific model hyperparameters. The performance evaluation was carried out using metrics such as accuracy, precision, recall, F1 score and area under the ROC curve. A performance of 75% is obtained with the Random Forest model using the cross-validation technique.

<https://github.com/MateoCa0397/Monografia-udea>

*Keywords***:** Customer classification, competition, machine learning, predictive modeling.

# Descripción del problema

Según NIELSEN[[1]](#footnote-2) el 94% del mercado de cervezas en Colombia pertenece a Cervecería Bavaria, pero cada día nuevos competidores surgen para acaparar el consumo de los colombianos (Forbes, 2023), cada punto porcentual de participación de mercado que pierde el gigante de la cerveza, representa casi 15 millones de dólares (Mejía, 2024), esto constituye una oportunidad enorme para la compañía y obliga a concentrar esfuerzos para cambiar la tendencia. La mejor forma de competir en una industria globalizada como la actual, es enfocarse en los clientes (Muñoz Cardona, 2007), en este sentido se propone seleccionar los clientes en los que se aplicará la estrategia de competencia de Bavaria (la cual no está en el alcance de este proyecto), a través de un modelo de clasificación que tenga en cuenta datos históricos de compra tanto de productos propios como de la competencia, tipologías de los clientes, entre otras variables significativas. Se utilizarán datos de históricos de clientes del canal tradicional atendidos por Bavaria, extraídos de la base de datos maestra del departamento de mercadeo para entrenar distintos modelos de clasificación. Las métricas de desempeño podrían incluir la precisión del modelo y el aumento de participación de mercado.

## Problema de negocio

Para una compañía de consumo masivo como Bavaria S.A. la participación de mercado es uno de los indicadores más importantes debido a que puede representar millonarias pérdidas o ganancias. En este orden de ideas, es consecuente que constantemente se traten de implementar estrategias para ganar dicha participación. Así, surge la necesidad de escoger los clientes en los que se ejecute la estrategia de competencia de manera mucho más eficiente y precisa, configurándose como una solución para detener la expansión de sus competidores. Dicha selección permitirá comprender más el mercado e identificar focos de competencia para posteriormente aplicar la estrategia de marketing predefinida.

## Aproximación desde la analítica de datos

Teniendo en cuenta las secciones anteriores, a Bavaria se le presenta un desafío importante al momento de retener a sus clientes y mantener su participación en el mercado. Los procesos que la compañía utiliza para la selección y segmentación de clientes en sus estrategias competitivas son mayormente manuales y basados en la intuición.

Ante esta situación, el presente proyecto se propone la exploración de diversos modelos de Machine Learning para la clasificación de clientes de competencia basados en datos históricos. La segmentación de clientes basado en datos históricos y comportamientos de compra ha sido un tema ampliamente estudiado y se puede asemejar a la técnica que se emplea en marketing para la predicción de abandono. En muchos casos como lo mencionan (Zadoo, Jagtap, Khule, Kedari, & Khedkar, 2022) esta es una tarea de clasificación binaria y es semejante al caso de estudio ya que el algoritmo debe seleccionar los clientes, basados en ciertas variables como disminución de ROS[[2]](#footnote-3) de cierto SKU, reporte de competencia y demás y arrojar una respuesta binaria, de si debe pertenecer o no a la base foco por su propensión a adquirir productos de la competencia.

Un buen acercamiento a la tarea de predicción de abandono y su importancia en el ámbito competitivo empresarial es hecha por (Agudelo, 2023). En su trabajo se destaca que “la capacidad de clasificar con precisión a los clientes le permite a las instituciones y a las empresas ofrecer servicios personalizados, opciones de crédito ajustados a su historial y campañas de marketing dirigidas”. Por lo tanto, como lo menciona el autor y aterrizándolo al objeto de estudio de este proyecto, el desarrollo de técnicas sólidas y sofisticadas para la clasificación de clientes, basadas en la información de historiales y tendencias de compra, ubicación geográfica y demás, podrían resultar muy útil para que Bavaria pueda tomar decisiones informadas, mitigue los riesgos y optimice los recursos al ser empleados en los clientes que generaran un mayor impacto en el objetivo de negocio (Marketshare).

## Origen de los datos

El conjunto de datos utilizado contiene información de más de 50 mil clientes del canal tradicional atendidos por Bavaria S.A en la regional ANDES. Dichos datos reflejan diferentes perfiles en aspectos de subcanal, tipología, como también registros históricos de volumen y cobertura de productos de Bavaria y su competencia. Este, fue obtenido de la base de datos maestra del departamento de mercadeo de la compañía.

## Métricas de desempeño

* **Evaluación de desempeño:**

**- Accuracy:**

La precisión (accuracy) en un problema de [clasificación](https://foqum.io/blog/termino/clasificacion/) se define como la proporción de predicciones correctas realizadas por un [modelo](https://foqum.io/blog/termino/modelo/) de Machine Learning entrenado con respecto al número total de predicciones. En otras palabras, es la medida de cuántos de los casos clasificados por el [modelo](https://foqum.io/blog/termino/modelo/) coinciden con las etiquetas reales en el conjunto de datos. La fórmula para calcular la precisión es:

**- Precisión:**

Fracción de instancias relevantes entre las instancias recuperadas, es decir, la calidad de las predicciones positivas realizadas por el modelo. Dicha métrica se define en la siguiente ecuación:

**- Recall:**

También reconocida como sensibilidad, esta métrica indica el porcentaje de muestras que un modelo de ML identifica correctamente como pertenecientes a una clase de interés (la positiva) del total de las muestras de esa clase. Esta, se muestra a continuación:

**- F1 Score:**

Métrica que combina las medidas de precisión y recall en clasificación binaria (promedio armónico), utilizada para evaluar el rendimiento predictivo de modelos de ML, especialmente útil para datos no balanceados.

