



Escuela de Administración

Transformación de la Industria Retail en Centros Comerciales: Modelos Avanzados de
Clasificación y Predicción

Proyecto Empresarial – Trabajo de Grado

Presentado por:

David José Cantillo Rocha, Juan Sebastián Cardona, Juan Pablo Marulanda y María Camila
Morales

Maestría en Business Analytics
Escuela de Administración

Bogotá, D.C. 14 de julio de 2024



**Universidad del
Rosario**

Escuela de Administración
Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Transformación de la Industria Retail en Centros Comerciales: Modelos Avanzados de
Clasificación y Predicción

Presentado por:

David José Cantillo Rocha, Juan Sebastián Cardona, Juan Pablo Marulanda y María Camila
Morales

Bajo la dirección de:

Daniel Leonardo Cruz Castro

14 de julio de 2024

Declaración de originalidad y autonomía	6
Declaración de exoneración de responsabilidad	7
Lista de figuras.....	8
Lista de tablas	9
Lista de Gráficos	9
Lista de Ecuaciones	9
Resumen Ejecutivo	10
Palabras clave	11
Abstract	12
Keywords	13
1. Introducción	14
2. Objetivos	16
3. Alcance del proyecto aplicado	16
4. Marco teórico	17
4.1.1. Regresión logística:	17
4.1.2. Función de regresión logística.....	18
4.1.3. Aprendizaje de redes neuronales:.....	19
4.1.4. Comparación.....	20
4.2. Técnicas de Analítica para Segmentación.....	21
4.2.1. K-means.....	21
4.2.2. ACM.....	22
4.2.3. Comparación.....	24
4.3. Otras Técnicas Aplicadas.....	24
4.3.1. Oversampling:	24
4.3.2. Undersampling:	25
4.3.3. Comparación.....	26
5. Metodología CRISP - DM	26
6. Cronograma.....	29
7. Análisis descriptivo.....	29
7.1. Comportamiento de la Población por CCs.....	36
7.1.1. Edad de Visitantes por CCs	36

7.1.2.	¿Con quién va al CC?	38
7.1.3.	Motivación Visita al CC	40
8.	Transformación de datos	41
9.	Descripción de la Situación organizacional donde se realizará el proyecto	44
10.	Descripción de la situación estudio de caso y/o problemática empresarial y método y/o estrategia a aplicar para su solución	48
11.	Evaluación de modelo candidato	49
11.1.	Regresión Logística Binomial.....	50
11.1.1.	Balanceo de Datos	51
11.1.2.	Selección de Variables.....	53
11.1.3.	Evaluación de Odds	56
11.1.4.	Pseudo-R ² de McFadden.....	63
11.1.5.	Evaluación Importancia de Variables (VAR IMP).....	64
11.1.6.	Variance Inflation Factor (VIF).....	66
11.2.	Evaluación Modelos:	76
11.3.	Resultados Regresión Logística por CC	78
11.3.1.	Barranquilla	79
11.3.2.	Medellín.....	80
11.3.3.	Bucaramanga	80
11.3.4.	Pereira.....	80
11.3.5.	Sopó.....	81
11.4.	Modelo de Regresión Logística por Tipo de pregunta	81
11.4.1.	Centro Comercial 1.....	82
11.4.2.	Centro Comercial 2.....	83
11.4.3.	Centro Comercial 3.....	84
11.4.4.	Centro Comercial 4.....	86
11.4.5.	Centro Comercial 5.....	87
12.	Perfiles de Clientes Por Centro Comercial	88
12.1.	Resultados Regresión Lineal.....	88
12.1.1.	Perfiles por Centro Comercial	90
12.2.	Análisis de Correspondencias Múltiples	94
12.2.1.	ACM Variables demográficas CC1	96

12.2.2.	ACM Variables demográficas CC2.....	97
12.2.3.	ACM Variables demográficas CC3.....	98
12.2.4.	ACM Variables demográficas CC4.....	99
12.2.5.	ACM Variables demográficas CC5.....	100
13.	Recomendaciones para Atracción de Clientes.....	100
13.1.	Barranquilla.....	102
13.2.	Medellín.....	102
13.3.	Bucaramanga.....	103
13.4.	Pereira.....	104
13.5.	Sopó.....	105
14.	Conclusiones.....	106
	Referencias bibliográficas.....	108
	Anexo I – Encuesta Realizada.....	111
	Anexo II – Regresión Logística por Centro Comercial.....	118

Declaración de originalidad y autonomía

Declaro(amos) bajo la gravedad del juramento, que he(mos) escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por mi(nuestra) propia cuenta y que, por lo tanto, su contenido es original.

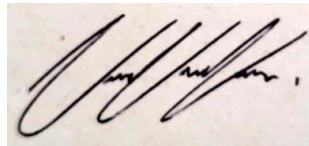
Declaro(amos) que he(mos) indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.



Juan Sebastian Cardona Nieto



David José Cantillo.



Juan Pablo Marulanda




María Camila Morales Ortiz

Firmado en Bogotá, D.C. el 16 de junio de 2024

Declaración de exoneración de responsabilidad

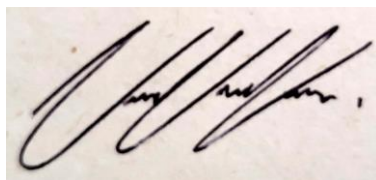
Declaro(amos) que la responsabilidad intelectual del presente trabajo es exclusivamente de su(s) autor(es). La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.



Juan Sebastián Cardona Nieto



David José Cantillo.



Juan Pablo Marulanda



Maria Camila Morales Ortiz

Firmado en Bogotá, D.C. el 14 de julio de 2024

Lista de figuras

Figura 1.	18
Figura 2.	18
Figura 3.	29
Figura 4.	30
Figura 5.	31
Figura 6.	33
Figura 7.	34
Figura 8.	35
Figura 9.	36
Figura 10.	38
Figura 11.	40
Figura 12.	46
Figura 13.	47
Figura 14.	50
Figura 15.	55
Figura 16.	64
Figura 17.	64
Figura 18.	66
Figura 19.	67
Figura 20.	71
Figura 21.	72
Figura 22.	75
Figura 23.	96
Figura 24.	97
Figura 25.	98
Figura 26.	99
Figura 27.	100
Figura 28.	101

Lista de tablas

Tabla 1.	20
Tabla 2.	28
Tabla 3.	42
Tabla 4.	43
Tabla 5.	44
Tabla 6.	52
Tabla 7.	52
Tabla 8.	56
Tabla 9.	57
Tabla 10.	79
Tabla 11.	88
Tabla 12.	95

Lista de Gráficos

Gráfico 1.	82
Gráfico 2.	83
Gráfico 3.	84
Gráfico 4.	86
Gráfico 5.	87

Lista de Ecuaciones

Ecuación 1. Odds Ratio	58
-------------------------------------	----

Resumen Ejecutivo

Parque Arauco es una empresa dedicada al sector de rentas inmobiliarias, enfocada en maximizar las ventas en sus centros comerciales y outlets mediante la oferta de productos, servicios y experiencias adaptadas a los nuevos hábitos de consumo de sus visitantes.

Teniendo como guía uno de los pilares organizaciones de Parque Arauco (*Junta Ordinaria de Accionistas 2024*, s. f.), este trabajo se centrará en comunicar estrategias alineadas con dos de estos pilares:

- **Rentabilidad:** Aumentar la rentabilidad de los activos inmobiliarios del portafolio de inversión (los Centros Comerciales).
- **Experiencia de Clientes:** Ser el operador de Real Estate recomendado por clientes y locatarios.

Además, la situación global actual demuestra que la pandemia ha generado cambios significativos en los patrones de consumo y ha impulsado el desarrollo de nuevos canales de compra.

Por estas razones, se utilizarán los resultados de una encuesta aplicada de manera independiente por Parque Arauco en el año 2023 a cinco de sus centros comerciales. Dicha encuesta, previamente validada por la entidad, servirá como insumo para la creación de un modelo predictivo de clientes con mayor probabilidad de compra para los centros comerciales en general, así como de modelos predictivos adicionales para cada uno de los centros comerciales de forma individual. Además, mediante técnicas de análisis de datos, como el análisis de correspondencias múltiples, se proporcionarán recomendaciones estratégicas basadas en el perfilamiento de los visitantes, optimizando los esfuerzos e

inversiones del grupo empresarial y proporcionando información basada en datos para mejorar la experiencia del cliente y, en consecuencia, su posible gasto en el centro comercial.

La ejecución de este proyecto se llevará a cabo mediante la integración de metodologías como CRISP-DM y Scrum. Durante el análisis de la estrategia de trabajo integrada, se emplearon técnicas de balanceo, análisis de correspondencias y análisis descriptivos para asegurar la validez de los datos evaluados. Posteriormente, se consideraron varios modelos analíticos. Sin embargo, tras la evaluación de los modelos, explicados a lo largo del documento, se observó que el modelo de regresión logística binomial demostró ser el más efectivo en la fase de evaluación. Esta conclusión se basa en los resultados de indicadores como el R^2 de McFadden, el VIF, la curva ROC y el área bajo la curva (AUC), lo que proporciona suficiente evidencia estadística para afirmar que los resultados son válidos y confiables para la subsecuente generación de recomendaciones, cuya implementación será decidida por la empresa.

Palabras clave

Modelo predictivo, perfilamiento de clientes, maximización de utilidad, Parque Arauco, Centros Comerciales (CC), compra, visita.

Abstract

Parque Arauco is a company dedicated to the real estate rental sector, focusing on maximizing sales in its shopping centers and outlets by offering products, services, and experiences adapted to the new consumption habits of its visitors. Guided by Parque Arauco's organizational pillars (Ordinary Shareholders' Meeting 2024, n.d.), this work will focus on communicating strategies aligned with two of these pillars:

- **Profitability:** Increase the profitability of the real estate assets in the investment portfolio (the Shopping Centers).
- **Customer Experience:** Be the recommended real estate operator by clients and tenants.

Furthermore, the current global situation demonstrates that the pandemic has generated significant changes in consumption patterns and has driven the development of new purchasing channels.

For these reasons, the results of a survey independently conducted by Parque Arauco in 2023 at five of their shopping centers will be utilized. This survey, previously validated by the entity, will serve as input for creating a predictive model of customers with a higher probability of purchase for shopping centers in general, as well as additional predictive models for each shopping center individually. Additionally, through data analysis techniques such as multiple correspondence analysis, strategic recommendations based on visitor profiling will be provided, optimizing the group's efforts and investments and providing data-driven information to improve the customer experience and, consequently, their potential spending at the shopping center.

The execution of this project will be carried out by integrating methodologies such as CRISP-DM and Scrum. During the analysis of the integrated work strategy, balancing techniques, correspondence analysis, and descriptive analyses were employed to ensure the validity of the evaluated data. Subsequently, several analytical models were considered. However, after evaluating the models, as explained throughout the document, it was observed that the binomial logistic regression model proved to be the most effective in the evaluation phase. This conclusion is based on the results of indicators such as McFadden's R^2 , VIF, the ROC curve, and the area under the curve (AUC), which provide sufficient statistical evidence to assert that the results are valid and reliable for the subsequent generation of recommendations, whose implementation will be decided by the company.

Keywords

Predictive model, customer profiling, utility maximization, purchase, Parque Arauco, Shopping Centers (CC), visit.

1. Introducción

El director ejecutivo de Acecolombia, en una entrevista concedida a La República (Fayad, 2023), informó que Colombia alberga actualmente 261 centros comerciales, con otros 70 en proceso de construcción en todo el territorio nacional. Esta situación demuestra el desarrollo de una industria en la cual convergen de manera progresiva el consumo de bienes, servicios y entretenimiento. En consecuencia, estos centros comerciales son espacios idóneos para la medición y proyección del consumo, pues no solo se tienen las transacciones explícitas realizadas en el establecimiento, sino también otras variables que confieren relevancia a este acto consumidor en el contexto de los centros comerciales, representando una significativa contribución para los mismos o un gasto adicional.

La industria en cuestión se ha enfrentado a desafíos tales como el cierre generalizado a nivel mundial provocado por la pandemia de COVID-19, inflación de dos dígitos, el aumento de la incertidumbre económica y política y una persistente tasa de desempleo cercana al 10%; sin embargo, el sector ha experimentado una recuperación parcial, alcanzando tasas de visita del 95%.

A pesar del panorama de mercado mencionado anteriormente y sumado al alza en los costos de arrendamiento en los centros comerciales, Parque Arauco ha revelado un incremento en la afluencia de visitantes y el aumento del tránsito en estos espacios. No obstante, dicho incremento no se traduce en un rendimiento proporcional en términos de consumo, considerando los gastos en inversión asociados con instalaciones, mantenimiento, seguridad y entretenimiento.

En este contexto, el grupo Parque Arauco contrató a una empresa externa la cual llevó a cabo la aplicación de múltiples encuestas, de las cuales después de la debida filtración, se obtuvieron 900 encuestas que cumplen los requisitos para ser sujetas de evaluación (respuestas completas y hay intención de compra). Estas encuestas fueron dirigidas a visitantes de los diversos centros comerciales, cada una compuesta por 42 preguntas diseñadas para proporcionar un perfil detallado de los encuestados en función de su demografía, hábitos, motivaciones y percepciones (mirar Anexo I). Estos resultados constituirán la base para la elaboración de una solución analítica integral, cuyo propósito será permitir a los centros comerciales implementar iniciativas eficientes de consumo para sus visitantes.

Este enfoque analítico se dará a conocer con recomendaciones estratégicas basadas en un modelo predictivo de compras que se hará para cinco centros comerciales y de manera independiente para cada uno; además, se entregará un perfilamiento de visitantes para cada centro comercial.

2. Objetivos

- Proponer un modelo predictivo que permita identificar las características de los clientes que mayor gasto realizan en los centros comerciales y outlets, de manera general; para poder diseñar estrategias que potencien la vinculación de clientes con esos atributos.
- Identificar a través del Análisis de correspondencias Múltiples las características demográficas de los clientes que tengan mayor probabilidad de comprar en los centros comerciales para poder segmentar y potenciar las compras de esos grupos de clientes.
- Sugerir estrategias comerciales y corporativas que permitan al grupo empresarial una toma de decisiones enfocada al perfilamiento de sus visitantes.

3. Alcance del proyecto aplicado

El presente proyecto empresarial tendrá las siguientes 2 entregas, que serán acumulativas y garantizarán el cumplimiento de cada uno de los objetivos propuestos en cada etapa.

- Entrega 1: diciembre de 2023: Documento que incluya el análisis de contexto empresarial de Parque Arauco, así como: Introducción, objetivos, problemática del proyecto y análisis descriptivo de los datos proporcionados por la empresa.
- Entrega 2: junio de 2024: Documento final del proyecto empresarial que incluye lo presentado en la entrega previa, además de la evidencia de la creación y evaluación

de modelos y técnicas analíticas, para la posterior formulación de estrategias y recomendaciones basadas en el perfilamiento de los visitantes de los centros comerciales de Parque Arauco.

4. Marco teórico

Este es un proyecto basado en técnicas para inteligencia de negocio, las herramientas, modelos y técnicas que se presentarán a continuación se seleccionaron como opciones más atractivas para una solución desde el ámbito analítico.

Dentro del ámbito del aprendizaje automático, se identifican diversos enfoques y técnicas que facultan a los sistemas automatizados para aprender y tomar decisiones basadas en datos. Dos de los paradigmas más prevalentes son los modelos de aprendizaje supervisado y no supervisado.

Con el fin de alcanzar los objetivos delineados en la presente investigación, se han examinado los siguientes modelos y técnicas analíticas para la predicción de la variable del cliente compra o no compra y la posterior segmentación de visitantes.

4.1.1. Regresión logística:

Este modelo estadístico es muy usado en la investigación, porque se concibe para elucidar la probabilidad de un evento específico que se manifiesta como binario, en valores de 0 o 1 (Falso o Verdadero, Sí o No, entre otros). La estructura logarítmica intrínseca del modelo facilita la incorporación de diversas variables explicativas (Jiménez et al., 2000) permitiendo así una amalgama de factores que influyen en la probabilidad del evento en

cuestión. La forma característica de "S" del modelo proporciona un punto de inflexión que demarca los límites en los cuales la probabilidad del evento transita desde 0 a 1, o viceversa. Este punto de corte permite determinar el rango específico en el cual la probabilidad del evento se sitúa en 0, o, en caso contrario, en 1.

4.1.2. Función de regresión logística

Definimos la variable aleatoria del modelo propuesto:

Figura 1.

Definición variable aleatoria

$$Y = \begin{cases} 1 & \text{Si el cliente Compra} \\ 0 & \text{Si el cliente no Compra} \end{cases}$$

Nota: Definición de posibles valores de la variable independiente Y para el modelo de regresión logística a aplicar para este ejercicio.

Fuente: Elaboración propia

La ecuación que define un modelo de regresión logística es la siguiente:

Figura 2.

Ecuación modelo de regresión logística

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

Nota: Ecuación general del modelo de regresión logística

Fuente: Adaptado de "La Predicción Del Rendimiento Académico: Regresión Lineal Versus Regresión Logística" por M. V. G. Jiménez, J. M. A. Izquierdo, y A. J. Blanco, 2000, p. 250.

Donde:

P es la probabilidad de que la variable respuesta (Y) sea igual a 1.

β_0 es el intercepto.

$\beta_1, \beta_2, \dots, \beta_p$ son los coeficientes que se asocian a las variables predictoras X_1, X_2, \dots, X_p consecutivamente.

Por lo tanto, después de la transformación Sigmoidal obtenemos:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)}}$$

Donde la función logarítmica se transforma por medio de la función logística (sigmoidal) para obtener la probabilidad $P(Y = 1)$

Y el siguiente termino es conocido como la función *logit* :

$$e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)}$$

Siendo β coeficientes de la regresión, que al final del ejercicio se espera que sean los valores que maximizan la probabilidad de ocurrencia de la variable objetivo, dadas las variables independientes.

Para este proyecto, este modelo es útil para responder la pregunta ¿Qué grupo/persona gastará en el centro comercial? en donde la variable objetivo será entonces el gasto esperado, siendo 0 el cliente no gastará en su visita en el centro comercial y 1, el cliente gastará en el centro comercial.

4.1.3. Aprendizaje de redes neuronales:

Las redes neuronales posibilitan la creación de modelos orientados a identificar la combinación óptima de parámetros que se ajuste a un contexto específico. Estos sistemas exhiben la capacidad de aprender y perfeccionarse mediante el proceso de entrenamiento.

Típicamente, una red neuronal se compone de tres partes principales: una capa de entrada que alberga unidades representativas de los campos de entrada, una o múltiples capas ocultas, y una capa de salida que contiene una o varias unidades representativas de los campos de destino (Schmidhuber, 2015). Las unidades establecen conexiones ponderadas entre sí. En el flujo de operación, los datos de entrada son presentados en la primera capa, y los valores se propagan desde cada neurona hasta todas las neuronas de la capa subsecuente, culminando con la emisión de un resultado desde la capa de salida.

El proceso de aprendizaje de la red implica el examen de registros individuales, generando predicciones para cada uno y ajustando las ponderaciones en caso de predicciones incorrectas. Este ciclo se repite en múltiples ocasiones, permitiendo que la red continúe mejorando sus predicciones hasta satisfacer uno o varios criterios de detención.

4.1.4. Comparación

Para el correcto desarrollo del presente proyecto empresarial, se comparan en la siguiente Tabla 1 los dos modelos analíticos descritos en esta sección en términos de interpretación, flexibilidad y mantenimiento. Como el objetivo de este ejercicio es la generación de estrategias, es imperativo asegurar el entendimiento del proceder del modelo seleccionado y de necesitarlo, la posibilidad de compartir los mismos con Parque Arauco para su futura replicación.