**- AUC-ROC:**

La Curva ROC es un gráfico que ilustra el rendimiento de un modelo de clasificación en función de todos los posibles umbrales de clasificación. En otras palabras, muestra cómo cambia la tasa de verdaderos positivos frente a la tasa de falsos positivos, al variar el umbral que se usa para clasificar una instancia como positiva o negativa (Gerón, 2019).

El AUC (Area Under the Curve) mide el área total debajo de la curva ROC. Este valor proporciona una única métrica que resume el rendimiento del modelo en todos los umbrales posibles.

Interpretación del AUC:

* AUC = 1.0: El modelo tiene un rendimiento perfecto, identificando correctamente todas las instancias positivas y negativas.
* AUC = 0.5: El modelo no tiene poder predictivo, es decir, hace predicciones aleatorias sin distinguir entre las clases.
* AUC < 0.5: El modelo está realizando predicciones peores que aleatorias, lo que indica un modelo muy malo o incorrecto.
* **Métricas de negocio**

**-PARTICIPACIÓN DE MERCADO (Marketshare):**

La participación de mercado es una medida fundamental en marketing que refleja la cuota que una empresa tiene dentro de su mercado objetivo. Se calcula como el ratio entre las ventas de la empresa y las ventas totales del mercado, y sirve como indicador de su posición competitiva (Chernev, 2014). Expresado en ecuación matemática según (Kotler & Keller, 2006):

En el caso particular de este proyecto, esta métrica brinda información acerca de la posición de la compañía en el mercado y puede sugerir la efectividad del modelo de clasificación en el contexto del negocio. No obstante, dicha clasificación de los clientes de competencia es solo uno de los pasos que compone el plan de share de Bavaria, puesto que, una vez escogidos los clientes, se debe diseñar una estrategia basada en datos que permita identificar el SKU ideal para hacerle frente a la competencia (dependiendo de las características e historial de cada cliente). Adicionalmente, la ejecución de la estrategia mencionada debe ser llevada a cabo por el equipo comercial, el monitoreo y cobro de dicha ejecución debería ser a través de dashboards que se actualicen de manera automática y estén disponibles para su rápida consulta. Sin embargo, tanto el diseño de esa estrategia como su ejecución no son objetos de estudio en esta propuesta y por lo tanto no serán tenidos en cuenta a la hora de evaluar la efectividad de la misma. De esta manera, la propuesta sólo interviene en la solución del problema desde la selección de los clientes (clasificación), un paso extremadamente importante, que tiene impacto directo en la identificación de focos de competencia lo que permite mitigar el riesgo de expansión de la misma y además proporciona seguridad en la inversión, ya que ésta retornará en ganancia de participación de mercado si se invierte la estrategia de mercadeo predefinida en los clientes correctos.

Para que el modelo proporcione un valor significativo, viable y confiable para la organización y se puedan tomar decisiones inteligentes basadas en ello, se podría optar por un nivel de F1 score del 75-90% como rendimiento aceptable en el modelo de predicción de clientes competencia.

# Objetivos

## Objetivo general

Desarrollar una herramienta de machine learning para segmentar los clientes competencia de Bavaria, que vaya en la vía de disminuir la incertidumbre y reprocesos en la selección, aumentando las posibilidades de encontrar focos de competencia.

## Objetivos específicos

* Identificar y describir el comportamiento de las variables más significativas para la selección de clientes competencia de Bavaria.
* Aplicar herramientas de machine learning y técnicas de preprocesamiento de datos que conduzcan a mejorar la eficiencia y precisión en el análisis de las variables escogidas como influyentes en la selección de clientes competencia para Bavaria.
* Definir y calcular métricas de diagnóstico que permitan analizar varios modelos de clasificación para elegir el más adecuado para las necesidades del contexto de negocio.

# Datos

## Datos originales

El conjunto de datos está contenido en un archivo csv que contiene 16 columnas y 60,000 registros, con peso aproximado de 16 MB. Cada fila proporciona información sobre características de distintos puntos de venta del canal tradicional de Bavaria en la región Andes. Incluyendo descripción geográfica, información asociada a las ventas, cobertura, volúmenes, compra de competencia entre otros. Esta información se extrae de la base maestra del departamento de mercadeo de la compañía en cuestión.

## Datasets

En el campo del aprendizaje automático, una de las tareas fundamentales es el diseño y la construcción de algoritmos capaces de aprender de los datos y realizar predicciones precisas. Estos algoritmos operan mediante la creación de modelos matemáticos que permiten tomar decisiones o hacer predicciones basadas en la información proporcionada. Este enfoque se aplica ampliamente en diferentes sectores, incluyendo el análisis del comportamiento de los consumidores, donde la clasificación precisa de clientes se ha convertido en una herramienta clave para las compañías de consumo masivo.

Para lograr esto, se sigue una metodología rigurosa que comienza con la limpieza y tratamiento de los datos. En este primer paso, se eliminan las columnas que no aportan información significativa para el modelo, como aquellas variables descriptivas (por ejemplo, código de cliente, subcanal, gerencia, regional, coordenadas, dirección, departamento, entre otras) que no influyen en la predicción. También se eliminan variables repetidas o aquellas relacionadas con otras categorías, como las maltas, que no están incluidas para el análisis en este estudio.

Como parte de la reducción de variables, se agrupan las columnas de volumen de cerveza correspondientes a los meses de junio, julio y agosto (columnas 20 a 25 de la figura 1), y se crea una nueva variable acumulada que representa la variación del volumen comparado con el mismo período del año anterior.

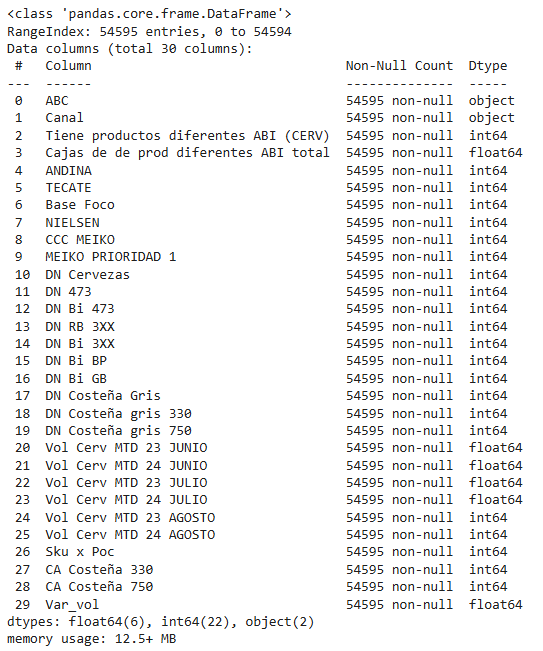


Figura 1. Listado de variables antes de la operación entre columnas

Este cambio se expresa como una cifra positiva si el volumen ha aumentado y negativa si ha disminuido. De esta manera, se genera una única variable denominada “Var\_vol”, que sintetiza esta información y facilita el análisis posterior (Figura 2). Teniendo como resultado las siguientes variables para el modelo:

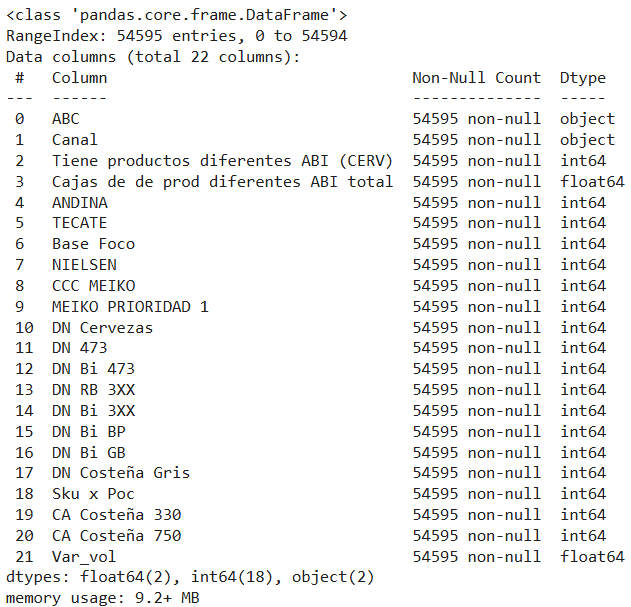


Figura 2. Listado de variables luego de obtener la variable "Var\_vol"

En el anexo 1 se encuentra la tabla 1 con todas las variables consideradas para el modelo final y su respectiva descripción.

Continuando con la preparación de los datos, se sigue un proceso de preprocesamiento, implementando diversas técnicas para garantizar la calidad y coherencia de los datos. En primer lugar, se llevó a cabo una revisión exhaustiva para eliminar registros duplicados, con el fin de evitar que la repetición de datos distorsione los resultados y afecte el rendimiento del modelo. Seguidamente, se verifica la presencia de valores nulos en el conjunto de datos. Dado que no se encuentran valores faltantes o nulos, no se hace necesario aplicar técnicas de imputación, lo que permite mantener la integridad de los registros sin la introducción de estimaciones. Finalmente, se aplica la normalización de los datos utilizando el método Min-Max scaler, escalando todas las características numéricas dentro del rango de 0 a 1:

 #ESCALAR VARIABLES NUMERICAS

numCols = df\_final.select\_dtypes(include = ['float64','float64','int32','int64']).columns.tolist()

numCols.remove('Base Foco')

num\_vars = numCols

print(num\_vars)

 pd.set\_option('display.float\_format', lambda x: '%.4f' % x)

 scaler = MinMaxScaler()

X\_train[num\_vars] = scaler.fit\_transform(X\_train[num\_vars])

X\_test[num\_vars] = scaler.transform(X\_test[num\_vars])

X\_test[num\_vars].head(2)

Este paso es crucial para garantizar que todas las variables tengan la misma escala, evitando que las variables con magnitudes mayores desproporcionen la influencia sobre el modelo. Estas técnicas de preprocesamiento son esenciales para asegurar que los datos estén listos para el modelamiento.

Para realizar la partición de los datos, se utilizó el método Split, que aplica una estratificación basada en una variable específica y distribuye los datos de forma aleatoria según el porcentaje determinado. En este caso, se asignó el 90% de los datos al conjunto de entrenamiento y el 10% restante al conjunto de evaluación, asegurando que el modelo se entrene con una cantidad suficiente de datos y se evalúe utilizando un conjunto independiente.

Para abordar el desbalanceo en la variable de salida “base foco”, se aplica una técnica de sobremuestreo utilizando la función oversampler, que ajusta el conjunto de entrenamiento generando muestras sintéticas de la clase minoritaria. Esto equilibra la distribución de las clases y permite que el modelo entrene con datos balanceados, mejorando su capacidad para predecir correctamente ambas respuestas.

 oversampler = RandomOverSampler()

X\_resampled, y\_resampled = oversampler.fit\_resample(df1.drop('Base Foco', axis=1), df1['Base Foco'])

df2 = pd.DataFrame(X\_resampled, columns=df.columns.drop('Base Foco'))

df3 = pd.DataFrame(y\_resampled)

df\_final = pd.concat([df2, df3], axis=1)

df\_final.shape

Posteriormente, dentro del 90% destinado al entrenamiento, se realizó una nueva partición en una proporción 80:20 para la fase de entrenamiento y pruebas de los modelos, permitiendo evaluar cada modelo con datos "vírgenes" no utilizados durante el proceso de ajuste. La elección de estas proporciones depende del volumen total de datos disponibles y de la necesidad de contar con un conjunto adicional para la optimización de los hiperparámetros del modelo. Este enfoque garantiza que los modelos sean entrenados de manera efectiva y evaluados de forma precisa con datos no vistos, lo que permite obtener una medición más fiable de su rendimiento.