Tabla 1.
Comparación Modelos

	Modelo de aprendizaje de redes neuronales	Modelo de regresión logística
--	--	--------------------------------------

Interpretación	Captura relación entre variables no lineales complejas	Directa y fácil de entender ya que los coeficientes indican la magnitud de la relación entre las variables predictoras y la variable objetivo.
Flexibilidad	Es altamente flexible por lo que se hace ideal para conjunto de datos donde no se puede establecer relación fácilmente, por lo mismo requiere más tiempo para su entrenamiento	El manejo de parámetros y datos iniciales es simple y flexible especialmente en conjuntos de datos pequeños
Mantenimiento	Requiere más recursos computacionales y pueden ser más difíciles de mantener y actualizar debido a su complejidad.	Es fácil de mantener y actualizar, ya que los coeficientes del modelo pueden ser interpretados y ajustados con relativa facilidad.

Nota: Cuadro comparativo de modelos analíticos según 3 categorías: interpretación, flexibilidad y mantenimiento.

Fuente: Elaboración propia

La tabla comparativa (Tabla 1) indica que el modelo de regresión logística es el mejor adaptado a las necesidades del presente Proyecto, ya que tiene una interpretación más fácil de comprender y dado el tamaño de los datos es un modelo práctico para mantener, analizar e interpretar.

4.2. Técnicas de Analítica para Segmentación

4.2.1. K-means

El método K-means constituye un enfoque de agrupación basado en centroides, cuyo objetivo principal radica en la partición de n variables en un número predeterminado de clústeres (x) (Kliemann & Sanders, 2016). Esta estrategia tiene como propósito la agrupación de datos más similares entre sí en conjuntos distintos. Un desafío inherente a

este método reside en la optimización de los centroides, los cuales desempeñan un papel crucial en la calidad de los grupos generados.

Asimismo, la identificación del número óptimo de agrupaciones representa otro desafío relevante en la implementación de este método. Una técnica frecuentemente utilizada con este propósito es la denominada "método del codo". Este método visualiza cómo disminuye la distancia entre los puntos a medida que se incrementa el número de clústeres.

Comúnmente, se observa un punto de inflexión o "codo", donde la reducción en la distancia entre grupos se vuelve mínima después de cierto número de agrupaciones, indicando un punto óptimo que minimiza la pérdida de eficiencia computacional.

4.2.2. ACM

El Análisis de Correspondencias Múltiples (ACM) es una técnica estadística multivariada utilizada para explorar y visualizar relaciones entre múltiples variables categóricas simultáneamente (Greenacre, 1984).

Características Principales:

- **Exploración de Datos Categóricos:** ACM permite explorar patrones y relaciones entre múltiples variables categóricas en tablas de contingencia.
- **Reducción de Dimensionalidad:** Ayuda a reducir la dimensionalidad de los datos categóricos al identificar y resaltar las dimensiones subyacentes más importantes.
- **Interpretación de Resultados:** ACM permite interpretar cómo las diferentes categorías de las variables se relacionan entre sí y con otras variables en el conjunto de datos.

4.2.3. Comparación

Al comparar las dos técnicas de segmentación se procede con el ACM ya que el conjunto de datos disponibles contiene variables principalmente categóricas, de esta manera la técnica de K – Means se convierte en un método poco útil debido a su incapacidad para manejar la naturaleza discreta y no métrica de las variables categóricas

4.3. Otras Técnicas Aplicadas

Adicionalmente, en el desarrollo de los modelos, se aplicaron técnicas de balanceo, tales como:

4.3.1. Oversampling:

Esta técnica se emplea con el propósito de abordar desafíos asociados al desequilibrio en conjuntos de datos, situación en la cual una o varias categorías presentan un número significativamente inferior de registros. Esta disparidad puede resultar en dificultades para que un modelo de aprendizaje automático aprenda y prediga con precisión las categorías menos representadas. La aplicación de técnicas de sobremuestreo se erige como una práctica destinada a mejorar la capacidad del modelo para aprender y generalizar patrones de las categorías minoritarias, minimizando el riesgo de sesgo hacia las categorías con mayor prevalencia de datos.

A continuación, se describen algunas técnicas de sobremuestreo:

- **Replicación:** Este método consiste en duplicar aleatoriamente ejemplos de las categorías minoritarias del conjunto de datos, con el objetivo de incrementar la representación de las observaciones asociadas a dichas categorías.
- **Generación de ejemplos sintéticos:** En lugar de duplicar registros, esta técnica implica la generación de ejemplos sintéticos que guarden similitud con los existentes en las categorías minoritarias. Un algoritmo comúnmente utilizado para este propósito es el Método de Sobremuestreo de Minorías Sintéticas (SMOTE), que crea nuevos ejemplos interpolando características de ejemplos vecinos.
- **Adición de ruido:** En determinadas ocasiones, se introduce variabilidad adicional a los ejemplos existentes en las categorías minoritarias mediante la incorporación de pequeñas perturbaciones.

4.3.2. Undersampling:

Este procedimiento se emplea con el propósito de abordar problemáticas vinculadas al desequilibrio observado cuando una o varias categorías en un conjunto de datos presentan una cantidad significativamente mayor de registros en comparación con otras, circunstancia que podría dificultar la capacidad de un modelo de aprendizaje automático para aprender y predecir de manera precisa.

El undersampling, consiste en la reducción del número de ejemplos pertenecientes a las categorías mayoritarias, con el objetivo de equilibrar el conjunto de datos. Al disminuir la representación de ejemplos de las categorías predominantes, se propicia que el modelo aprenda patrones de manera más equitativa entre las diversas categorías.

A continuación, se describen algunas técnicas de undersampling:

- **Muestreo aleatorio:** En esta estrategia, se procede a la eliminación aleatoria de ejemplos de las categorías mayoritarias hasta alcanzar el equilibrio deseado. Este proceso se lleva a cabo sin considerar la importancia o relevancia de los ejemplos eliminados.
- **Muestreo basado en densidad:** En lugar de eliminar ejemplos de manera aleatoria, esta técnica propone la supresión de ejemplos pertenecientes a las categorías mayoritarias que se encuentran cercanos a las fronteras entre agrupaciones. La finalidad de este enfoque es conservar ejemplos informativos y mitigar el riesgo de pérdida de información valiosa.

4.3.3. Comparación

Con base a las técnicas de balanceo expuestas anteriormente y a la cantidad de datos que tienen las clases del modelo, se toma la decisión de aplicar oversampling a la clase minoritaria. Esta clase tiene muy pocas muestras en comparación con la clase mayoritaria, por lo que el modelo podría sesgarse hacia la clase mayoritaria. De esta manera, se hace un sobre muestreo que implica generar más registros de la clase minoritaria para equilibrar el conjunto de datos.

5. Metodología CRISP – DM

Según IBM 2021, esta metodología traduce *Cross-Industry Standard Process for Data Mining*, y básicamente es un marco de trabajo para proyectos orientados hacia la minería de datos. A continuación, están las diferentes fases de esta metodología, y lo que se realizara en cada una de estas:

Tabla 2.
Metodología Crisp Dm

Actividad	Descripción
Business Understanding	Los centros comerciales buscan tener el mayor número de personas posible, con intención de compra, ya que esto incrementa la probabilidad de gasto en alguno de los comercios que a su vez paga al centro comercial por sus servicios.
Data Understanding	La información suministrada ha sido recolectada mediante encuestas. Las respuestas han sido tabuladas debidamente dentro de un archivo de Excel. Son 42 preguntas de hábitos, costumbres, motivaciones e información sociodemográfica de las personas. De la misma manera, la información del archivo ya ha sido previamente filtrada: solo hay personas que fueron al centro comercial por ocio (nada relacionado con temas laborales), y personas que consideran que han estado un buen tiempo en los CCs a fin de poder proveer un poco más de detalles sobre los mismos, con el fin de poder definir cuál es el perfil de clientes que con mayor probabilidad realizarán una compra
Data Preparation	Al comprender que la base de datos contiene información de carácter categórico, fue necesaria su transformación a través de hot-coding. Además, se realizó un posterior balanceo de registros para el correcto aprendizaje de los modelos candidatos y la creación de variables adicionales que permitieran la agrupación de respuestas para un mejor análisis y entrenamiento de modelos y técnicas analíticas.
Modelling	Para cumplir con el objetivo de este proyecto, se propusieron dos tipos de modelos analíticos: aprendizaje de redes neuronales y regresión logística binomial. Previamente, se seleccionaron mediante el método stepwise las variables que más impactan en la decisión del visitante de realizar o no una compra. Paralelamente, para complementar la descripción de estos perfiles, se determinaron sus principales características demográficas mediante las técnicas analíticas de ACM y K-means.
Evaluation	A través de la aplicación de diversas técnicas estadísticas y teniendo en cuenta el objetivo del ejercicio de determinar con la mayor especificidad posible los verdaderos visitantes compradores, se procedió a la elección del modelo de regresión logística binaria y la técnica de ACM, dada la naturaleza categórica de los datos disponibles.
Deployment	Presentar recomendaciones estratégicas con base al perfilamiento de visitantes por centro comercial resultado del modelo de regresión logística y ACM elegidos.

Nota: Breve descripción de cada etapa del proyecto según la metodología Crisp DM

Fuente: Elaboración propia

6. Cronograma

Figura 3.
Cronograma de trabajo

Actividad	Oct				Dic				Feb				Abr				Jun	
	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2
Análisis descriptivo de la base de datos	★																	
Definición del marco teórico	★																	
Elaboración de los objetivos y el alcance	★																	
Evaluación de modelos analíticos				★														
Implementación de modelos elegidos				★														
Distribución de temas pendientes del P.E.				★														
Elaborar introducción				★														
Diseñar cronograma				★														
Elaborar bibliografía				★														
Definir metodología CRISP-DM	★																	
Elaborar PPT de segunda entrega del P.E.				★														
Documentar resultados del modelo en el P.E.									★									
Perfilamiento por centros comerciales									★									
Propuesta de estrategias corporativas									★									
Revisión del trabajo del P.E.									★									
Envío del word del P.E.																★		
Elaborar PPT de entrega final del P.E.									★									
Sustentación P.E.									★									

★ David C. ★ Sebastian C. ★ Juan M. ★ Camila M.

Nota: Distribución de tareas y objetivos a desarrollar a lo largo de la maestría para poder cumplir con los diferentes hitos.

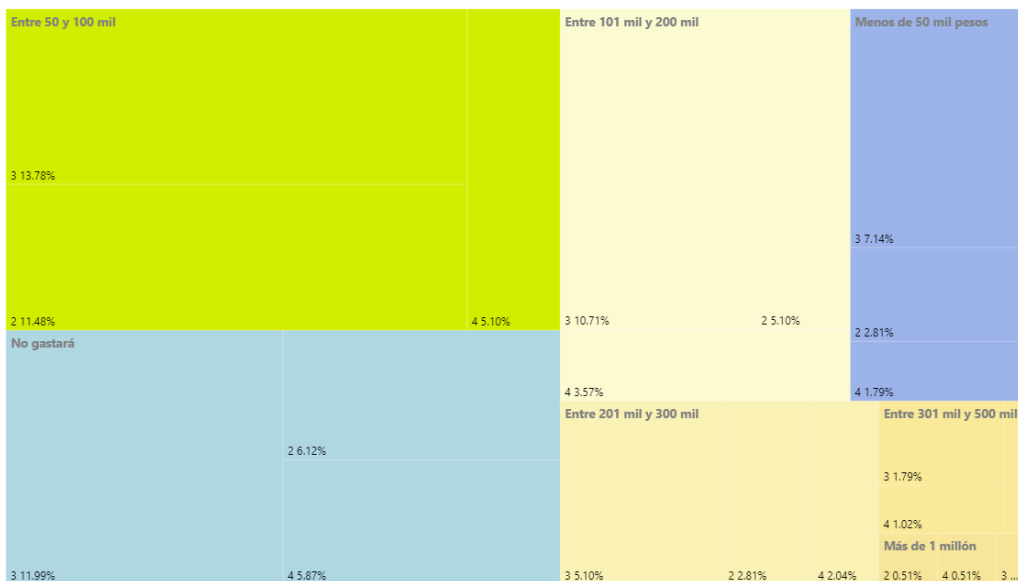
Fuente: Elaboración propia.

7. Análisis descriptivo

A continuación, se detalla el comportamiento demográfico de la población objeto de la encuesta y su interrelación con el gasto anticipado. Esta última variable representa uno de

los objetivos finales del ejercicio en términos de predicción. Adicionalmente, se detallarán algunos escenarios específicos identificados en el análisis demográfico, los cuales consideramos cruciales para una comprensión más profunda del entorno empresarial.

Figura 4.
Treemap de gasto



Nota: Distribución de disposición de gasto de todos los clientes encuestados de los distintos centros comerciales.

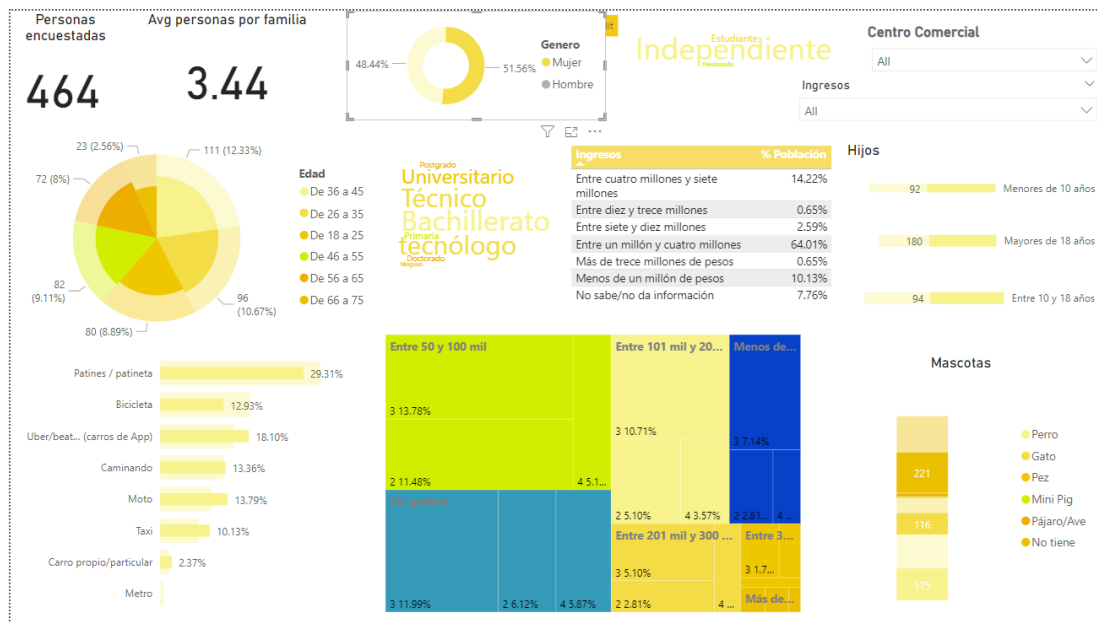
Fuente: Elaboración propia

La Figura 4, presenta un Treemap de la intención de gasto de visitantes en los Centros Comerciales en función de su estrato socioeconómico, solo está mostrando los estratos socioeconómicos 2, 3 y 4, ya que el 69.5% de la población encuestada pertenece a estos estratos. De esta manera observamos que tenemos individuos que tendrán un gasto 0 (No gastará) hasta individuos que esperan realizar un gasto hasta por 500 mil pesos (Entre 301 y 500 mil). Analizando los grupos presentes en el Treemap, es notable observar que, en términos generales, los participantes de la encuesta tienen la intención de gastar hasta \$200

mil pesos en su visita al centro comercial. Este comportamiento no implica una afluencia significativa de ingresos al establecimiento. Además, se destaca que el cuadrante de "No gastará" abarca un 24% de la intención expresada a lo largo de la encuesta. En resumen, se enfrenta a una población de estratos 2 y 3 que no percibe la visita a un centro comercial como motivo para aumentar sus gastos de manera sustancial.

A continuación, a fin de separar cualquier posible comportamiento que pueda estar sesgado por el género, se analizarán los resultados de la encuesta para cada uno de estos de acuerdo *con* los posibles resultados de la encuesta: masculina y femenina.

Figura 5.
Descripción demográfica de población femenina



Nota: Captura de pantalla de un tablero realizado en Power BI en donde se muestran los resultados a algunas preguntas de la encuesta únicamente para los que respondieron ser Mujer.

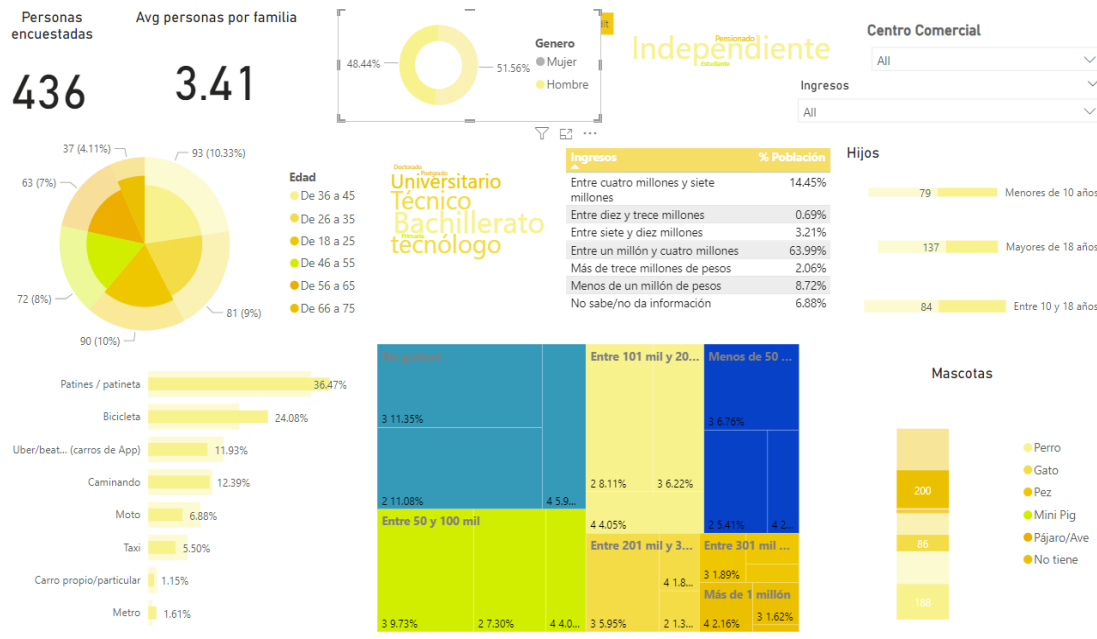
Fuente: Elaboración propia

Dado que esta muestra fue previamente balanceada para garantizar su representatividad en relación con la población, se evidencia que el 51.56% de los participantes en la encuesta son mujeres mayores de edad. La mayor concentración de este grupo demográfico, un 44%, se ubica en edades comprendidas entre los 26 y 36 años, y pertenecen a los estratos socioeconómicos 2 y 3, con educación secundaria completada, situación laboral independiente y un ingreso mensual promedio predominante en el rango de 1 a 4 millones de pesos.

Adicionalmente, se destaca que estas mujeres forman parte de familias con una media de aproximadamente 4 personas, y alrededor del 50% de esta población tiene al menos una mascota. Se presume que residen en zonas cercanas al centro comercial donde se llevó a cabo la encuesta, ya que un 29% de ellas se desplaza utilizando patines, seguido por un 18% que utiliza vehículos con servicio de conductor.