#DIVISION DE LOS DATOS EN ENTRENAMIENTO Y TEST EN UN 80% TRAIN Y 20% TEST

X = df\_final.drop(columns ='Base Foco')

y = df\_final['Base Foco']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y.values.reshape(-1,1),

                                        train\_size   = 0.8,

                                        random\_state = 123,

                                        shuffle      = True)

## Analítica descriptiva

En las figuras 3 y 4 se muestran las distribuciones de las variables categóricas del dataset:

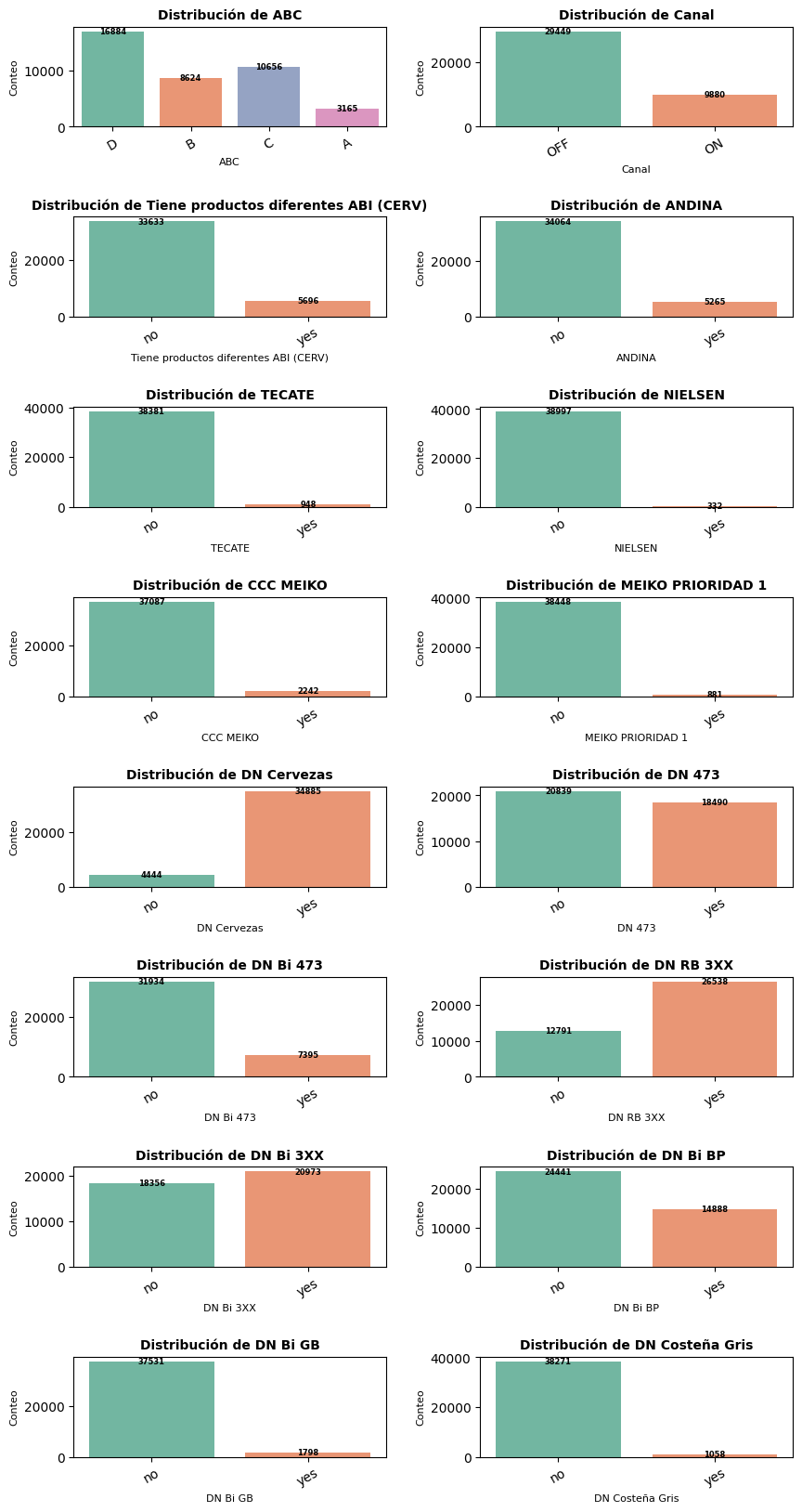


Figura 3. Distribución de variables categóricas

Se puede observar un desbalanceo de registros para las variables que corresponden al comportamiento de la competencia, lo que puede ser entendido desde el punto de vista de que Bavaria ocupa el 95% del mercado y como es de esperarse no en todos los clientes se encuentra competencia, esto da una claridad en cuanto a la veracidad de la información que en su mayor parte es recolectada manualmente o por terceros. Adicionalmente y como también es de esperarse, basados en la figura ubicada en la fila 1 y columna 1 de la figura 3, podemos concluir que los registros encontrados en el dataset son mayormente de clientes de tamaño mediano a pequeño (C y D), como sucede en el canal tradicional y en su mayoría (fila 1, columna 2, figura 3) corresponden a clientes del canal OFF (consumo en el hogar).