El comportamiento predominante en esta población demográfica es la disposición a gastar entre 50 y 100 mil pesos en su visita al centro comercial, supuestamente influenciado por el número de hijos, miembros familiares y mascotas en el hogar.

Figura 6.
 Descripción demográfica de población masculina

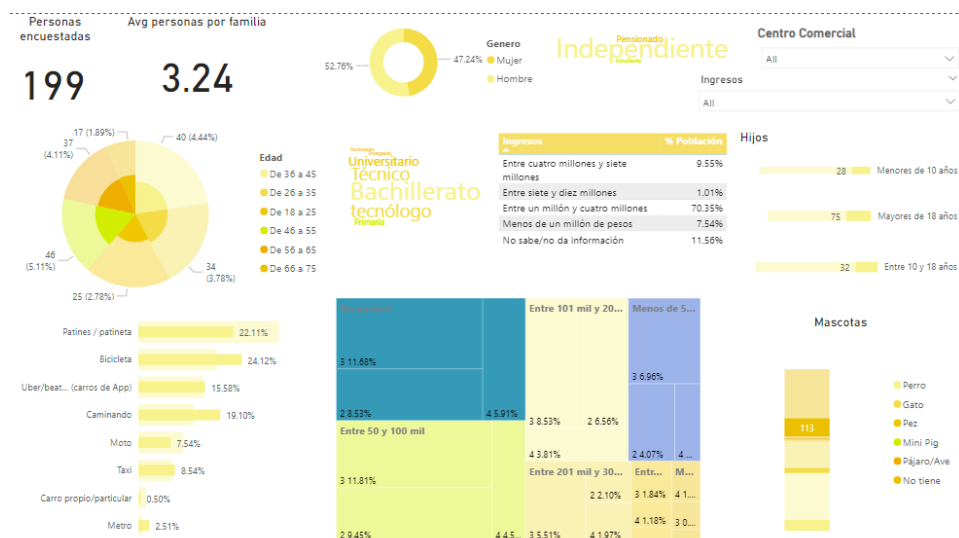


Nota: Captura de pantalla de un tablero realizado en Power BI en donde se muestran los resultados a algunas preguntas de la encuesta únicamente para los que respondieron ser Hombre.

Fuente: Elaboración propia

Al mirar la Figura 6, en donde se observan las variables demográficas y comportamentales de la población masculina podemos evidenciar que a pesar de presentar similitudes en la situación socioeconómica respecto a la Figura 5 donde analizamos las mismas variables para el género femenino, se observa un menor apetito de gasto. Este fenómeno puede atribuirse a cambios en el comportamiento masculino, evidenciados por un aumento en el uso de patinetas, una menor cantidad de mascotas e hijos.

Figura 7.
 Descripción demográfica de población que no gastará

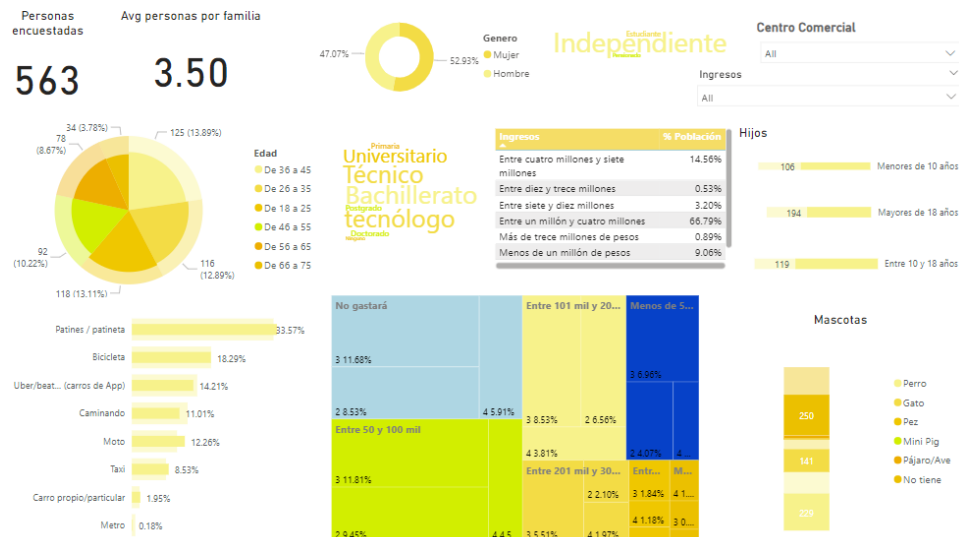


Nota: Captura de pantalla de un tablero realizado en Power BI en donde se muestran los resultados a algunas preguntas de la encuesta únicamente para los que respondieron que no van a realizar compras en el centro comercial.

Fuente: Elaboración propia

La población que ha expresado su intención de no realizar gastos se distingue por estar compuesta en un 53% por individuos del género masculino, con edades comprendidas en el rango de 36 a 55 años. Este grupo se caracteriza por no poseer mascotas, haber completado sus estudios secundarios, desplazarse hacia el centro comercial mediante bicicleta, patines o a pie, presentar un ingreso mensual situado en el intervalo de 1 a 4 millones de pesos y conformar hogares con una media de 3.2 personas, donde el 66.7% de la población tiene hijos.

Figura 8.
Descripción demográfica de población que gastará



Nota: Captura de pantalla de un tablero realizado en Power BI en donde se muestran los resultados a algunas preguntas de la encuesta únicamente para los que respondieron que van a realizar compras en el centro comercial.

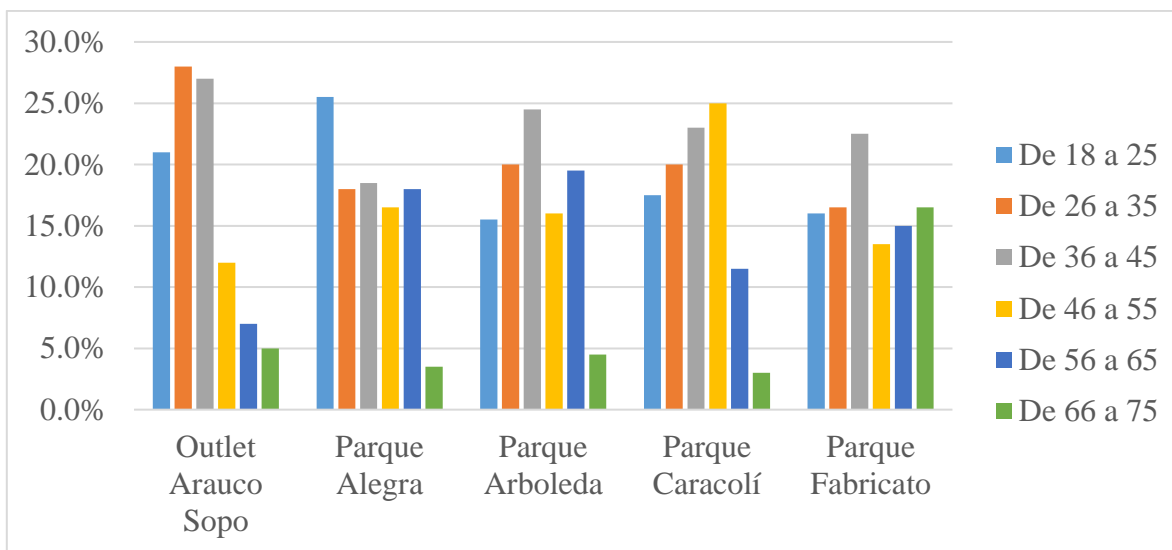
Fuente: Elaboración propia

La población que tiene la intención de realizar gastos en su visita al centro comercial está compuesta en un 53% por individuos del género femenino, con edades comprendidas entre 18 y 45 años. Este grupo demográfico se distingue por poseer niveles de educación superior. Se observa un incremento del 5% en comparación con la población que no tiene planeado realizar gastos, en términos de ingresos situados entre 4 y 7 millones de pesos. Estos individuos conforman hogares con una media de 3.5 personas, en los cuales las mascotas predominantes son perros y gatos. Además, el 74.4% de esta población pertenece a familias con hijos, registrándose un aumento del 7.8% en esta característica en comparación con la población anterior.

7.1. Comportamiento de la Población por CCs

7.1.1. Edad de Visitantes por CCs

Figura 9.
Edad de visitantes



Nota: Histograma con la distribución de los rangos de edad de los diferentes encuestados para cada uno de los CCs.

Fuente: Elaboración propia

1. El CC que tuvo mayor proporción de respondientes entre los 26 y 45 años de edad fue el CC parque Arauco. Adicionalmente, el 88% de los respondientes tiene menos de 56 años.
2. El parque Alegra fue el CC que tuvo su mayor cantidad de respondientes en la categoría de edad menor a los 25 años. Curiosamente, tiene una distribución similar en todos los rangos de edad, de aproximadamente el 18%, a excepción en el rango de mayores a 66 años, en donde tiene solo el 3,5%.

3. El CC Parque Arboleda es el que tiene la mayor proporción de respondientes entre los 56 a 65 años. Sin embargo, a partir de los gráficos, su mayor concentración está en el grupo de los 36 a 45 años.
4. El CC Parque Caracolí agrupa el 85% de su población menor a 66 años.
5. El Parque Fabricato es el que tiene la mayor concentración en el grupo de 66 a 75 años. Sin embargo, su mayor concentración de respondientes está ubicado en el rango de 36 a 45 años.

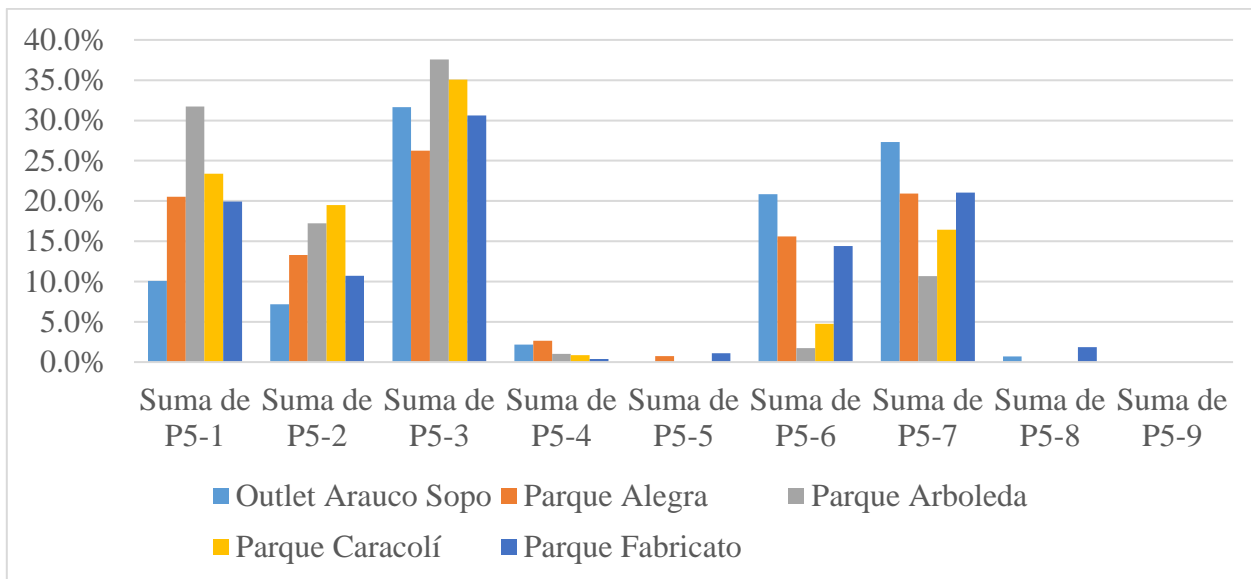
Hallazgos generales:

Todos los centros comerciales registran una participación significativa de entre el 18.5% y el 27% del total de encuestados dentro del rango de edad comprendido entre 36 y 45 años, lo que confiere a esta franja demográfica un atractivo generalizado. De manera similar, se podría afirmar que las dos categorías siguientes corresponden a los grupos de edades que les siguen, es decir, aquellos ubicados entre los 18 y 35 años. En conjunto, estas tres categorías, de las seis disponibles, aglutinan un poco más del 60% de los participantes en la encuesta.

En virtud de lo expuesto, resultará intrigante examinar la disposición de gasto de estos grupos en etapas posteriores del análisis.

7.1.2. ¿Con quién va al CC?

Figura 10.
Acompañantes



Nota: Gráfico de histogramas para detallar la participación de los distintos tipos de acompañantes

con los que acuden al CC los clientes.

Fuente: Elaboración propia

Con el análisis presentado en el gráfico, buscamos identificar el acompañante más común entre los visitantes de cada centro comercial. Antes de realizar el análisis del gráfico, a continuación, se dispone la leyenda para cada una de las respuestas:

P5-1 Solo (a), **P5-2** Con amigo/a(s), **P5-3** Familiar / familia (diferentes a pareja o hijo/nieto), **P5-4** Compañeros trabajo / estudio, **P5-5** Colega / compañero / cliente, **P5-6** Hijo(a) / nieto(a), **P5-7** Con su Pareja, **P5-8** Mascota, **P5-9** Otro.

Los resultados muestran lo siguiente:

1. Outlet Arauco Sopo: Este centro comercial concentra aproximadamente el 80% de sus visitantes en tres categorías principales, siendo la más predominante la de ir con

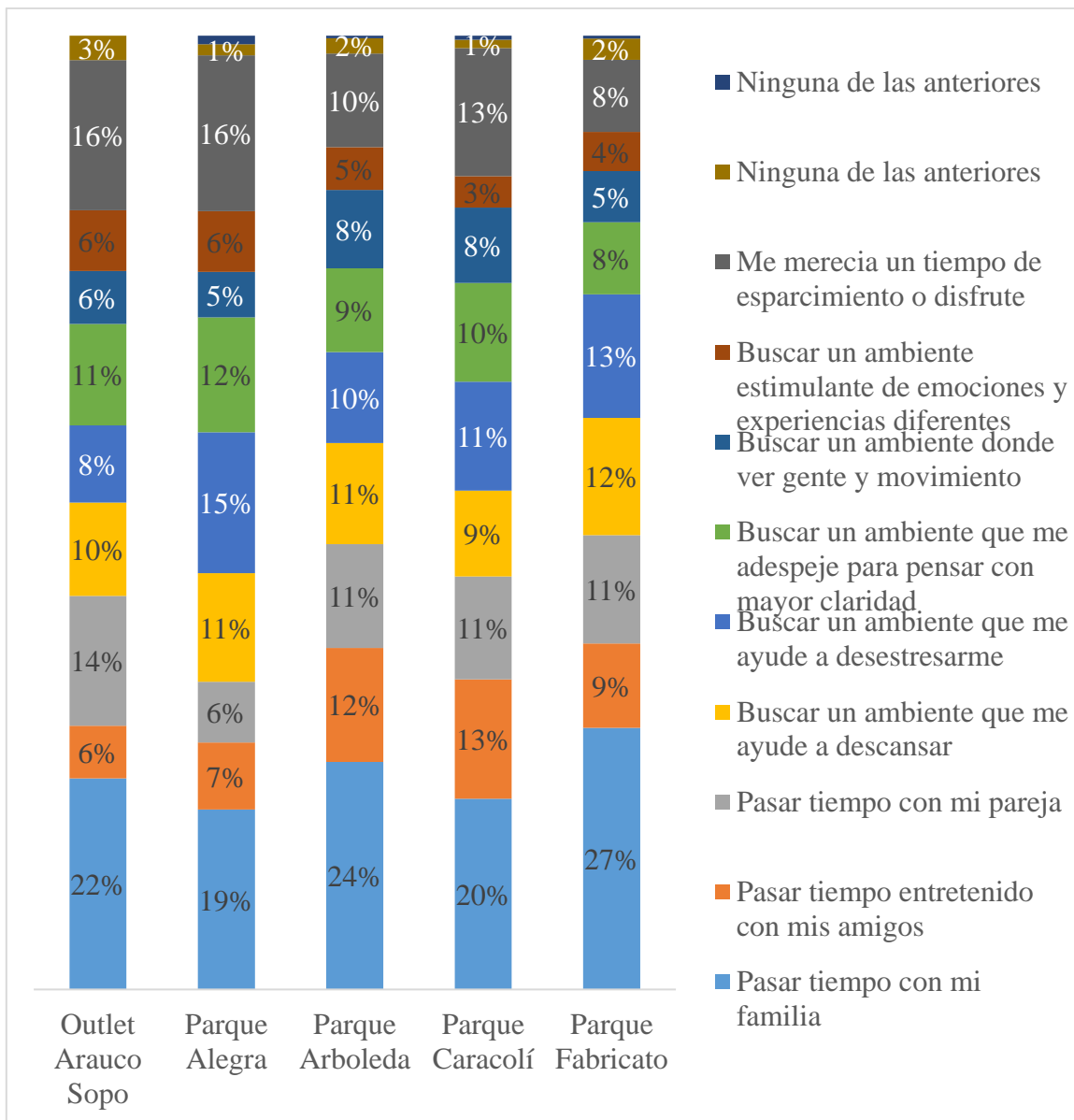
familiares, seguida por la opción de asistir con pareja y con hijos. Estos resultados indican que la mayoría de los clientes prefieren visitar este establecimiento acompañados, aunque cuenta con amplios espacios, siendo menos frecuente la presencia de mascotas, amigos o colegas.

2. Parque Alegre: Cerca del 48% de los visitantes acuden con pareja o hijos, mientras que aproximadamente el 20% opta por visitar el centro comercial de manera individual.
3. Parque Arboleda: La mayoría de los encuestados que visitan este centro comercial indican que lo hacen con familiares, seguido por la opción de asistir solo y, en menor medida, con amigos. Ningún participante menciona ir con mascotas, y solo el 10.7% señala acudir con su pareja.
4. Parque Caracolí: Este centro comercial presenta la mayor proporción de encuestados que indican asistir con amigos, seguido por visitas en compañía de familiares y la opción de asistir solo.
5. Parque Fabricato: Aunque solo el 1.8% de los encuestados menciona asistir con mascotas, este centro comercial comparte patrones de comportamiento similares con Parque Alegre y Parque Caracolí.

Para todos los centros comerciales se evidencio un alto % de respuestas para la opción #3. Sería interesante investigar si hay una mayor propensión de estas personas a responder en comparación con otros grupos o si simplemente representan la mayoría de los asistentes.

7.1.3. Motivación Visita al CC

Figura 111.
Motivaciones de visitas



Nota: Gráfico de Columnas apiladas en donde se quiere observar cual es la respuesta con mayor participación, independiente del número de respondientes teniendo en cuenta que hay más encuestados en algunos CCs que en otros.

Fuente: Elaboración propia

En todos los centros comerciales analizados, la mayoría de los encuestados manifestaron que su principal motivo de visita es pasar tiempo con la familia, esta tendencia es particularmente notable en el Parque Fabricato, donde casi el 28% de los encuestados señalaron dicho motivo como predominante. Por otro lado, los centros comerciales Parque Arboleda y Parque Caracolí presentaron una proporción significativa de encuestados que indicaron visitar el lugar con amigos, lo cual podría estar relacionado con la facilidad de acceso en comparación con el Outlet, donde la vía de acceso más conveniente suele ser mediante vehículo propio.

8. Transformación de datos

Esta encuesta se entregó al grupo Parque Arauco en un archivo .xlsx junto con el formulario Arquetipo para convertir variables de ser necesario, este ejercicio lo ejecutó una empresa externa experta en la realización encuestas, por lo que no necesitó la depuración de los datos.

Esta base de datos cuenta inicialmente con 900 registros anonimizados, cada uno con un total de 208 columnas, una por cada opción de respuesta, que para fácil entendimiento los encuestadores asignaron un número a cada opción de respuesta, sin embargo, esta conversión no permite el análisis de las variables ya que sujeta su peso en el valor numérico asignado.

Por lo anterior, se realizó la recodificación de las variables convirtiéndolas en binarias, en donde 1 corresponde a se selecciona esa opción de respuesta y 0 en caso contrario.

Adicional a esto, se observó que la pregunta 20 relacionada con identificación de características fue formulada a través de 23 literales calificados mediante una escala de 1 a 4, siendo 1 “Estoy totalmente de acuerdo con la afirmación” y 4 “Estoy totalmente en desacuerdo con la afirmación” tal como lo muestra la siguiente tabla:

Tabla 3.
Pregunta 20 - Posibles respuestas

Afirmación	Totalmente de acuerdo	Algo de acuerdo	Algo en desacuerdo	Totalmente en desacuerdo
A. En esta etapa de mi vida estoy más enfocado(a) en producir ingresos que en mi bienestar físico o emocional	1	2	3	4
B. En vez de viajar prefiero comprar productos	1	2	3	4
C. Como te ves te juzgan	1	2	3	4
D. Dicen que la plata no es felicidad, pero la verdad es que sí es muy importante para ser feliz	1	2	3	4
E. Prefiero construirlo o arreglarlo, que comprarlo nuevo	1	2	3	4
F. Suelo antojarme de experiencias excitantes	1	2	3	4
G. Disfruto salir a distraerme o entretenerme yo solo(a), sin la compañía de otras personas	1	2	3	4
H. Cuando me siento bajo(a) de ánimo o cuando algo me intranquiliza, prefiero quedarme en casa en vez de salir	1	2	3	4
I. Compartir con mis amigo(a) es demasiado importante para mi felicidad y estabilidad emocional	1	2	3	4
J. Me considero más intelectual que social	1	2	3	4
K. Es sano evitar estar todo el tiempo con la familia	1	2	3	4
L. Me considero una persona introvertida	1	2	3	4
M. Me tienen sin cuidado la moda y las tendencias	1	2	3	4
N. Me gusta mucho experimentar cosas nuevas	1	2	3	4
O. Me considero de mentalidad conservadora	1	2	3	4
P. Mi perfil es más del lado humano y de las artes que del lado analítico y numérico	1	2	3	4
Q. Me gustaría vivir en otros países	1	2	3	4
R. Me gusta estar siempre en actividad, preferiblemente fuera de mi casa	1	2	3	4
S. Soy más práctico que filosófico	1	2	3	4
T. Las redes sociales, en general, son negativas para la sociedad	1	2	3	4

U. Me enfoco más en precio que en calidad o marca	1	2	3	4
V. En este momento de mi vida, la prioridad es mi familia	1	2	3	4
W. Casi siempre lo mejor es lo más caro	1	2	3	4
X. Cuando debo contactar al servicio al cliente de una empresa, prefiero que sea en persona o, en su defecto, por teléfono	1	2	3	4
Y. Me considero una persona digital	1	2	3	4

Nota: Listado de afirmaciones expuestas a encuestados con su escala de evaluación según afinidad de este.

Fuente: Arquetipo encuesta, Parque Arauco, 2023.

Esta pregunta arroja datos que transformados y agrupados de la manera adecuada significan información valiosa para la definición de perfiles de los visitantes, por lo que las afirmaciones que describen cierta orientación/perfil se agruparon en 9 secciones enlistadas a continuación y se promediaron el valor de las respuestas para disminuir el número de variables evaluadas, de esta manera, se generan recomendaciones acotadas en un mejor espectro.

Tabla 4.

Orientaciones identificadas pregunta 20 Cuestionario arquetipos grupo de Centros Comerciales

Código orientación	Nombre/Descripción orientación	Literales incluidos
ORIDIIN	Individual	A-D
ORIPRE	Interés en el precio	U-W
ORIEXP	Interés en experiencias	N-Q-R
ORIIMA	Interés por cuidar su imagen	C-M
ORIHUMA	Humano	E-P-S
ORISOC	Social	G-H-I-J-L
ORISTA	Status Quo - Conservador	O-T
ORIFAM	Familiar	K-V
ORIDIG	Digital	X-Y

Nota: Orientaciones de comportamiento identificadas según segmentación de respuestas en afinidad con afirmaciones mencionadas.

Fuente: Elaboración propia

Una vez realizada la adición de estas 9 orientaciones, se obtiene un archivo xlsx (Anexo BDD_Rosario) con un total de 217 variables, clasificadas en la Tabla 3 según el tipo de pregunta:

Tabla 5.
Clasificación de pregunta

Tipo de pregunta	Preguntas incluidas
Demográfica	A, E, F, G, 23, 24, 25, 26, 27, 28, 29, 30, 31,32, 33, 34, 35.
Hábitos de consumo	B, C, D, 2, 3, 5, 6, 13, 14, 15, 16, 21
Motivación	7, 8, 10,11
Percepción	1, 9, 12, 19, 20

Nota: División en 4 categorías de las 42 preguntas del arquetipo de encuesta aplicado por Parque Arauco, estas categorías permiten evaluar las respuestas desde frentes delimitados y con diferentes rutas de acción.

Fuente: Elaboración propia

9. Descripción de la Situación organizacional donde se realizará el proyecto

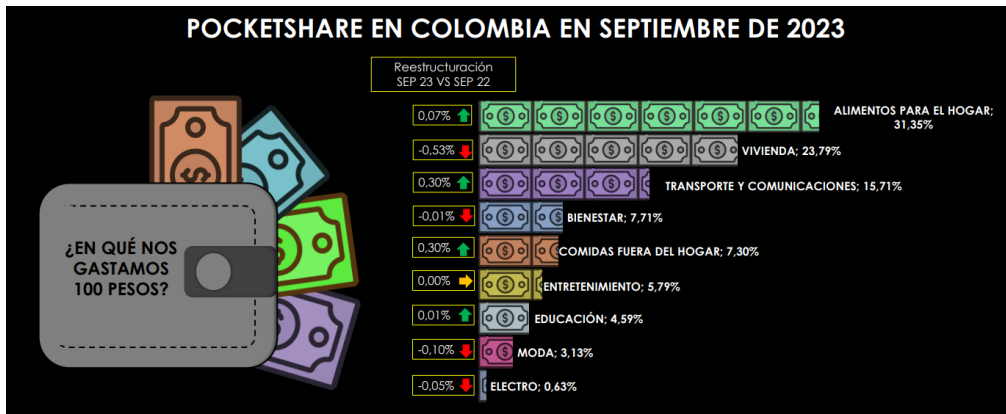
En la actualidad, a nivel global, nos encontramos frente a un fenómeno económico sin precedentes que surgió a raíz de la pandemia. Posterior al impacto inicial de la crisis sanitaria, se ha observado un cambio significativo en los hábitos de consumo, donde factores como la experiencia, el gusto y la calidad han ganado relevancia. Anteriormente, los determinantes clave para realizar una compra se limitaban al precio y la satisfacción de necesidades específicas.

Esta transformación ha llevado a las empresas del sector comercial a adaptarse de manera acelerada, reconfigurando tanto los beneficios asociados a la adquisición de productos como las estrategias de comercialización. Aunque inicialmente el impacto fue positivo, con un mantenimiento del poder adquisitivo y la disponibilidad de productos, tres años después de este fenómeno, aún experimentamos sus consecuencias. La escasez de mano de obra operativa, la falta de materia prima, conflictos bélicos, el desempleo y desplazamientos forzados de naciones enteras son algunas de las ramificaciones que persisten.

En este contexto, Colombia se sitúa como el segundo país de Latinoamérica con las tasas de interés más elevadas, según el informe de Portafolio publicado en septiembre de 2023 (Gomez Guasca, 2023). Con un aumento de la tasa de interés real al 7,23%, el país se ve superado únicamente por Brasil, que registra un 8,96%. Esta situación ha planteado desafíos significativos para el sector de consumo, focalizando la atención en la búsqueda de estrategias para recuperar utilidades y mantenerse relevante en el mercado.

Considerando estos factores reunidos, podemos obtener una comprensión más abarcadora del pocketshare de Colombia para septiembre de 2023, según el informe de Raddar (Olaya, 2023). El análisis compara dicho mes con el mismo periodo del año 2022, evidenciando una transformación en los hábitos de consumo de los colombianos. Esta evolución se traduce en una mayor inversión en la satisfacción de necesidades primarias, dado que la situación actual limita las posibilidades de los hogares. En consecuencia, se observa una renuncia a actividades asociadas con el entretenimiento, el bienestar o el confort que va más allá de lo básico.

Figura 122.
Pocketshare en Colombia.



Nota: Evolución de la intención de gasto por tipo de servicio/bien de los colombianos medido entre septiembre 2022 y septiembre 2023

Fuente: Olaya, 2023

Para abordar este fenómeno, junto con el aumento de las compras en línea, es imperativo que los centros comerciales redefinan su actividad económica. Un estudio realizado por Fenalco (Analitik, 2022) revela que el comportamiento de las personas ha experimentado cambios significativos, transformándose de lugares principalmente dedicados a la adquisición de bienes a espacios destinados a la visita de establecimientos donde se sirven comidas y bebidas.

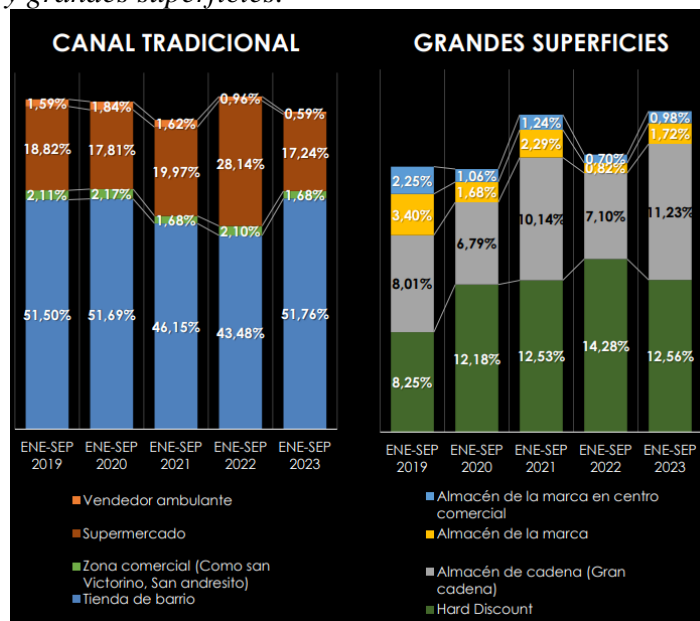
En 2023, Colombia cuenta con 254 centros comerciales activos, lo que equivale a 11.5 m² de centro comercial por cada 100 habitantes. En este contexto, la clave para destacar en este sector no solo radica en ofrecer una variedad de locales disponibles, sino también en proporcionar nuevas experiencias, servicios y productos diversos. Esta importancia se refleja en la Figura 13, con gráficos de barras que muestran los lugares preferidos por los hogares colombianos para realizar compras y sus fluctuaciones entre

2019 y septiembre de 2023 además se puede observar la variabilidad de porcentaje de distribución de compra a lo largo del periodo de tiempo mencionado anteriormente dividida en canales tradicionales y almacenes de grandes superficies,

Se observa un cambio en la tendencia, donde los hogares tienden a optar por lugares de compra más cercanos a sus lugares de residencia, como las tiendas de barrio. Sin embargo, cuando las compras superan el volumen de consumo diario, los consumidores se desplazan hacia almacenes de grandes superficies. Estos lugares ofrecen una variedad de productos, marcas privadas y artículos adicionales a la canasta familiar, sin incurrir en costos significativos de desplazamiento.

Figura 133.

Canal tradicional y grandes superficies.



Nota: Evolución de plazas de compra de los consumidores colombianos

Fuente: Olaya, 2023

Dadas estas condiciones, los centros comerciales deben abordar las nuevas dinámicas de consumo de manera integral, transformándose en espacios que no solo ofrezcan productos de primera necesidad, sino que también proporcionen servicios que vayan más allá del simple intercambio comercial, convirtiéndose en experiencias para los consumidores. Es crucial mantener la proximidad con el cliente final y reconocer que estas alternativas requieren un esfuerzo significativo por parte de cualquier grupo empresarial. Estas iniciativas deben ser diseñadas estratégicamente, orientadas a un grupo específico de clientes que estén dispuestos a retribuir la inversión mediante un mayor gasto en sus visitas. En este contexto, los modelos integrales de analítica desarrollados en este proyecto desempeñarán un papel fundamental al influir en la toma de decisiones de este grupo empresarial.

10. Descripción de la situación estudio de caso y/o problemática empresarial y método y/o estrategia a aplicar para su solución

Parque Arauco es una compañía latinoamericana con una trayectoria de más de 40 años en el sector de rentas inmobiliarias, especializándose en el desarrollo y operación de propiedades multiformato, con un enfoque particular en Centros Comerciales (CCs, en plural o CC, en singular). La empresa, con sede principal en Santiago, Chile, cuenta con alrededor de 300 empleados.

En el contexto colombiano, Parque Arauco inició sus operaciones en 2008 (Bejarano, 2023) con la apertura del Centro Comercial Parque Arboleada en la ciudad de Pereira. Para el año 2021, la compañía ya contaba con tres Centros Comerciales, un Outlet y, tras una transacción de acciones en el Parque La Colina, inauguró su cuarto Centro

Comercial, Parque Alegra, ubicado en la ciudad de Barranquilla. Esta expansión representó una inversión y adecuación aproximada de \$434.000 millones de COP (Argote, 2021).

Este movimiento estratégico, sumado a los planes de expansión en el país y los cambios en los hábitos de consumo postpandemia, resalta la importancia no solo de análisis financieros para la toma de decisiones, sino también de un análisis detallado de los visitantes de los Centros Comerciales. Aspectos como sus patrones de compra, intereses y características demográficas se vuelven cruciales para determinar la mejor manera de consolidar la marca Parque Arauco en Colombia y garantizar un aumento sostenible del gasto por parte de la población visitante en los centros comerciales actuales y futuros.

Este análisis, como ha sido previamente expuesto en el presente informe, se abordará utilizando principalmente los datos obtenidos a través de la encuesta implementada por el grupo empresarial. El propósito es generar recomendaciones fundamentadas en los resultados derivados de modelos predictivos y estrategias analíticas. Estas recomendaciones estarán orientadas hacia la identificación de grupos de interés específicos. De esta manera, los líderes y directivos del grupo empresarial estarán en posición de tomar decisiones estratégicas, así como abordar tareas tácticas y operativas con el objetivo de potenciar sus ingresos y asegurar un crecimiento sostenido en el ámbito nacional.

11. Evaluación de modelo candidato

En los siguientes capítulos se realizó una comparación de posibles modelos a aplicar para este trabajo. Esto con el fin de usar la herramienta que demuestre ser la mejor (tanto en

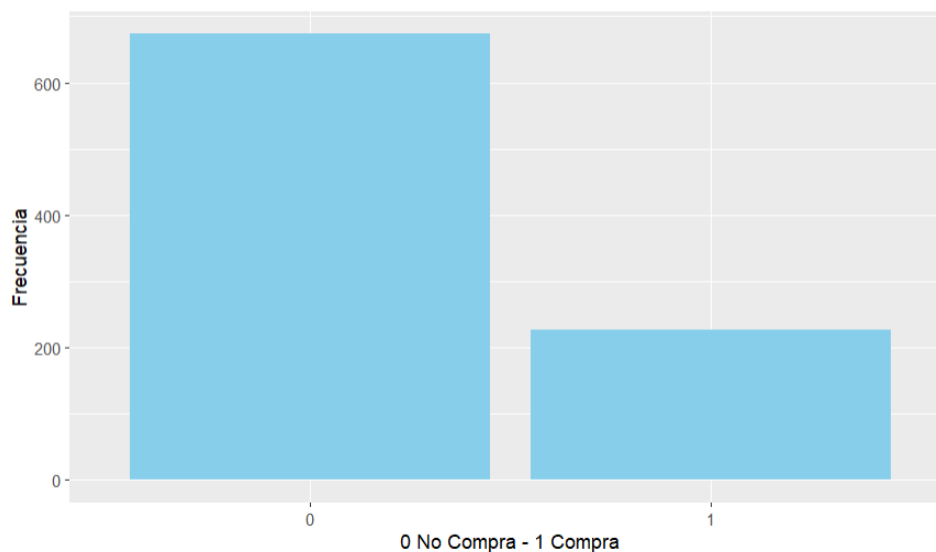
indicadores como en su entendimiento de los procesos que realizó) para poder realizar las recomendaciones y análisis a la información suministrada por Parque Arauco.

11.1. Regresión Logística Binomial

La base de datos cuenta con la siguiente información para la construcción de un modelo de regresión logística donde la variable dependiente es binaria (compra o no compra), las clases se representan por los valores 0 y 1 en donde:

- Clase 0 (no compra): 674 observaciones
- Clase 1 (compra): 226 observaciones

Figura 144.
Comportamiento variable dependiente



Nota: Histograma de clasificación de las variables para Compra (1) y No Compra (0)

Fuente: Elaboración propia

Es importante tener en cuenta la proporción entre las clases, ya que un desequilibrio significativo puede afectar el rendimiento del modelo. En este caso, la clase 0 tiene una proporción más alta en comparación con la clase 1 por lo que se aplicará la técnica de oversampling para balancear las clases, vale la pena mencionar que, de no utilizarse la técnica de oversampling se genera un sesgo que sobre estima la clase mayoritaria y subestima la clase minoritaria, es decir, las predicciones del modelo con los datos originales estarán predeterminados por la clase mayoritaria, en este caso, la clase 0.

11.1.1. Balanceo de Datos

Para poder trabajar de forma adecuada el modelo de regresión logística se tendrán en cuenta las siguientes consideraciones:

Detección de desequilibrio entre clases: El desequilibrio de clases puede afectar la capacidad del modelo para generalizar patrones en la clase minoritaria. Se probaron técnicas como el undersampling y el oversampling de la clase mayoritaria en donde se evidencia una mejora del rendimiento del modelo cuando se aplica el sobremuestreo. Además, para comprobar la correcta aplicación de esta técnica se construyó el modelo de Regresión Logística utilizando la fuente de datos sin balancear y, posteriormente se utilizaron la muestra de datos balanceada, a continuación, se presentan los principales hallazgos que respaldan la decisión de proceder con una muestra de datos balanceada a través de la técnica de sobremuestreo.

Tabla 6.
Matrices de Confusión Modelo

Matriz de Confusión Modelo Sin Balancear			Matriz de confusión Modelo Balanceado		
	0	1		0	1
0	23	28	0	135	38
1	30	103	1	8	89

Nota: Valores de resultados de predicciones de modelo de regresión logística binaria con el conjunto de datos sin balancear y balanceado clasificados en verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

Fuente: Elaboración propia

Las matrices de confusión permiten comparar la cantidad de aciertos y errores para cada modelo (Fawcett, 2006), en este caso, el modelo de Regresión Logística sin muestras balanceadas presenta más falsos positivos y negativos, ligado a la mayor presencia de datos de clase 0 (El cliente o compra).