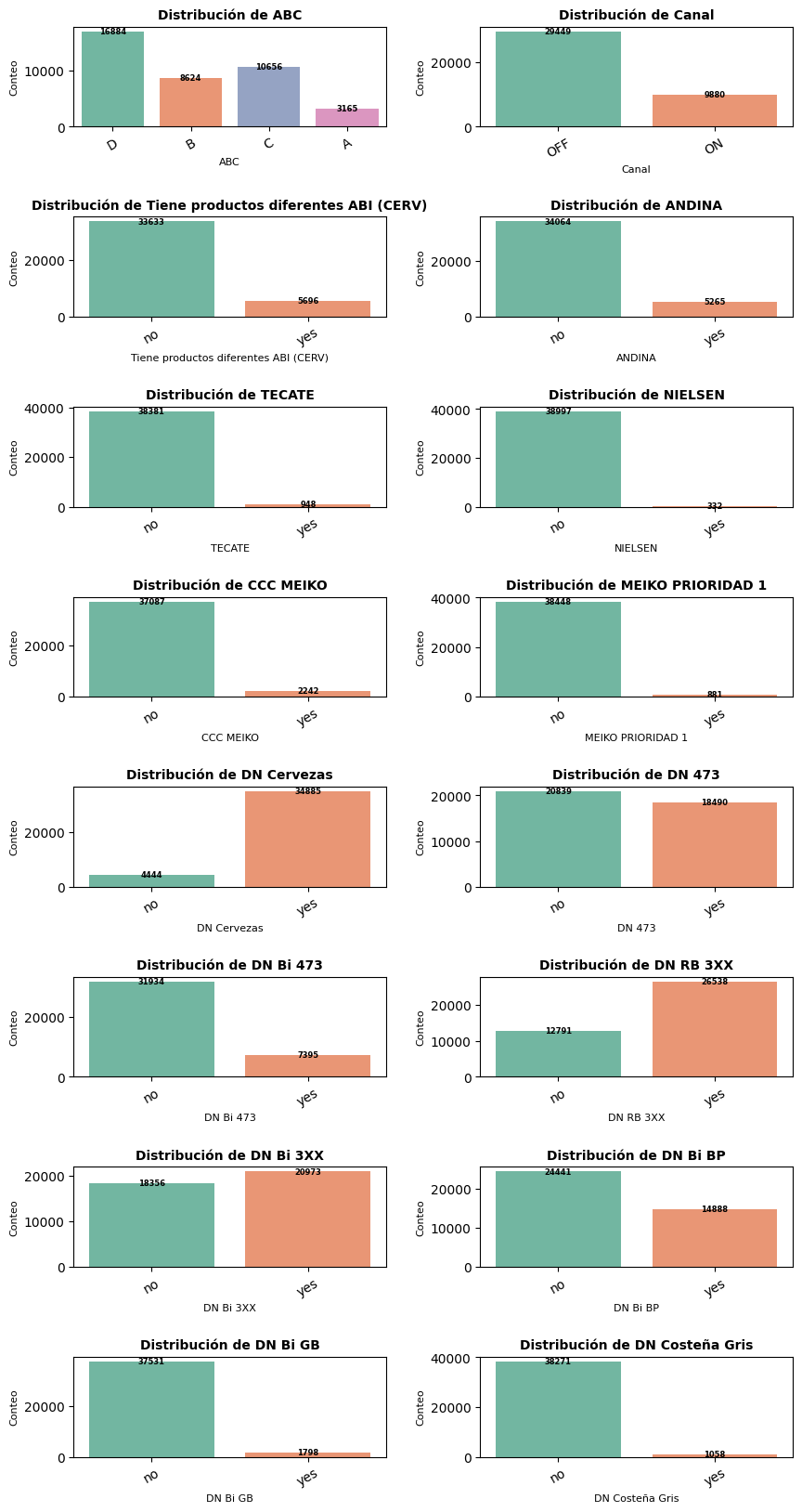


Figura 4. Distribución de variables categóricas (Continuación)

Cuando revisamos las distribuciones de las variables correspondientes a históricos de compra de productos Bavaria (figura 4) se puede identificar un mayor balanceo, excepto por las variables que describen la cobertura de skus como globales o costeña gris (última fila), algo normal desde el contexto del negocio ya que son coberturas de skus flankers (pensados para competir como estrategia en clientes puntuales) o son skus con coberturas históricamente más reducidas ya que son de un valor adquisitivo mayor.

En la figura 5 se presentan los gráficos de boxplot de las variables numéricas del dataset

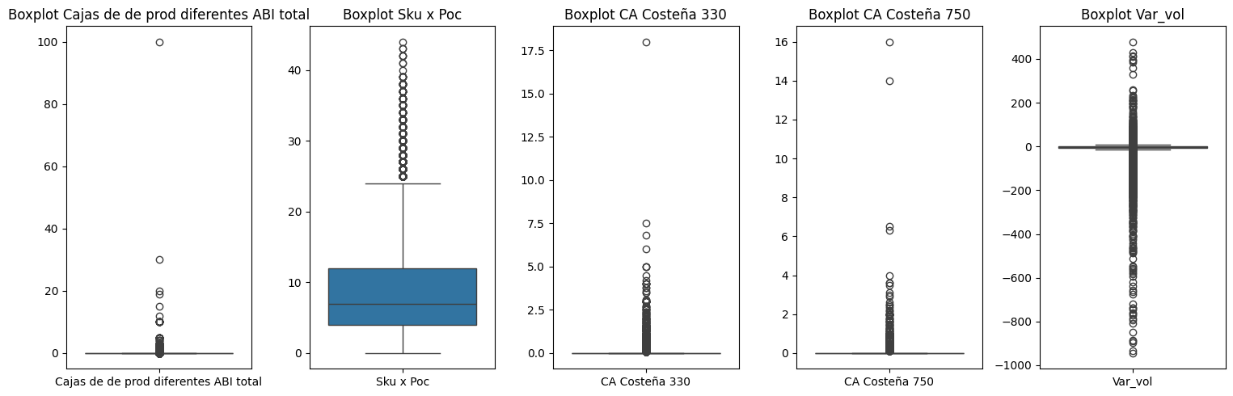


Figura 5. Boxplots de variables numéricas

Se puede evidenciar que en una mayor o menor medida todas las variables se encuentran muy dispersas y presentan valores atípicos que deben ser eliminados. Para esto se emplea la técnica de LocalOutlierFactor con 15 vecinos (n) ya que se trata de un dataset de más de 1000 registros y por heurística se recomienda este valor (10< n <20). Se presentan las líneas de código usadas:

lof = LocalOutlierFactor(n\_neighbors=15)

y\_pred = lof.fit\_predict(df[var\_num])

outliers = y\_pred == -1

print('Indices de valores atípicos: ', df[outliers].index.tolist())

len(outliers)

# Eliminar los valores atípicos del dataframe original

df = df.loc[~outliers]

df.reset\_index(drop=True)

df.shape

De esta forma se eliminan un total de 4900 registros con valores atípicos y se obtienen los histogramas de distribución de la figura 6, en ellos se puede observar que aún existe dispersión en los datos pues se ocupan ejes logarítmicos para la representación, pero es inherente al contexto de los datos y estos ya se encuentran un poco más concentrados. En cuanto a la forma de los histogramas encontramos que la variable “var\_vol” tiene una forma leptocúrtica con altos picos en bajas variaciones, la variable “skus por poc” tiene una distribución sesgada a la izquierda ya que se concentra en valores más pequeños y las variables referentes a volúmenes de costeña se comportan con asimetría negativa por el mismo motivo explicado en la revisión de las variables categóricas.

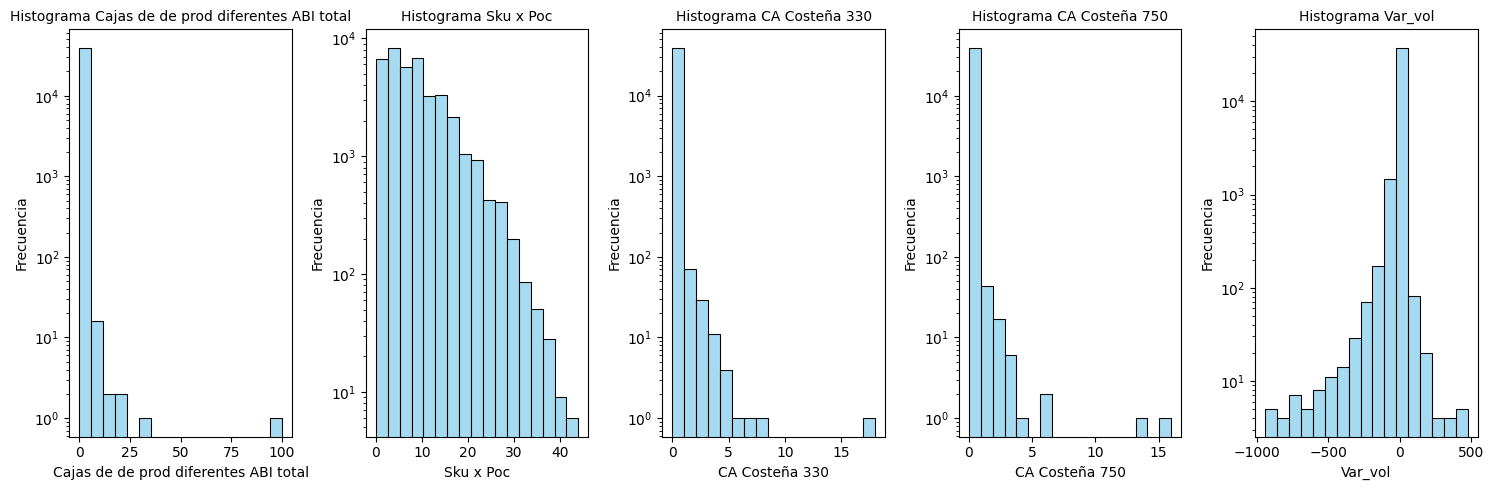


Figura 6. Histogramas de frecuencias de variables numéricas

Se procede a evaluar la correlación posible entre las variables numéricas, con el fin de entender si hay riesgos de multicolinealidad y encontrar posibles relaciones en el contexto del negocio. Los resultados son mostrados en la figura 7:

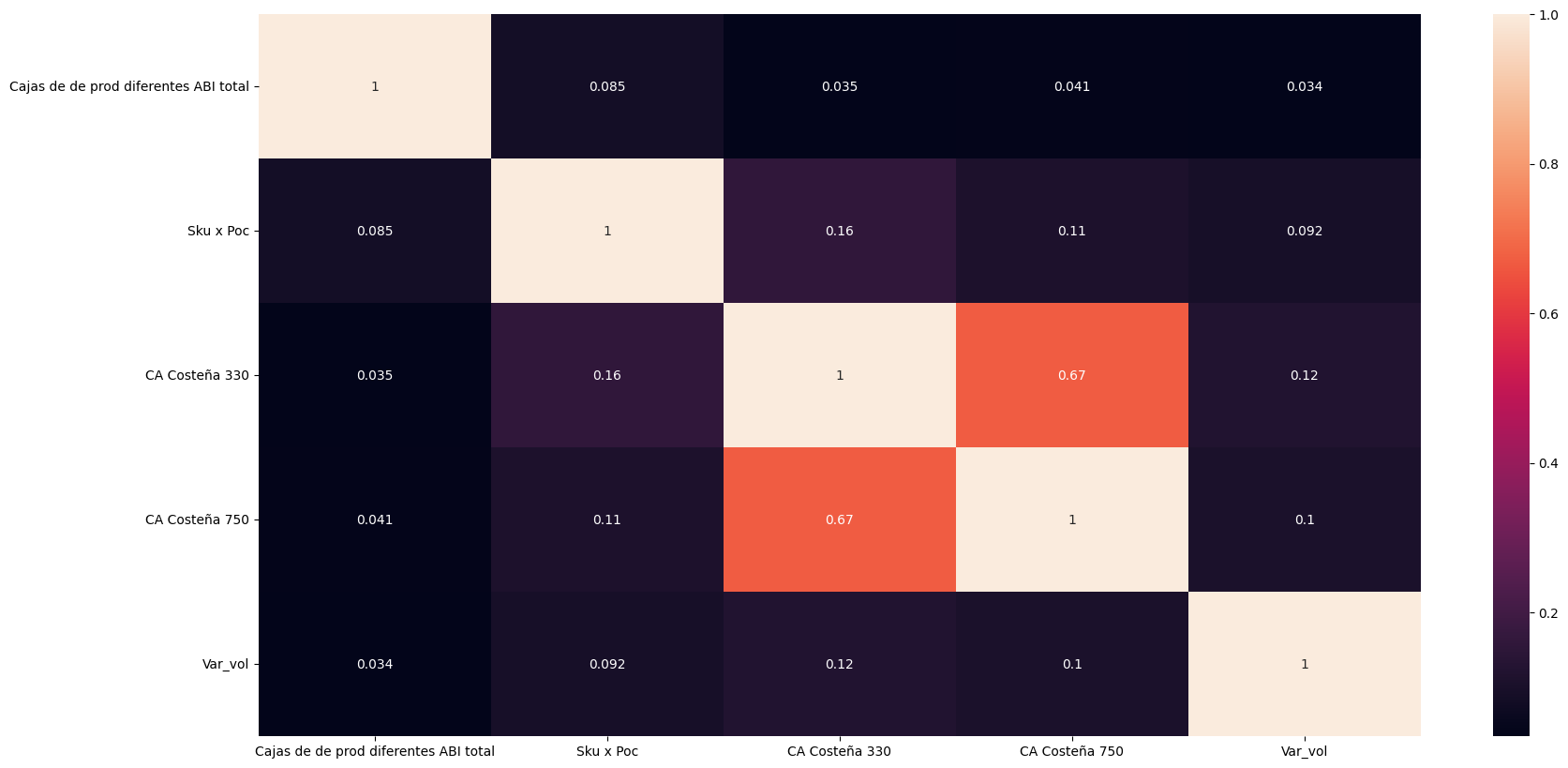


Figura 7. Matriz de correlaciones de variables numéricas

No se observa alta correlación sino en solo dos variables que describen el comportamiento de las cajas compradas de costeña 330 y 750, esto es de esperarse ya que casi a la mayoría de los clientes se les habilitan ambos skus flanker cuando son seleccionados para estrategia de competencia.

Finalmente se calculan los estadísticos principales de las variables numéricas, en la figura 8 y 9 se observan los resultados:



Figura 8. Resultados de estadísticos principales para variables numéricas

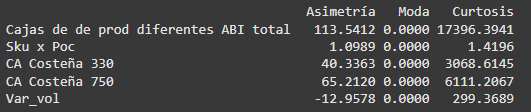


Figura 9. Asimetría, moda y curtosis de variables numéricas

De las figuras 8 y 9 se puede inferir lo siguiente:

* La cantidad promedio de skus que maneja un punto de venta en la regional ANDES es 8.39.
* La regional ANDES en promedio decrece 5.24 hL en volumen de cervezas.
* La desviación estándar de la variable var\_vol indica que los datos están muy dispersos como se mencionó con anterioridad y se puede explicar ya que los registros corresponden a clientes con características muy distintas, desde clientes mayoristas hasta tiendas básicas.
* En cuanto a la curtosis y asimetría se refuerzan las conclusiones mencionadas anteriormente. Los datos de las variables diferentes a sku x poc tienen sesgos hacia la derecha o izquierda y en general presentan leptocurtosis, lo que indica que los datos tienen una mayor concentración alrededor de la media y es más probable que haya **valores extremos** (outliers).

# Proceso de analítica

## Pipeline principal

Describe con un gráfico el flujo de trabajo general de los datos en tu proyecto. Añade o modifica secciones si lo consideras necesario

## Preprocesamiento

Describe las alternativas de preprocesamiento de datos que consideraste, aumentación de datos, etc.

## Modelos

Describe los distintos modelos que consideraste, sus configuraciones, etc.

## Métricas

Describe cómo calculas las métricas de desempeño ML y de negocio (Por ejemplo: con qué funciones de sklearn, tensorflow, etc.)

# Metodología

## Baseline

Describe tu primera iteración, sus resultados y los problemas técnicos que tuviste que resolver

## Validación

Describe tu proceso de validación, particiones de train/test/validación, etc.

## Iteraciones y evolución

Describe en qué te enfocaste en las iteraciones subsiguientes. No es necesario hacer un listado exhaustivo de las iteraciones, pero sí has de describir qué perseguías y qué obtuviste con cada grupo de iteraciones. P.ej. primero hicimos un conjunto de iteraciones para definir las opciones de preprocesamiento, luego hicimos unas iteraciones probando distintos modelos de tal tipo, etc.

## 5.4 Herramientas

Menciona las herramientas que usaste para tu proyecto

# Resultados y discusión

En los resultados se comunican los hallazgos y descubrimientos del estudio. Se incluyen tablas, figuras, diagramas y demás material demostrativo.