Tabla 7.
Estadísticos de los Modelos

	Modelo Sin Balancear	Modelo Balanceado
Accuracy	0.6848	0.8296
95% CI	(0.6123, 0.7512)	(0.7794, 0.8725)
No Information Rate	0.712	0.5296
P-Value [Acc > NIR]	0.8153	0.0000000000000022
Kappa	0.2227	0.6535
Mcnemar's Test P-Value	0.8955	0.00001904
Sensitivity	0.4340	0.9441
Specificity	0.7863	0.7008
Pos Pred Value	0.4510	0.7803
Neg Pred Value	0.7744	0.9175
Prevalence	0.2880	0.5296
Detection Rate	0.1250	0.5000
Detection Prevalence	0.2772	0.6407
Balanced Accuracy	0.6101	0.8224

Nota: Comparativo de principales métricas estadísticas para el modelo de regresión logística usando datos balanceados y no balanceados

Fuente: Elaboración propia

Las métricas presentadas en la Tabla 7 juegan un papel crítico en la evaluación de estos modelos (Sheskin, 2020), ya que proporcionan información adicional y permiten tomar decisiones fundamentadas basadas en datos. En el contexto de este proyecto, se destaca la precisión, que indica la proporción de predicciones correctas, así como la sensibilidad, que mide la tasa de verdaderos positivos. Además, el coeficiente kappa (Carletta, 2012) es notable, especialmente para el modelo de regresión logística balanceado, ya que indica un desempeño significativamente superior en la clasificación precisa de datos respecto a una clasificación aleatoria y al modelo sin balancear.

En resumen, por los estadísticos presentados en la Tabla 7 y las matrices de confusión de la Tabla 6 hay evidencia estadística suficiente para el desarrollo de recomendaciones y estrategias basadas en el resultado del Modelo Balanceado.

11.1.2. Selección de Variables

Una vez seleccionado el modelo analítico, se procede a seleccionar las variables predictoras a través del método stepwise cuya combinación óptima permite mejorar el rendimiento del modelo aplicado.

El método de selección de variables paso a paso (stepwise) utilizado para refinar el modelo, considera tanto la adición como la eliminación de variables en el proceso de selección (*Análisis de Regresión Múltiple*, s. f.). La función stepwise se utiliza para realizar la selección de variables paso a paso en modelos lineales generalizados (GLM), basándose en el criterio de información de Akaike (AIC) o el criterio de información bayesiano (BIC). El modelo resultante incluye las variables $X_2 + X_3 + X_{39} + X_{30} + X_{80} + X_{81} + X_{142} + X_{252} + X_{250} + X_{244}$. Las variables seleccionadas son aquellas que el método stepwise considera más relevantes para el modelo.

La figura 15 permite observar las respuestas más relevantes para predecir la variable dependiente X_{161} (El cliente compra o no compra) y en la tabla 7 se definen las variables asociadas a dichas respuestas para mayor comprensión del lector.

Figura 155.*Método de selección de variables*

```

Call:
glm(formula = X161 ~ X2 + X3 + X39 + X30 + X80 + X81 + X142 +
      X252 + X250 + X244, family = binomial(link = "logit"), data = train,
      maxit = 100)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.49621 0.78992 -3.160 0.001577 **
X22 0.50205 0.23227 2.161 0.030661 *
X23 -0.34780 0.22426 -1.551 0.120932
X24 -0.86939 0.22876 -3.800 0.000144 ***
X25 2.08618 0.49132 4.246 0.0000218 ***
X32 0.08024 0.15764 0.509 0.610746
X391 1.31038 0.60710 2.158 0.030895 *
X301 -0.75053 0.18633 -4.028 0.0000563 ***
X80Positivo 0.97449 0.22751 4.283 0.0000184 ***
X81Positivo 0.69818 0.19232 3.630 0.000283 ***
X1422 18.92944 456.96730 0.041 0.966958
X1423 18.61110 882.52833 0.021 0.983175
X1424 17.70401 1626.84554 0.011 0.991317
X1425 18.44848 3497.74836 0.005 0.995792
X1428 19.87510 6522.63860 0.003 0.997569
X252 -0.20956 0.12397 -1.690 0.090940 .
X250 0.38448 0.11158 3.446 0.000569 ***
X244 0.22929 0.11740 1.953 0.050816 .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1494.2 on 1077 degrees of freedom
Residual deviance: 963.7 on 1060 degrees of freedom
AIC: 999.7

Number of Fisher Scoring iterations: 17

```

Nota: Variables seleccionadas del modelo de regresión logística binomial para predecir la probabilidad de compra bajo la metodología stepwise,

Fuente: Elaboración propia

El resultado final es un modelo de regresión logística simplificado que utiliza las variables $X2 + X3 + X39 + X30 + X80 + X81 + X142 + X252 + X250 + X244$. En la siguiente tabla, se encuentra la leyenda asociada a cada variable:

Tabla 8.
Descripción de Variables Significativas

Código variable	Descripción
X2	Ubicación de Centro Comercial
X3	Género del visitante
X39	Motivación de visita al Centro Comercial
X30	Compañía en visita al Centro Comercial
X80	Sentimiento optimista al visitar el Centro Comercial
X81	Motivado al visitar el Centro Comercial
X142	Tipo de compra en el Centro Comercial
X252	Orientación digital
X250	Orientación social
X244	Orientación individual

Nota: Código y descripción de las variables significativas para el modelo de regresión logística aplicado.

Fuente: Elaboración propia

11.1.3. Evaluación de Odds

De manera continua con el análisis que se viene realizando, en la Tabla 9 se encuentra un resumen de comportamiento de las variables independientes con respecto a la probabilidad de compra que tiene un cliente. Cabe resaltar que dentro de la tabla, la columna Odds hace referencia a la medida estadística utilizada para evaluar la relación entre variables dependientes e independientes (Hosmer et al., 2013).

Tabla 9.
Odds Ratios

Variable	Respuesta asociada a la variable	Odds	Std.Error	Statistic	P.Value
(Intercept)		-2,50	0,79	-3,16	0,00
X2-2	Medellín	0,50	0,23	2,16	0,03
X2-3	Bucaramanga	-0,35	0,22	-1,55	0,12
X2-4	Pereira	-0,87	0,23	-3,80	0,00
X2-5	Sopó	2,09	0,49	4,25	0,00
X3-2	Mujer	0,08	0,16	0,51	0,61
X39-1	Comer/Almorzar	1,31	0,61	2,16	0,03
X30-1	Solo	-0,75	0,19	-4,03	0,00
X80Positivo	Si	0,97	0,23	4,28	0,00
X81Positivo	Si	0,70	0,19	3,63	0,00
X142-2	Intención de compra en 2 categorías del Centro Comercial	18,93	456,97	0,04	0,97
X142-3	Intención de compra en 3 categorías del Centro Comercial	18,61	882,53	0,02	0,98
X142-4	Intención de compra en 4 categorías del Centro Comercial	17,70	1626,85	0,01	0,99
X142-5	Intención de compra en 5 categorías del Centro Comercial	18,45	3497,75	0,01	1,00
X142-8	Intención de compra en 8 categorías del Centro Comercial	19,88	652,26	0,00	1,00
X252	Orientación digital	-0,21	0,12	-1,69	0,09
X250	Orientación social	0,38	0,11	3,45	0,00
X244	Orientación individual	0,23	0,12	1,95	0,05

Nota: Listado de variables significativas con sus respectivos valores de intercepto y p value, lo que permite determinar cuánto y como la variable independiente afecta la probabilidad de ocurrencia de la variable dependiente.

Fuente: Elaboración propia

- Intercepto (Intercept)

Un intercepto de -2.50 con un valor p de 0.0016 indica que las odds de compra son significativamente bajas cuando todas las demás variables en el modelo son cero, es decir, sin la presencia o impacto de las demás variables predictoras analizadas a continuación, la

probabilidad que el cliente realice una compra al momento de la visita es de aproximadamente 7.69% dadas la siguiente ecuaciones (Hosmer et al., 2013):

Ecuación 1. Odds Ratio

$$Odds = e^{intercept} \rightarrow probabilidad = \frac{Odds}{1+Odds}$$

- X22 – El Centro Comercial se encuentra en Medellín

Un odds ratio de 0.502 indica que la razón de probabilidad de que un cliente realice una compra es aproximadamente la mitad entre aquellos que visitan el centro comercial de Medellín en comparación con aquellos que no visitan, teniendo todo lo demás constante. El valor p (0.0307) indica que este efecto es estadísticamente significativo, lo que significa que la asociación entre visitar el centro comercial de Medellín y la probabilidad de realizar una compra es improbable que se deba al azar.

- X23 – El Centro Comercial se encuentra en Bucaramanga

Un coeficiente negativo (-0.348) sugiere que, en promedio, la variable dependiente es 0.348 unidades más baja para el centro comercial en Barranquilla en comparación con los centros comerciales de las demás ubicaciones. Sin embargo, el valor p (0.1209) es mayor que 0.05, indicando que el efecto de estar en Barranquilla no es estadísticamente significativo.

- X24 – El Centro Comercial se encuentra en Pereira

Un coeficiente negativo (-0.869) sugiere que, en promedio, la variable dependiente es 0.869 unidades más baja para el centro comercial en Pereira en comparación con los centros comerciales de las demás ubicaciones. Ahora bien, el valor p (0.0001) es mucho menor que el umbral de 0.05, indicando que el efecto de estar en Pereira es estadísticamente significativo.

- X25 – El Centro Comercial se encuentra en Sopó

Un coeficiente de (208.618) sugiere que, en promedio, la variable dependiente es 208.618 unidades más alta para el centro comercial de Sopó en comparación con los centros comerciales de las demás ubicaciones. Ahora bien, el valor p (0.00002) es mucho menor que el umbral de 0.05, indicando que el efecto de la diferencia es estadísticamente significativo.

- X32 - Género mujer

El coeficiente positivo de (0.080) sugiere un ligero incremento en la variable dependiente para mujeres, sin embargo, este aumento no es estadísticamente significativo dado el valor p (0.6107). Esto implica que no hay evidencia suficiente para afirmar que el género mujer tiene un impacto significativo.

- X391 – El motivo de la visita es Comer/Almorzar

Un coeficiente de (131.038) sugiere que, en promedio, la variable dependiente es 131.038 unidades más alta cuando el motivo de la visita es comer/almorzar en comparación

con otros motivos. Este efecto es estadísticamente significativo, dado el valor p (0.0309), lo que implica que hay evidencia suficiente para afirmar que el motivo de la visita tiene un impacto positivo en la variable dependiente.

- X301 – Visita solo el centro comercial

Un coeficiente de (-0.751) sugiere que, en promedio, la variable dependiente es 0.751 unidades más baja cuando la persona visita el centro comercial sola en comparación cuando la visita acompañada. Este efecto es estadísticamente significativo, dado que el valor p (0.00006) es muy bajo, lo que implica que hay evidencia suficiente para afirmar que el motivo de visitar solo el centro comercial tiene un impacto negativo significativo en la variable dependiente.

- X80Positivo - Sentimiento optimista al visitar el Centro Comercial

Un coeficiente de (0.974) sugiere que, en promedio, la variable dependiente es 0.974 unidades más alta cuando el sentimiento del visitante es optimista en comparación con otros sentimientos. Este efecto es estadísticamente significativo, dado que el valor p (0.000018) es muy bajo, lo que implica que hay evidencia suficiente para afirmar que un sentimiento optimista tiene un impacto positivo significativo en la variable dependiente.

- X81Positivo - Motivado al visitar el Centro Comercial

Un coeficiente de (0.698) sugiere que, en promedio, la variable dependiente es 0.698 unidades más alta cuando el visitante está motivado. Este efecto es estadísticamente

significativo, dado que el valor p (0.000283) es muy bajo, lo que implica que hay evidencia suficiente para afirmar que un visitante motivado tiene un impacto positivo significativo en la variable dependiente.

- X1422- Intención de compra en 2 categorías del Centro Comercial

Como el coeficiente es tan extremadamente grande (1.892.943.848) se sugiere revisar los datos para descartar que no haya errores. Adicionalmente, un valor p de (0.967) indica que esta estimación no es estadísticamente significativa. Por lo que se concluye que no hay evidencia suficiente para afirmar que la intención de compra en 2 categorías del Centro Comercial tiene un impacto significativo en la variable dependiente.

- X1423 - Intención de compra en 3 categorías del Centro Comercial

Como el coeficiente es tan extremadamente grande (1.861.109.575) se sugiere revisar los datos para descartar que no haya errores. Adicionalmente, un valor p de (0.983) indica que esta estimación no es estadísticamente significativa. Por lo que se concluye que no hay evidencia suficiente para afirmar que la intención de compra en 3 categorías del Centro Comercial tiene un impacto significativo en la variable dependiente.

- X1424 - Intención de compra en 4 categorías del Centro Comercial

Como el coeficiente es tan extremadamente grande (1.770.401.020) se sugiere revisar los datos para descartar que no haya errores. Adicionalmente, un valor p de (0.991) indica que esta estimación no es estadísticamente significativa. Por lo que se concluye que no hay

evidencia suficiente para afirmar que la intención de compra en 4 categorías del Centro Comercial tiene un impacto significativo en la variable dependiente.

- X1425 - Intención de compra en 5 categorías del Centro Comercial

Como el coeficiente es tan extremadamente grande (1.844.848.330) se sugiere revisar los datos para descartar que no haya errores. Adicionalmente, un valor p de (0.996) indica que esta estimación no es estadísticamente significativa. Por lo que se concluye que no hay evidencia suficiente para afirmar que la intención de compra en 5 categorías del Centro Comercial tiene un impacto significativo en la variable dependiente.

- X1428 - Intención de compra en 8 categorías del Centro Comercial

Como el coeficiente es tan extremadamente grande (1.987.510.078) se sugiere revisar los datos para descartar que no haya errores. Adicionalmente, un valor p de (0.998) indica que esta estimación no es estadísticamente significativa. Por lo que se concluye que no hay evidencia suficiente para afirmar que la intención de compra en 5 categorías del Centro Comercial tiene un impacto significativo en la variable dependiente.

- X252 - Orientación digital

El coeficiente (-0.210) sugiere una ligera disminución en la variable dependiente, ante un aumento en la orientación digital del visitante, esta disminución no es estadísticamente significativa dado el valor p de (0.0909). Esto implica que no hay evidencia suficiente para afirmar que la orientación digital tiene un impacto significativo en la variable dependiente.

- X250 - Orientación social

Un coeficiente de (0.384) sugiere que, en promedio, la variable dependiente aumenta en 0.384 unidades por cada unidad de aumento en la orientación social del visitante. Este efecto es estadísticamente significativo dado el valor bajo de su valor p (0.000569). Lo que implica que hay evidencia suficiente para afirmar que una mayor orientación social, tiene un impacto positivo significativo en la variable dependiente.

- X244 - Orientación individual

Un coeficiente de (0.229) sugiere que, en promedio, la variable dependiente aumenta en 0.229 unidades por cada unidad de aumento en la orientación social del visitante. Este efecto no es estadísticamente significativo dado su valor p (0.0508). Lo que implica que no hay evidencia suficiente para afirmar que una mayor orientación social tiene un impacto significativo en la variable dependiente.

11.1.4. Pseudo-R² de McFadden

A continuación, se realiza el análisis del pseudo-R² de McFadden. Este indicador es una medida de bondad de ajuste utilizada en el modelo de regresión logística (Hu et al., 2006). Este tipo de R² es "pseudo" porque no tiene la misma interpretación directa que el R² clásico en modelos lineales, pero proporciona una forma de cuantificar la mejora en el ajuste del modelo de regresión logística en comparación con un "modelo nulo" o "modelo base"

Figura 166.*Pseudo-R² de McFadden.*

```

fitting null model for pseudo-r2
McFadden
0.3550319

```

Nota: Se indica el nivel de ajuste del modelo a los datos disponibles a través del R² McFadden, es decir el poder predictivo del modelo según el entrenamiento y los datos disponibles para el mismo

Fuente: Elaboración propia

El valor de pseudo-R² de McFadden obtenido para el modelo de 0.3550319 implica que el modelo ajustado explica aproximadamente un 14.8% la variabilidad en los datos en comparación con el modelo nulo que solo incluiría la constante. Es decir, el Modelo Balanceado tiene una mejor capacidad de predicción frente al Modelo Sin Balancear, por lo cual para los objetivos planteados para este trabajo significa un mayor grado de confiabilidad en los resultados otorgados.

11.1.5. Evaluación Importancia de Variables (VAR IMP)

Considerando el número de variables independientes involucradas en el modelo, es crucial evaluar su relevancia y su impacto para priorizar los esfuerzos futuros destinados a incrementar la probabilidad de compra en cada uno de los CC (Hosmer et al., 2013).

Figura 177.*Importancia de las variables.*

Variable	VAR IMP	Respuesta Asociada a la Variable
X2-2	2,1614	Medellín
X2-3	1,5509	Bucaramanga
X2-4	3,8005	Pereira

X2-5	4,2460	Sopó
X3-2	0,5090	Mujer
X39-1	2,1584	Comer/Almorzar
X30-1	4,0279	Solo
X80Positivo	4,2832	Si
X81Positivo	3,6304	Si
X142-2	0,0414	Intención de compra en 2 categorías del Centro Comercial
X142-3	0,0211	Intención de compra en 3 categorías del Centro Comercial
X142-4	0,0109	Intención de compra en 4 categorías del Centro Comercial
X142-5	0,0053	Intención de compra en 5 categorías del Centro Comercial
X142-8	0,0030	Intención de compra en 8 categorías del Centro Comercial
X252	1,6905	Orientación digital
X250	3,4457	Orientación social
X244	1,9530	Orientación individual

Nota: Tabla donde se enumeran las variables significativas, su importancia, y la respuesta asociada a la Variable

Fuente: Elaboración propia

Entre las variables más significativas se encuentran la ubicación geográfica de los encuestados, con Sopó y Pereira, así como factores personales como el estado civil y la orientación social. También se observa una marcada inclinación hacia la respuesta positiva ante ciertas situaciones (X80Positivo y X81Positivo con valores de 4,2832 y 3,6304 respectivamente). Por otro lado, las intenciones de compra en distintas categorías dentro del centro comercial presentan valores menores de importancia, indicando una menor influencia en el comportamiento global del consumidor. Estos resultados subrayan la relevancia de factores personales y contextuales sobre las decisiones de compra, proporcionando una valiosa perspectiva para el desarrollo de estrategias y personalización de ofertas en el entorno de un centro comercial.

11.1.6. Variance Inflation Factor (VIF)

El VIF cuantifica cuánto aumenta la varianza de la estimación de un coeficiente de regresión debido a la multicolinealidad que se presenta cuando dos o más variables predictoras en un modelo de regresión están altamente correlacionadas con otras variables predictoras. En otras palabras, si el VIF es alto para una variable, significa que esa variable está altamente correlacionada con otras variables predictoras en el modelo (Hosmer et al., 2013).

Figura 188.
Índice VIF

Variable	VIF	DF
X2	1,0297	4
X3	1,0091	1
X39	1,0211	1
X30	1,0110	1
X80	1,0240	1
X81	1,0316	1
X142	1,0000	5
X252	1,0229	1
X250	1,0576	1
X244	1,0488	1

Nota: Resultado del VIF para cada una de las variables significativas a ser incorporadas en el modelo. El valor observado nos permite indicar que hay evidencia estadística para afirmar que las variables están bajamente correlacionadas.

Fuente: Elaboración propia

GVIF (Generalized VIF): Este es el valor del VIF para cada variable predictora en el modelo. Las variables X2, X3, X39, X30, X80, X81, X142, X252, X250 y X244 tienen valores puntuales de GVIF.

Df (Degrees of Freedom): Este es el número de grados de libertad asociados con cada variable o el número de parámetros estimados para cada variable en el modelo.

$\text{GVIF}^{(1/(2*\text{Df}))}$: Es una transformación del GVIF que se utiliza para interpretar la magnitud del efecto de la multicolinealidad entre variables. Un valor cercano a 1 indica que la multicolinealidad no es un problema importante.

En el modelo propuesto, los valores de GVIF son relativamente bajos para todas las variables (cerca de 1), lo cual es positivo porque nos indica ausencia de multicolinealidad. Valores bajos de VIF sugieren que la multicolinealidad no es un problema significativo en el modelo candidato. Las variables X2, X3, X39, X30, X80, X81, X142, X252, X250 y X244 tienen GVIF cercanos a 1, lo que indica que la multicolinealidad no afecta negativamente a esas variables.

Matriz de Confusión y Estadísticos

Predicción	Actual	
	0	1
0	135	38
1	8	89

Figura 199.

Coefficientes de rendimiento del Modelo

Accuracy	0.8296
95% CI	(0.7794, 0.8725)
No Information Rate	0.5296
P-Value [Acc > NIR]	< 0.000000000000000022
Kappa	0.6535
McNemar's Test P-Value	0.00001904
Sensitivity	0.9441
Specificity	0.7008
Pos Pred Value	0.7803

Neg Pred Value	0.9175
Prevalence	0.5296
Detection Rate (Recall)	0.5000
Detection Prevalence	0.6407
Balanced Accuracy	0.8224

Nota: Presenta los principales coeficientes de rendimiento del modelo de clasificación binaria

Fuente: Elaboración propia

A continuación, analizaremos a mayor profundidad los coeficientes de Rendimiento del Modelo Propuesto de Clasificación Binaria:

Accuracy (Precisión): 0.8296

La precisión indica la proporción de predicciones correctas en relación con todas las observaciones, el modelo propuesto tiene una precisión del 82.96%, lo que indica que aproximadamente dos tercios de las predicciones estas clasificadas de forma correcta.

Intervalo de Confianza del 95%: 0.77942, 0.8725

El intervalo de confianza proporciona una estimación de la variabilidad con respecto a la precisión. Con un 95% de confianza, se espera que la precisión real del modelo esté entre el rango de 77.94% a 87.25%.

No Information Rate (Tasa de No Información): 0.5296

La tasa de no información muestra la proporción de la clase más frecuente en los datos del modelo en la variable respuesta. En este caso, aproximadamente el 52.96% de las observaciones pertenecen a la clase más frecuente.

P-Value [Acc > NIR]: 0.000000000002

El valor p compara la precisión del modelo con la tasa de no información. Un valor p bajo indica que el modelo supera significativamente la tasa de no información.

Kappa: 0.6535

El coeficiente Kappa de Cohen cuantifica la concordancia de las predicciones del modelo propuesto y las clases reales, ajustando la concordancia esperada al azar. Un valor de Kappa de 0.6535 indica una concordancia bastante buena.

Mcnemar's Test P-Value: 0.00001904

La prueba de McNemar contrasta las clasificaciones erróneas entre dos modelos. Un valor p alto sugiere que no hay una diferencia significativa entre las clasificaciones incorrectas de este modelo en comparación con otro modelo propuesto.

Sensitivity (Sensibilidad): 0.9441

La sensibilidad evalúa que tan capaz es el modelo para identificar correctamente los casos de la clase positiva. El modelo identifica correctamente el 94.41% de los casos positivos.

Specificity (Especificidad): 0.7008

La especificidad calcula la capacidad del modelo propuesto para identificar de forma correcta las instancias de la clase negativa. Para el modelo propuesto, el modelo propuesto identifica correctamente el 70.08% de las instancias negativas.

Pos Pred Value (Valor Predictivo Positivo): 0.7803

El valor predictivo positivo es la proporción de predicciones positivas correctas en relación con todas las predicciones positivas. Para el modelo propuesto el 78.03% de las predicciones positivas son correctas.

Neg Pred Value (Valor Predictivo Negativo): 0.9175

El valor predictivo negativo es la proporción de predicciones negativas correctas en relación con todas las predicciones negativas. Para el modelo propuesto, el 91.75% de las predicciones negativas son correctas.

Prevalence (Prevalencia): 0.5296

La prevalencia es la proporción de la clase positiva en los datos. Para el modelo propuesto es aproximadamente del 52.96% de las observaciones pertenecen a la clase positiva.

Detection Rate (Tasa de Detección - Recall): 0.5000

La tasa de detección es la proporción de instancias positivas que el modelo ha identificado correctamente. Para el modelo propuesto se identifica correctamente el 50% de las instancias positivas.

Detection Prevalence (Prevalencia de Detección): 0.6407

La prevalencia de detección es la proporción de instancias que el modelo clasifica como positivas. Alrededor del 64.07% de las observaciones se clasifican como positivas.

Balanced Accuracy (Precisión Equilibrada): 0.8224

La precisión equilibrada es la media aritmética de la especificidad y sensibilidad y obteniendo así una medida balanceada del rendimiento del modelo propuesto. La precisión equilibrada es del 82.24%.

Según las métricas del modelo, este exhibe una Tasa de Detección del 50.00%, lo que indica que puede capturar correctamente la mitad de los casos positivos reales. Aunque esta métrica podría parecer modesta a primera vista, es respaldada por una alta Sensibilidad del 94.41%, reflejando una capacidad robusta para identificar verdaderos positivos entre los casos positivos reales. Esta combinación sugiere que, si bien el modelo no detecta todos los casos positivos, su enfoque en maximizar la identificación de verdaderos positivos es efectivo. Además, el modelo logra una Precisión global del 82.96%, lo que indica que es preciso en general en sus predicciones. Esta Tasa de Detección, junto con otras métricas sólidas como la Sensibilidad y la Precisión, posiciona al modelo como adecuado para

aplicaciones donde la detección precisa de casos positivos es crítica, balanceando eficacia en la detección con una sólida precisión global.

La Figura 21 muestra los resultados del modelo de regresión logística desarrollado con diferentes puntos de corte. Al realizar esta técnica en el modelo nos permitirá ajustar el mismo para poder encontrar la mayoría de clientes que compran ya que permitirá maximizar a los verdaderos positivos dentro de la capacidad de predicción.

Figura 200.

Comparación de la precisión del modelo con diferentes puntos de corte

Threshold	Accuracy	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
Optimal (46%)	82,96%	94,41%	70,08%	78,03%	91,75%	52,96%	50,00%	64,07%	82,24%
10%	9,26%	3,92%	100,00%	100,00%	5,77%	94,44%	3,70%	3,70%	51,96%
20%	20,37%	16,08%	93,33%	97,62%	6,14%	94,44%	15,19%	15,56%	54,71%
30%	37,41%	35,69%	66,67%	94,79%	5,75%	94,44%	33,70%	35,56%	51,18%
40%	52,22%	52,94%	40,00%	93,75%	4,76%	94,44%	50,00%	53,33%	46,47%
50%	64,44%	66,67%	26,67%	93,92%	4,49%	94,44%	62,96%	67,04%	46,67%
60%	68,15%	70,59%	26,67%	94,24%	5,06%	94,44%	66,67%	70,74%	48,63%
70%	70,00%	72,55%	26,67%	94,39%	5,41%	94,44%	68,52%	72,59%	49,61%
80%	71,85%	74,90%	20,00%	94,09%	4,48%	94,44%	70,74%	75,19%	47,45%
90%	74,07%	77,26%	20,00%	94,26%	4,92%	94,44%	72,96%	77,41%	48,63%

Nota: Muestra la evaluación de diferentes indicadores con respecto a diferentes puntos de corte

Fuente: Elaboración propia

La figura anterior evalúa el modelo de regresión logística en diferentes puntos de corte, lo que permite determinar el mejor umbral de probabilidad. Este umbral define cuándo el modelo de regresión logística binaria predice la pertenencia a la clase positiva (por encima del umbral, el visitante realiza una compra) o a la clase negativa (por debajo del umbral, el visitante no realiza una compra).

De manera paralela, los resultados de esta Figura se pueden analizar en función de las siguientes métricas::

Precisión: La proporción de casos correctamente clasificados.

Sensibilidad: La proporción de casos positivos correctamente clasificados.

Especificidad: La proporción de casos negativos correctamente clasificados.

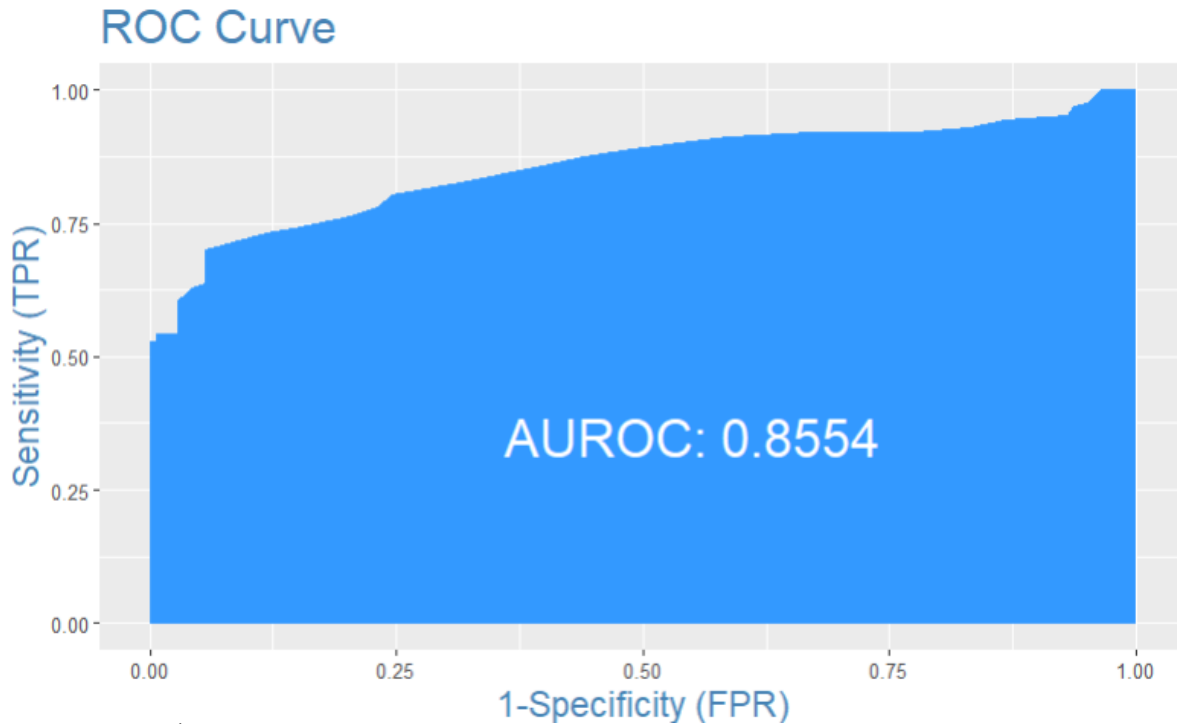
Valor predictivo positivo: La probabilidad de que un caso clasificado como positivo sea realmente positivo.

Valor predictivo negativo: La probabilidad de que un caso clasificado como negativo sea realmente negativo.

Precisión equilibrada: La media ponderada de la precisión y la sensibilidad.

En general, los resultados del modelo son buenos. El punto de corte óptimo es 45,99%, que proporciona una precisión del 82,96%, una sensibilidad del 94,41% y una especificidad del 70,08%. Este punto de corte también proporciona una precisión equilibrada del 82,24%. Por lo tanto, el punto de corte óptimo proporciona un muy buen equilibrio entre la sensibilidad y la especificidad. También proporciona una precisión equilibrada del 82.24%, que es un valor muy bueno para el modelo propuesto.

Figura 21.
Curva ROC



Nota: Área bajo la curva del modelo de regresión logística binario aplicado a la totalidad de datos balanceados disponibles, en el que se evidencia el buen rendimiento del modelo propuesto para detectar verdaderos positivos.

Fuente: Elaboración propia

La curva ROC (Receiver Operating Characteristic) y su métrica asociada, el Área bajo la curva ROC (AUROC), permite evaluar el modelo propuesto en el proyecto. La AUROC varía entre 0 y 1, donde un valor de 0.5 indica un rendimiento similar al azar y un valor de 1.0 indica un rendimiento aproximadamente perfecto.

El modelo propuesto arroja un valor de AUROC de 0.8554, lo que sugiere que puede destacar la capacidad de discriminación para clasificar las clases positivas y negativas respectivamente.

Para una capacidad de discriminación óptima, se busca que el AUROC se acerque a 1, indicando un rendimiento cercano a la perfección en la discriminación entre clases. Un valor de 0.8554 indica que el modelo está cerca de ese objetivo, pero aún hay margen de mejora.

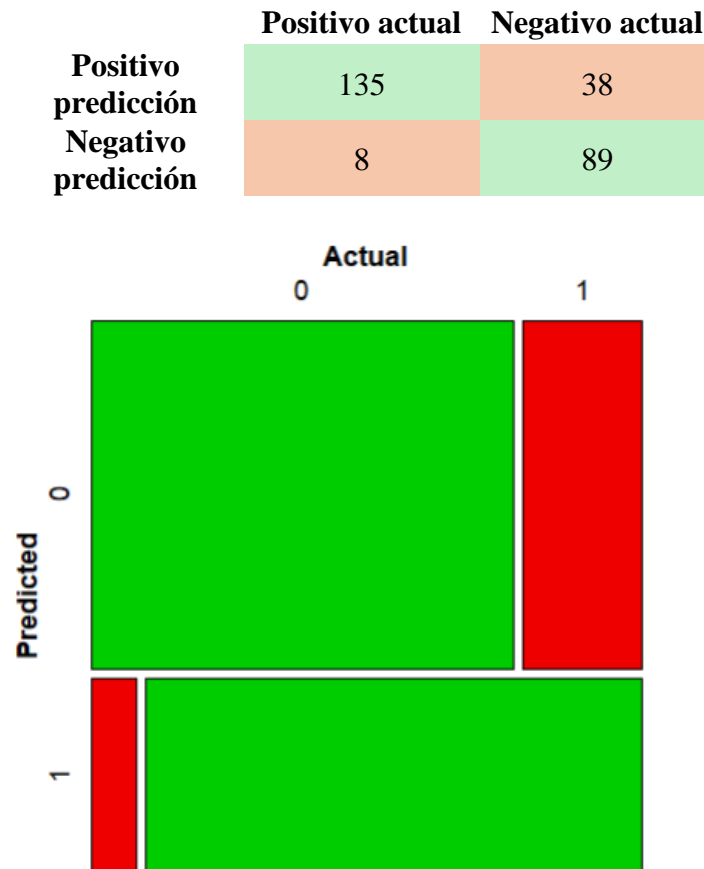
Se recomienda explorar posibles mejoras en el modelo, como la inclusión de nuevas variables o la modificación en la codificación de las existentes. También se puede considerar probar otros tipos de algoritmos de clasificación para verificar si pueden ofrecer un rendimiento aún mejor.

En comparación con otros modelos, se realizaron pruebas con diferentes puntos de corte en modelos de regresión logística, y se encontró que el punto de corte óptimo calculado es de 0.4599, logrando el mejor accuracy entre las distintas configuraciones evaluadas.

En resumen, un valor de AUROC de 0.8554 indica un rendimiento destacado del modelo.

Figura 22.
Matriz de confusión

Matriz de Confusión y Estadísticos



Nota: Matriz de confusión del modelo de regresión logística binaria aplicado a todos los Centros Comerciales, donde se evidencia el buen comportamiento en predecir los verdaderos positivos y los verdaderos negativos.

Fuente: Elaboración propia

Según el rendimiento del modelo, la matriz de confusión presenta los siguientes resultados:

- **True Positive (TP): 135** Clientes que realmente compraron y fueron predichos como compradores.

- **True Negative (TN): 89** Clientes que realmente no compraron y fueron predichos como no compradores
- **False Positive (FP): 8** Clientes que realmente no compraron, pero fueron predichos como compradores
- **False Negative (FN): 38** Clientes que realmente compraron, pero fueron predichos como no compradores.

11.2. Evaluación Modelos:

Durante la fase de implementación de modelos predictivos para el conjunto de datos, se exploraron dos enfoques principales: un modelo de Regresión Logística Binomial y uno de Redes Neuronales. Para determinar el modelo más adecuado, se hicieron análisis exhaustivos de indicadores clave para evaluar y seleccionar la opción que mejor se ajustara a los datos y proporcionara predicciones más precisas.

Ambos modelos, el de Regresión Logística Binomial y el de Redes Neuronales, fueron entrenados y evaluados utilizando métricas fundamentales como precisión, sensibilidad, especificidad y la curva ROC. En última instancia, se concluye que el modelo de Regresión Logística Binomial exhibe un rendimiento predictivo superior al modelo de Redes Neuronales.

Razones para escoger el modelo binario:

- **Superioridad en Precisión:** Se evidenció una precisión notablemente superior en el modelo de regresión logística binomial (85%) en comparación con el modelo de

redes neuronales (65%). Esta diferencia destaca la mayor eficacia y fiabilidad del modelo binomial en la predicción precisa de las clases en nuestra tarea específica.

- **Mejora en Sensibilidad:** El modelo de regresión logística binomial exhibió una sensibilidad del 94%, indicando una capacidad óptima para identificar correctamente los casos de clase positiva. Esta sensibilidad es significativamente superior a la alcanzada por el modelo de redes neuronales, que registró un nivel del 62% para los casos positivos.
- **Buena Especificidad:** El modelo de regresión logística binomial exhibe una capacidad del 70% para identificar correctamente las instancias de la clase negativa, en comparación con el modelo de redes neuronales, que logró un nivel de especificidad del 68%.
- **Mejora en Curva ROC:** La curva ROC del modelo de regresión logística binomial (86%) muestra un área bajo la curva (AUC) más amplia en comparación con el modelo de redes neuronales (77%). Este indicador resulta crucial para evaluar la capacidad de discriminación del modelo, donde un mayor nivel de AUC sugiere un mejor rendimiento en la clasificación.
- **Simplicidad Interpretativa:** El modelo de regresión logística binomial se caracteriza por su simplicidad y facilidad de interpretación, en contraste con el modelo de redes neuronales. Esta simplicidad interpretativa resulta valiosa, especialmente al comunicar resultados a equipos no técnicos y al exponer los resultados del proceso.

El modelo de regresión logística seleccionado predice de manera excepcional a los verdaderos compradores (alta sensibilidad), aunque a costa de su especificidad. Dado que el

objetivo principal de este proyecto empresarial es identificar a los compradores, se desarrollarán perfiles detallados utilizando el modelo resultante. La situación sería distinta si el objetivo fuera clasificar correctamente a los visitantes que no compran (clase negativa).

11.3. Resultados Regresión Logística por CC

Para seleccionar las métricas con mejor desempeño de cada ciudad y explicarlas, podemos considerar aquellas métricas que indiquen el mejor desempeño del modelo en aspectos clave como la precisión, la capacidad de distinguir entre clases, y la robustez de las predicciones. En la Tabla 9 se ubica un resumen donde se muestra la distribución de variables significativas según cada modelo ejecutado, con la evaluación de las métricas de rendimiento para cada uno de los modelos podemos confirmar que el resultado de estos modelos es confiable y podemos proceder a la generación de estrategias soportadas en los mismos.