La discusión es la interpretación crítica y el análisis de los resultados, que surgen de las preguntas de investigación.

## Métricas

Describe los resultados numéricos de las métricas de las iteraciones que consideres más relevantes, junto con las configuraciones. Usa tablas o gráficas siguiendo el formato mostrado a continuación.

## Evaluación cualitativa

Realiza una evaluación cualitativa de los resultados, indicando casos de overfitting, underfitting, utilidad de los resultados, relación obtenida entre la métrica de ML y la de negocio, etc.

## Consideraciones de producción

Menciona consideraciones técnicas de una posible puesta en producción. Por ejemplo: condiciones de monitoreo del desempeño de los modelos, integración con streams de datos, servicios en la nube, etc.

# Conclusiones

Son las interpretaciones finales que recopilan los datos de la investigación, describe lo que se obtuvo, qué se logró y cuáles son los resultados. Guardan relación directa con lo que se mencionó en el planteamiento del problema y en los objetivos. Haz un breve juicio crítico de tu proyecto.

# Recomendaciones

Las recomendaciones son las futuras y posibles líneas de investigación que llevarán a resolver problemas relacionados con la presente investigación.

# Referencias

Agudelo, H. (2023). *ANÁLISIS DE RETENCIÓN DE CLIENTES EN INSTITUCIONES BANCARIAS.* Medellín: Universidad de Antioquia.

Chernev, A. (2014). *Strategic marketing management.* Cerebellum Press.

Forbes. (2023). Venta de cervezas artesanales crece 51% en Colombia. *Forbes Colombia*.

Gerón, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.*

Kotler, P., & Keller, K. L. (2006). *Marketing Management.* Pearson.

Mejía, F. (5 de Agosto de 2024). Mercado de la cerveza mueve cerca de $22 billones al año, es una industria resiliente. (C. Jaramillo, Entrevistador)

Muñoz Cardona, Á. E. (2007). Los Retos de la competencia industrial. *Semestre Económico*.

Ramirez Prada, S. (19 de Enero de 2024). Cerveza Andina celebra un exitoso 2023 y prepara sus nuevos desafíos para el 2024. (A. Lugo, Entrevistador)

Zadoo, A., Jagtap, T., Khule, N., Kedari, A., & Khedkar, S. (2022). A review on Churn Prediction and Customer Segmentation using Machine Learning. *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*.

# Anexos

## Anexo 1. Tabla de explicación de las variables

|  |  |
| --- | --- |
| **ABC** | Tipología del cliente, A para grandes B para medianos y C y D pequeños. Se clasifican según volumen que manejan de productos de Bavaria |
| **Canal** | Se refiere al canal de venta del cliente hacia el consumidor final, se clasifican en OFF si el cliente vende productos para llevar o en ON si tiene adaptado su negocio para el consumo en sitio |
| **Tiene productos diferentes ABI (CERV)** | Indica si el representante de ventas contestó en la encuesta si el cliente tiene o no marcas de la competencia |
| **Cajas de de prod diferentes ABI total** | Indica la cantidad de competencia (volumen)que el representante de ventas contestó en la encuesta |
| **ANDINA** | Indica si el cliente tiene o no cobertura de cervezas de la marca ANDINA |
| **TECATE** | Indica si el cliente tiene o no cobertura de cervezas de la marca TECATE |
| **Base Foco** | Variable de salida, indica si el cliente pertenece a la base foco (competencia) para Bavaria |
| **NIELSEN** | Indica si el cliente es reportado por NIELSEN (proveedor de análisis de mercado) como comprador de competencia |
| **CCC MEIKO** | Indica si el cliente es reportado por MEIKO (proveedor de análisis de mercado) como comprador de competencia |
| **MEIKO PRIORIDAD 1** | Indica si el cliente es reportado por MEIKO como comprador de competencia y pertenece a cuadrantes cercanos donde hay concentraciones de competencia |
| **DN Cervezas** | Indica si el cliente tiene cobertura de cualquier tipo de producto de Bavaria |
| **DN 473** | Indica si el cliente tiene cobertura de skus de calibre 473 ml |
| **DN Bi 473** | Indica si el cliente tiene cobertura de dos o más skus de calibre 473 ml pero diferente marca |
| **DN RB 3XX** | Indica si el cliente tiene cobertura de dos o más skus de botella retornable 310 ml en adelante hasta 355 ml |
| **DN Bi 3XX** | Indica si el cliente tiene cobertura de dos o más skus de botella retornable 310 ml en adelante hasta 355 ml |
| **DN Bi BP** | Indica si el cliente tiene cobertura de al menos dos skus de calibres grandes (750, 850 y 1000 ml) |
| **DN Bi GB** | Indica si el cliente tiene cobertura de al menos dos skus de marcas globales (importadas-Corona, budweiser, Stella artois, Michelob) |
| **DN Costeña Gris** | Indica si el cliente tiene cobertura de alguno de los dos skus de costeña gris (330 o 750 ml) |
| **Sku x Poc** | Cantidad de skus de Bavaria que maneja el cliente |
| **CA Costeña 330** | Volumen en cajas de costeña 330 que maneja el cliente |
| **CA Costeña 750** | Volumen en cajas de costeña 750 que maneja el cliente |
| **Var\_vol** | Volumen en HL (hectolitros) que pierde o gana el cliente comprando los meses de junio, julio y agosto de 2023 vs 2024 |

Tabla 1. Resumen de variables finales

1. Es el proveedor número uno a nivel mundial de información de marketing, medición de audiencia y productos y servicios de medios comerciales. [↑](#footnote-ref-2)
2. Facturación mensual de cajas de cierto SKU por cliente [↑](#footnote-ref-3)