Tabla 10.
Resumen Resultados Regresiones Logísticas

	Barranquilla	Medellin	Bucaramanga	Pereira	Sopo
Accuracy	0.6538	0.7895	0.6944	0.7073	0.7419
95% CI	(0.5091, 0.7803)	(0.6611, 0.8862)	(0.5747, 0.7976)	(0.5446, 0.8387)	(0.5539, 0.8814)
No Information Rate	0.5192	0.5263	0.6528	0.6341	0.5806
P-Value [Acc > NIR]	0.03477	0.00003578	0.2708	0.2104	0.04809
Kappa	0.3004	0.5793	0.2997	0.4029	0.4918
Mcnemar's Test P-Value	0.23859	0.7728	0.5224	0.3865	0.28884
Sensitivity	0.7778	0.7667	0.4800	0.7333	0.8462
Specificity	0.5200	0.8148	0.8085	0.6923	0.6667
Pos Pred Value	0.6364	0.8214	0.5714	0.5789	0.6471
Neg Pred Value	0.6842	0.7586	0.7451	0.8182	0.8571
Prevalence	0.5192	0.5263	0.3472	0.3659	0.4194
Detection Rate	0.4038	0.4035	0.1667	0.2683	0.3548
Detection Prevalence	0.6346	0.4912	0.2917	0.4634	0.5484
Balanced Accuracy	0.6489	0.7907	0.6443	0.7128	0.7564

Nota: Estas métricas son un resumen del comportamiento del modelo de regresión logística para cada uno de los Centros Comerciales, confirmando la validez de aplicación de este para cada uno de ellos.

Fuente: Elaboración propia

De manera siguiente, se muestra y explica las mejores métricas para cada ciudad:

11.3.1. Barranquilla

- Sensitivity (Sensibilidad) - 0.7778

Explicación: La sensibilidad del modelo en Barranquilla es alta, indicando que es efectivo en la identificación de verdaderos positivos. Esto es crucial en aplicaciones donde detectar la clase positiva es más importante.

- Detection Rate (Tasa de Detección) - 0.4038

Explicación: Esta métrica muestra que el modelo detecta correctamente el 40.38% del total de observaciones, lo cual es un buen indicador de rendimiento en la detección.

11.3.2. Medellín

- Accuracy (Precisión) - 0.7895

Explicación: El modelo en Medellín tiene una alta precisión general, indicando que el 78.95% de las predicciones totales son correctas.

- Kappa - 0.5793

Explicación: Un valor de Kappa de 0.5793 sugiere que el modelo tiene una buena concordancia entre las predicciones y las observaciones reales más allá del azar.

11.3.3. Bucaramanga

- Specificity (Especificidad) - 0.8085:

Explicación: La especificidad del modelo en Bucaramanga es alta, lo que significa que el modelo es eficaz en la identificación de verdaderos negativos, minimizando falsos positivos.

- Neg Pred Value (Valor Predictivo Negativo) - 0.7451

Explicación: Este valor muestra que el 74.51% de las predicciones negativas son correctas, lo que es útil en contextos donde es importante minimizar la tasa de falsos negativos.

11.3.4. Pereira

- Balanced Accuracy (Precisión Balanceada) - 0.7128

Explicación: La precisión balanceada es un buen indicador en situaciones de clases desbalanceadas, mostrando un rendimiento general equilibrado del modelo en Pereira.

- Neg Pred Value (Valor Predictivo Negativo) - 0.8182

Explicación: Con un valor predictivo negativo de 0.8182, el modelo es muy confiable en sus predicciones negativas.

11.3.5. Sopó

- Sensitivity (Sensibilidad) - 0.8462:

Explicación: Una alta sensibilidad del 84.62% indica que el modelo en Sopó es muy eficaz en la detección de verdaderos positivos.

- Balanced Accuracy (Precisión Balanceada) - 0.7564:

Explicación: La precisión balanceada alta indica que el modelo tiene un buen rendimiento general, manejando bien tanto verdaderos positivos como verdaderos negativos.

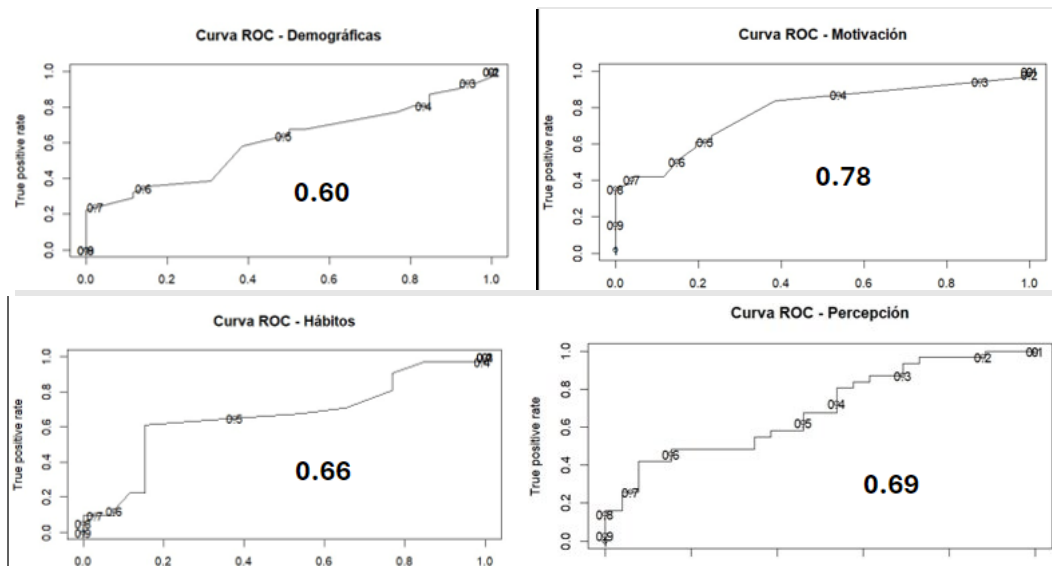
11.4. Modelo de Regresión Logística por Tipo de pregunta

En esta sección se presentarán las conclusiones principales derivadas de la aplicación del modelo de regresión logística binomial como fundamento para los modelos analíticos asociados a cada tipo de pregunta en la encuesta (Tabla 4), es decir, a partir de la segmentación de las preguntas en 4 distintas clases (Demográficas, Hábitos, Percepción, Motivación), se aplicara una regresión línea binomial que permitirá identificar cuál de estas categorías tiene una mayor capacidad de predicción para responder la pregunta si una

persona compra o no. Este ejercicio se realizará de manera individual con los datos obtenidos de cada CC. Esto permitirá a Parque Arauco, en futuros análisis, diseñar cuestionarios específicos para cada ubicación, centrándose en las preguntas relacionadas con los aspectos que demuestren una mayor capacidad predictiva de manera individual.

11.4.1. Centro Comercial 1

Gráfico 1.
Curvas ROC CCI



Nota: Curvas ROC en donde se evidencia el tipo de variable que mejor capacidad predictiva tiene en el modelo de regresión logística aplicado para los datos disponibles en el Centro Comercial 1.

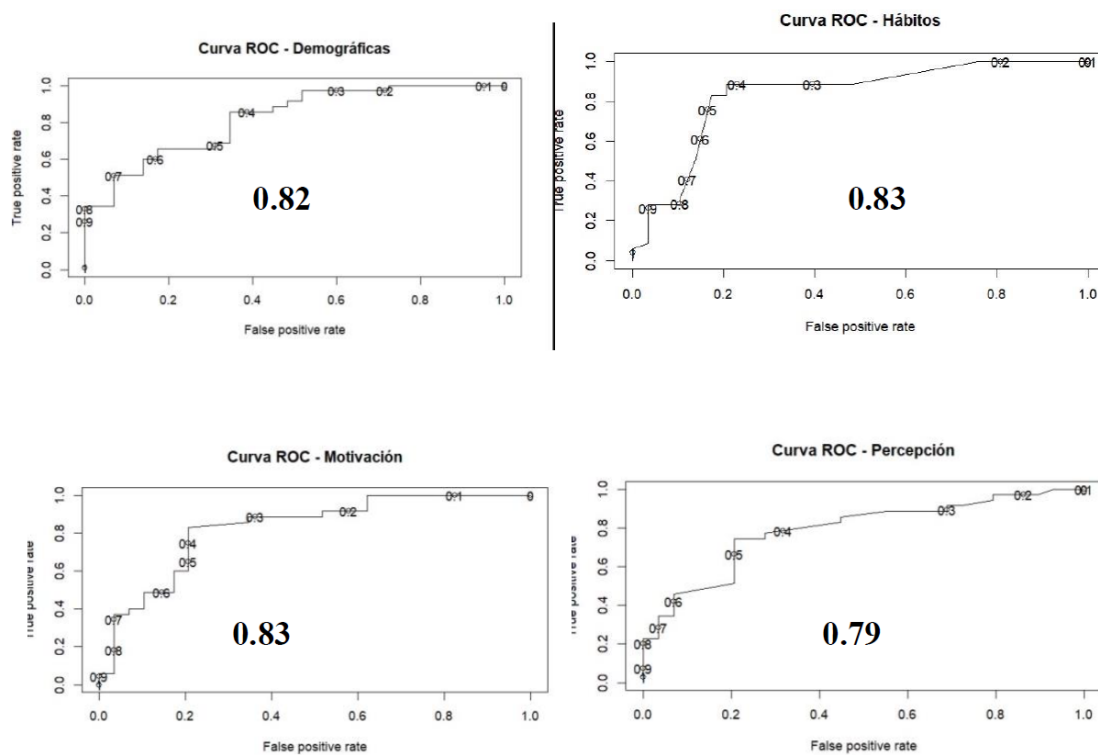
Fuente: Elaboración propia

Para el Centro Comercial 1 en Barranquilla, se observa que las variables más influyentes en el pronóstico de compra son las relacionadas con la motivación, específicamente aquellas vinculadas a las principales razones de visita y los sentimientos de los visitantes. En este contexto, es crucial que el centro comercial mantenga su carácter

multiplataforma, permitiendo que los visitantes realicen múltiples actividades diarias de manera eficiente, al tiempo que se garantiza su bienestar durante toda su experiencia. El incumplimiento de estas condiciones puede resultar en visitas sin gastos y en la posible pérdida de clientes potenciales.

11.4.2. Centro Comercial 2

Gráfico 2.
Curvas ROC CC2



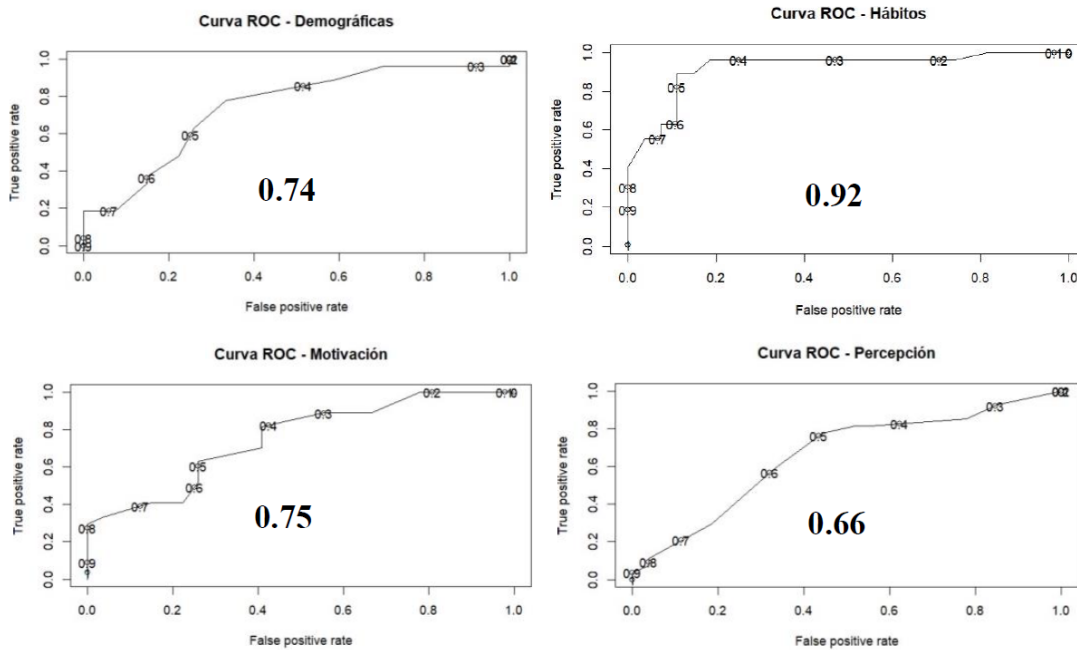
Nota: Curvas ROC en donde se evidencia el tipo de variable que mejor capacidad predictiva tiene en el modelo de regresión logística aplicado para los datos disponibles en el Centro Comercial 2.

Fuente: Elaboración propia

Para el Centro Comercial 2 ubicado en Medellín se identifica que, entre las variables consideradas, los factores predominantes son la motivación del consumidor y sus hábitos de compra. Esto sugiere que los visitantes de dicho centro comercial priorizan la conveniencia de encontrar una amplia gama de productos y servicios en un solo lugar, por encima de consideraciones como la proximidad y facilidad de acceso. En consecuencia, se infiere que la satisfacción de las necesidades y preferencias del consumidor, así como la diversidad y calidad de las ofertas comerciales, son elementos determinantes en la elección de realizar una transacción en este entorno comercial.

11.4.3. Centro Comercial 3

Gráfico 3.
Curvas ROC CC3



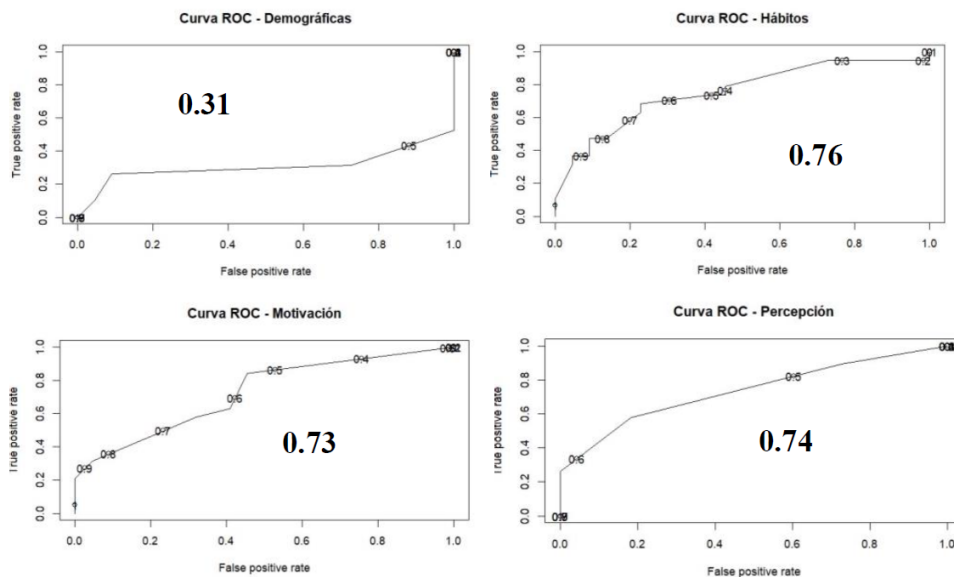
Nota: Curvas ROC en donde se evidencia el tipo de variable que mejor capacidad predictiva tiene en el modelo de regresión logística aplicado para los datos disponibles en el Centro Comercial 3.

Fuente: Elaboración propia

En el contexto del Centro Comercial 3, ubicado en Bucaramanga, se destaca la influencia significativa de las variables relacionadas con los hábitos de consumo en el ajuste óptimo para predecir la probabilidad de compra entre los visitantes. Este hallazgo subraya la importancia de comprender en profundidad cómo se comportan los consumidores en términos de sus hábitos de visita, incluyendo la frecuencia, la compañía durante la visita y el canal de compra utilizado. Estos aspectos son cruciales para el diseño de estrategias efectivas, como activaciones y eventos, que se alineen con la frecuencia de visita, el canal de compra preferido y los productos de interés, lo que puede resultar en un aumento del gasto promedio por transacción.

11.4.4. Centro Comercial 4

Gráfico 4.
Curvas ROC CC4



Nota: Curvas ROC en donde se evidencia el tipo de variable que mejor capacidad predictiva tiene en el modelo de regresión logística aplicado para los datos disponibles en el Centro Comercial 4.

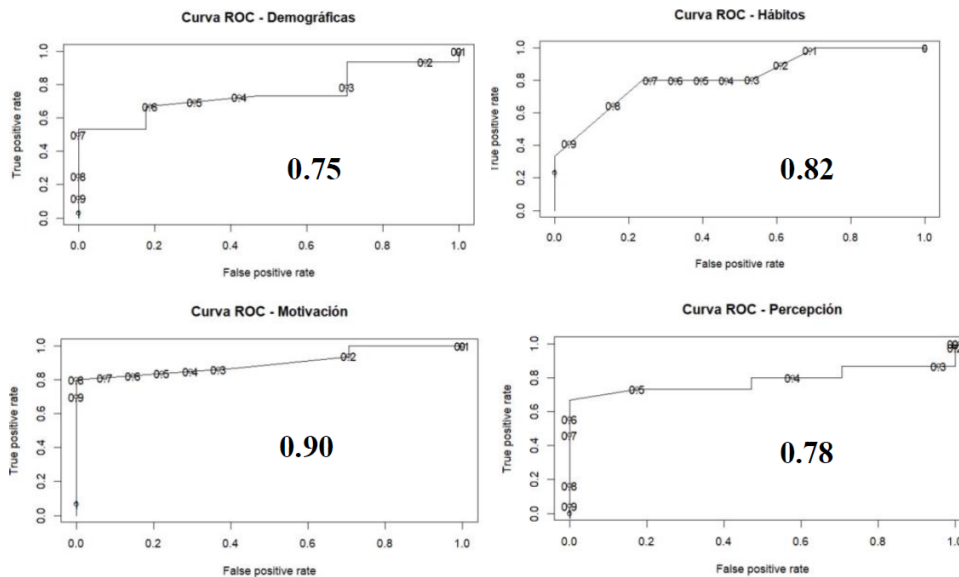
Fuente: Elaboración propia

En relación con el Centro Comercial 4 situado en Pereira, se observa una disminución significativa en la relevancia de las variables demográficas para la predicción de la compra. Esta disminución podría atribuirse a la alta densidad poblacional y a las cortas distancias características de esta ciudad, lo que sugiere que dichas variables no constituyen un factor determinante en el proceso de compra. Por el contrario, se evidencia que las variables relacionadas con los hábitos de consumo y la percepción del centro comercial son las que poseen una mayor influencia en la predicción de nuestra variable de interés. En consecuencia, se sugiere que el centro comercial oriente sus estrategias hacia el fortalecimiento de su imagen como destino preferido para realizar compras, entretenimiento y gestiones diversas. Para los visitantes de este establecimiento, resulta crucial la

percepción de que están accediendo a un espacio reconocido en el mercado por su versatilidad en términos de oferta disponible.

11.4.5. Centro Comercial 5

Gráfico 5.
Curvas ROC CC5



Nota: Curvas ROC en donde se evidencia el tipo de variable que mejor capacidad predictiva tiene en el modelo de regresión logística aplicado para los datos disponibles en el Centro Comercial 5.

Fuente: Elaboración propia

En el contexto del Centro Comercial 5, situado en Sopó y caracterizado como un centro outlet, se destaca que la motivación inicial de visita presenta una alta capacidad predictiva para la compra. Este fenómeno se atribuye al hecho de que la mayoría de los visitantes acuden a este tipo de establecimiento con la intención primaria de realizar compras, aprovechando los precios por debajo del mercado que ofrece. En este sentido, la influencia de factores como la composición familiar o los ingresos de los visitantes resulta poco significativa en la predicción de la compra. Por consiguiente, las estrategias dirigidas a este

caso particular deberían enfocarse en destacar la diversidad de opciones de entretenimiento y gastronomía, dado que la fase de establecimiento del primer contacto con el centro comercial está plenamente satisfecha. Se sugiere, por tanto, fomentar el gasto adicional en otras categorías de productos o servicios para maximizar el rendimiento económico del establecimiento.

12. Perfiles de Clientes Por Centro Comercial

12.1. Resultados Regresión Lineal

Teniendo en cuenta el ejercicio realizado en la sección 11.3, en donde se realizó un modelo para cada Centro Comercial con la finalidad de predecir de manera individual la probabilidad de compra de los clientes, a continuación, se muestran las variables significativas para cada uno de los modelos de cada CC. De manera subsecuente, se analiza el grupo de variables significativas resultantes de la aplicación de estos modelos para cada CC y se realiza un perfilamiento, de acuerdo con las mismas para cada uno de estos.

Tabla 11.

Variable Significativa de Perfil

No.	VARIABLES RELEVANTES	CC1	CC2	CC3	CC4	CC5
1	Género	Si				
2	Visita Solo el Centro Comercial	Si	Si			
3	Intención de Visita para Comer	Si				
4	Optimista Antes de Visita	Si		Si		
5	Motivado Antes de Visita	Si			Si	
6	Orgullosa Antes de Visita					Si
7	Compra en 1 o más locales	Si				

8	Noticias por Televisión				Si	
9	Empleado/Tiene Ingresos					Si
10	Perfil Individual	Si				Si
11	Perfil Social				Si	
12	Perfil Conservador		Si			
13	Perfil Digital	Si	Si	Si		

Nota: Nombres de las variables significativas según el modelo de regresión logística binaria aplicado, además

se clasifican las mismas según su significancia para cada Centro Comercial.

Fuente: Elaboración propia

Con el fin de tener un entendimiento de lo que representan las variables previamente presentadas, a continuación, se encuentran las definiciones de estas:

1. Género: Que sea hombre influye positivamente
1. Va solo/a al Centro Comercial: La mayoría de las ocasiones suele asistir al CC sin acompañantes.
2. Intención de Visita: La principal motivación por la que suele ir al CC no está relacionada con comida.
3. Optimista Antes de Visita: Se sentía optimista antes de acudir al Centro Comercial o no.
4. Motivado Antes de Visita: Se sentía motivado antes de acudir al Centro Comercial o no.
5. Orgulloso Antes de Visita: Se sentía orgullosos antes de acudir al Centro Comercial o no.
6. Compra en 1 o más locales: Cantidad de locales/objetos que terminara comprando, como calzado, accesorios, Cine, comida, entre otros.

7. Noticias por Televisión: Principal fuente de noticias del país/ciudad las obtiene a través de la televisión.
8. Empleado / Tiene Ingresos: Es una persona que recibe ingresos, sea un empleado, independiente, pensionado, entre otras.
9. Perfil Individual: En estos momentos de su vida el dinero no es la principal preocupación o enfoque.
10. Perfil Social: Es una persona que en busca activamente compartir con otras personas.
11. Perfil Conservador: Es una persona que no se considera conservadora y piensa que las redes sociales no son negativas para la sociedad.
12. Perfil Digital: Son personas que cuando deben realizar alguna PQR la prefieren hacer por algún medio digital, sin embargo, no se consideran una persona digital.

12.1.1. Perfiles por Centro Comercial

12.1.1.1. CC1 – Barranquilla

Para este CC, se quiere implementar una estrategia que permita no solo atraer más visitantes con los atributos del perfil identificado, sino también fidelizar y aumentar las compras de los clientes que ya tienen este perfil. En ese orden de ideas se deben tener en cuenta las siguientes variables significativas:

1. Que sea hombre.
2. Personas que buscan algo distinto a comida/almuerzo.
3. Se sienten desmotivados y negativos antes de ir al centro comercial.

4. El dinero en este momento no es su prioridad.
5. No se considera digital.

Adicionalmente, sabemos que las variables más significativas para el pronóstico de compra son las relacionadas con la motivación, específicamente aquellas vinculadas a las principales razones de visita y los sentimientos de los visitantes. En este orden de ideas, se sugiere implementar una experiencia sensorial, que inicie con un ambiente acompañado con música, aromas agradables y diferentes temáticas a lo largo del centro comercial. Es importante mencionar que, si no se potencia este grupo de trabajo motivacional, es muy probable que las visitas no hagan compras y se genere una fuga de clientes.

12.1.1.2. CC2 – Medellín

Para este CC, el perfil que se debe tener en cuenta al momento de idear la estrategia que se debe implementar con el fin de atraer a más clientes con el potencial de compra debe tener en cuenta las siguientes variables que resultaron significativas:

1. Personas que acuden acompañadas al CC.
2. No se consideran conservadoras.
3. No ven la tecnología como algo negativo para la sociedad, de manera que prefieren realizar cualquier PQR con una persona en vez de una red social, y se consideran digitales

Adicionalmente, para el Centro Comercial 2 ubicado en Medellín se identifica que, entre las variables consideradas, los factores predominantes son la motivación del consumidor y sus hábitos de compra. Esto nos anima a sugerir la organización de eventos y

actividades que generen interés y entusiasmo en los visitantes y durante el desarrollo de estas, se pueden promocionar ofertas y descuentos que promuevan la opción de compra de los visitantes.

12.1.1.3. CC3 – Bucaramanga

Para este CC, el perfil que se debe tener en cuenta al momento de idear la estrategia que se debe implementar con el fin de atraer a más clientes con el potencial de compra debe tener en cuenta las siguientes variables que resultaron significativas:

1. Sentimiento optimista antes de ir al CC.
2. No ven la tecnología como algo negativo para la sociedad, de manera que prefieren realizar cualquier PQR con una persona en vez de una red social, y se consideran digitales.

En el Centro Comercial 3, ubicado en Bucaramanga, se destaca la influencia significativa de las variables relacionadas con los hábitos de consumo para predecir la probabilidad de compra entre los visitantes. Este hallazgo subraya la importancia de desarrollar programas de fidelización que recompensen a los clientes frecuentes ya sea con beneficios exclusivos como lavado de carro gratuito o simplemente un descuento en el pago de parqueadero. Adicionalmente, se sugiere coordinar sinergias comerciales entre la administración del centro comercial y algunas tiendas o marcas para ofrecer promociones en productos especiales para los clientes habituales.

12.1.1.4. CC4 – Pereira

Para este CC, el perfil que se debe tener en cuenta al momento de idear la estrategia que se debe implementar con el fin de atraer a más clientes con el potencial de compra debe tener en cuenta las siguientes variables que resultaron significativas:

1. Sentimiento de motivado antes de acudir al CC.
2. Personas que ven noticias a través de la televisión.
3. Les gusta compartir y socializar con los demás en general, incluso cuando no se sienten bien.

La percepción del centro comercial es la variable que posee una mayor influencia en la predicción de compra de un visitante. En consecuencia, se sugiere que el centro comercial oriente sus estrategias hacia el fortalecimiento de su imagen como destino principal para realizar compras, actividades de esparcimiento y de entretenimiento. Una recomendación que sugerimos para cumplir con este objetivo es la de mejorar los canales de comunicación con los clientes, garantizando que estén enterados de las promociones y eventos que el centro comercial realiza para promover el bienestar de sus clientes. Esto podría ser mediante la implementación de una aplicación móvil o la notificación por mensajería sms.

12.1.1.5. CC5 – Sopó

Para este CC, el perfil que se debe tener en cuenta al momento de idear la estrategia que se debe implementar con el fin de atraer a más clientes con el potencial de compra debe tener en cuenta las siguientes variables que resultaron significativas;

1. No se sienten orgullosas antes de acudir al CC.

2. Cuentan con una seguridad financiera, es decir, tienen ingresos asegurados por diferentes conceptos, independientes, empleados, pensionados, entre otros.
3. En este momento de su vida la plata/dinero no es una prioridad o preocupación.

Para el contexto de este centro comercial, situado en Sopó y caracterizado por ser un centro outlet, se destaca que la motivación inicial de visita presenta una alta capacidad predictiva para la compra. Por consiguiente, las estrategias dirigidas a este caso particular deberían enfocarse en destacar la diversidad de opciones de entretenimiento y gastronomía, dado que la fase de establecimiento del primer contacto con el centro comercial está plenamente satisfecha. Se sugiere, por tanto, fomentar el gasto adicional en otras categorías de productos o servicios para maximizar el rendimiento económico del establecimiento.

12.2. Análisis de Correspondencias Múltiples

A continuación, se encuentran los resultados de un ejercicio de Análisis de Correspondencias Múltiples (ACM). La interpretación de los resultados de un análisis como este es crucial para comprender las relaciones entre las variables categorizadas y las categorías de respuesta. Una vez realizado el análisis, es fundamental examinar los mapas de puntos y las tablas de coordenadas para visualizar la estructura de las relaciones entre las categorías.

En el mapa de puntos, cada categoría se representa como un punto, y la proximidad de los puntos indica la fuerza de la asociación entre ellas. Las categorías que están más cerca unas de otras en el mapa de puntos tienen una mayor asociación, mientras que aquellas que están más separadas están menos relacionadas. Además, las dimensiones del

mapa de puntos muestran la estructura subyacente de los datos, lo que puede ayudar a identificar patrones y tendencias.

Las tablas de coordenadas proporcionan información detallada sobre la contribución de cada variable y categoría a las dimensiones del ACM. Esto permite identificar qué variables y categorías están más fuertemente asociadas con cada dimensión y cómo contribuyen a la estructura general del análisis.

El ACM que se realizó para cada CC fue con algunas de las variables demográficas previamente definidas en la sección 8 (se escogieron aquellas que permitan de mayor manera distinguir a los individuos), que son las siguientes:

Tabla 12.
Subgrupo Variables Demográficas

No.	Pregunta
E	¿Qué estrato aparece en la factura de energía de su hogar?
F	¿En cuál de los siguientes rangos de edad se encuentra usted hoy?
G	Género del entrevistado
25	¿Qué medio de transporte es el que usualmente utiliza para venir a este centro comercial?
26	¿Tiene hijos?
27	¿En su hogar tiene mascota? ¿Cuáles?
30	¿Cuál es el promedio de los ingresos mensuales de su hogar?
31	Cuál es su actividad actual principal
32	¿Qué nivel de estudios máximo ha completado usted?

Nota: Las preguntas mencionan características demográficas de cada visitante a los Centros Comerciales.

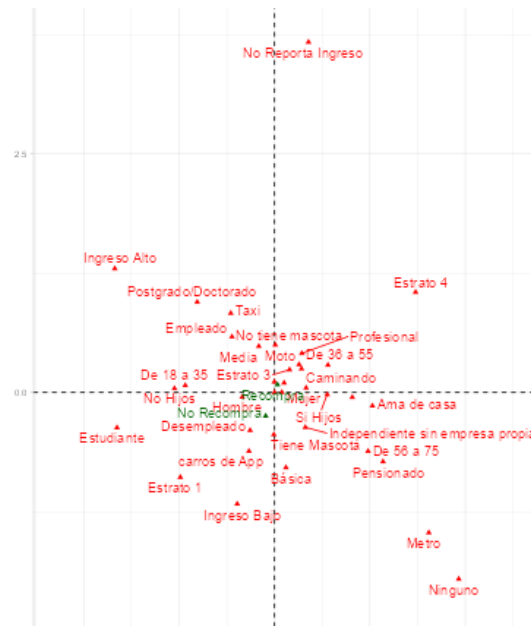
Fuente: Arquetipo encuesta, Parque Arauco, 2023.

Para cada uno de los gráficos que se van a presentar a continuación se hará un listado de aquellas variables que resultaron ser las más cercanas con el objetivo que el

cliente compre. Con este listado se realizarán recomendaciones al grupo Parque Arauco en la sección 13.

12.2.1. ACM Variables demográficas CC1

Figura 23.
ACM CC1



Nota: Gráfico de análisis de Correspondencias múltiples de variables demográficas y su cercanía con la variable objetivo de compra o no compra para el Centro Comercial 1.

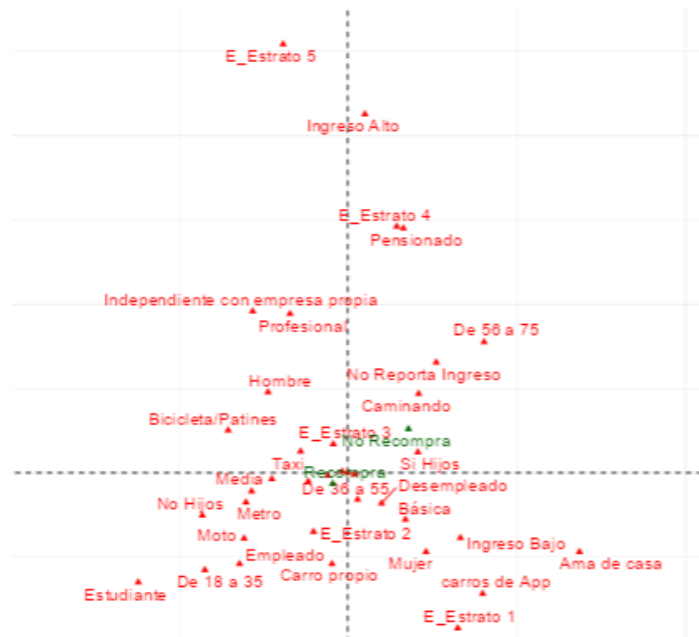
Fuente: Elaboración propia

Listado de Variables:

- Hombre o Mujer profesional,
- Edad: 18 y 35 años,
- Con hijos y mascotas,
- Estrato 2 y 3,
- Se acerca al Centro comercial caminando o por servicio público

12.2.2. ACM Variables demográficas CC2

Figura 24.
ACM CC2



Nota: Gráfico de análisis de Correspondencias múltiples de variables demográficas y su cercanía con la variable objetivo de compra o no compra para el Centro Comercial 2.

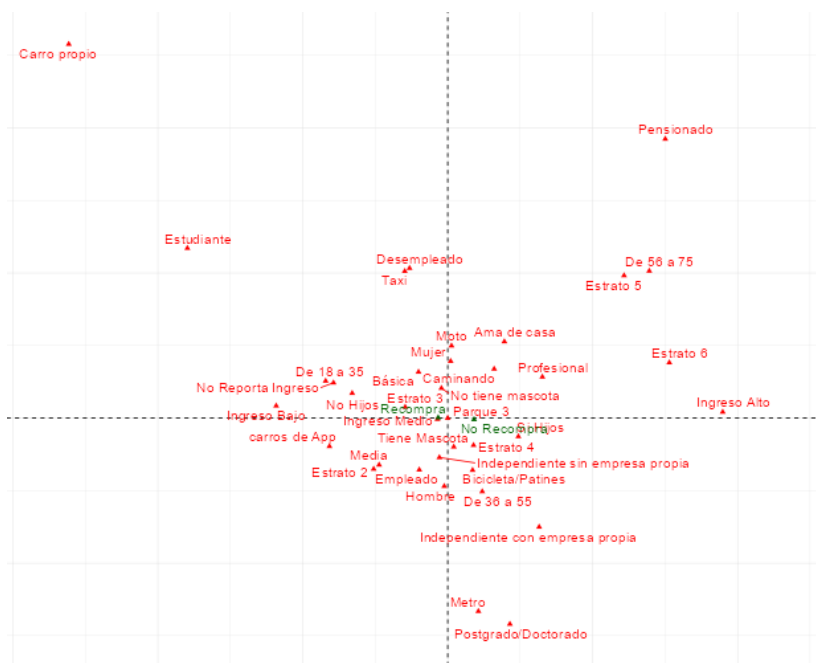
Fuente: Elaboración propia

Listado de Variables

- Hombre o mujer estrato 2,
- Edad: entre 36 y 55 años,
- Usa taxi o camina para ir al centro comercial,
- Educación media (Técnico/Tecnólogo)

12.2.3. ACM Variables demográficas CC3

Figura 25.
ACM CC3



Nota: Gráfico de análisis de Correspondencias múltiples de variables demográficas y su cercanía con la variable objetivo de compra o no compra para el Centro Comercial 3.

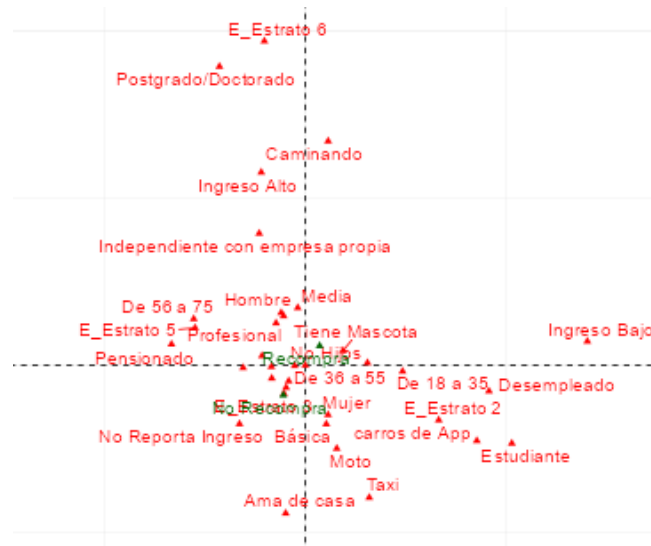
Fuente: Elaboración propia

Listado de Variables

- Hombre o mujer estrato 3 y 4,
- Edad: entre 36 y 55 años,
- Con mascotas y sin hijos,
- Tiene un rango de ingresos medio (de 1 a 4 millones de pesos),
- Usa carros de aplicaciones para ir al centro comercial,
- Educación media y profesional (Técnico/Tecnólogo).

12.2.4. ACM Variables demográficas CC4

Figura 26.
ACM CC4



Nota: Gráfico de análisis de Correspondencias múltiples de variables demográficas y su cercanía con la variable objetivo de compra o no compra para el Centro Comercial 4.

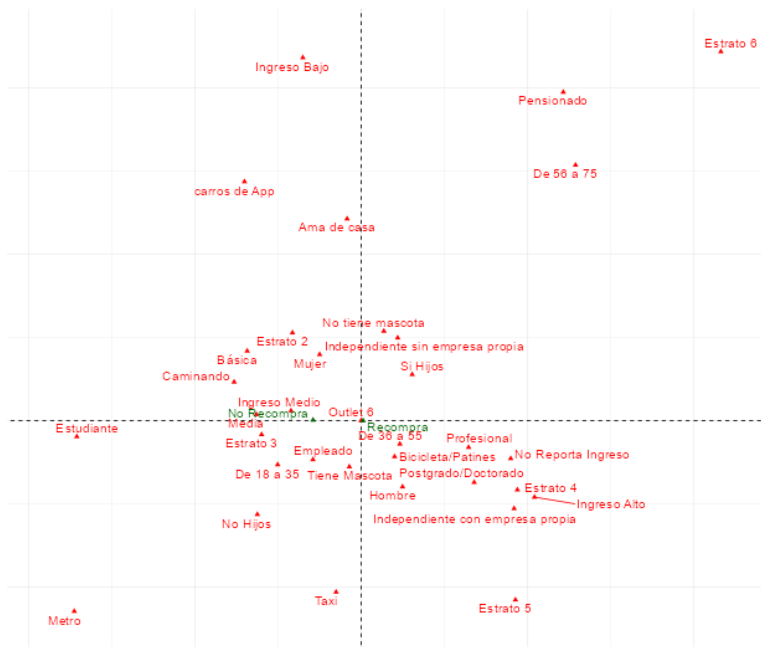
Fuente: Elaboración propia

Listado de Variables

- Hombre o mujer estrato 4 y 5,
- Edad: 36 a 55 años,
- Con mascotas y sin hijos,
- Independiente con rango de ingresos altos (Mas de 7 millones),
- Usa carros de aplicación para ir al centro comercial,
- Educación profesional y pensionados.

12.2.5. ACM Variables demográficas CC5

Figura 27.
ACM CC5



Nota: Gráfico de análisis de Correspondencias múltiples de variables demográficas y su cercanía con la variable objetivo de compra o no compra para el Centro Comercial 5.

Fuente: Elaboración propia

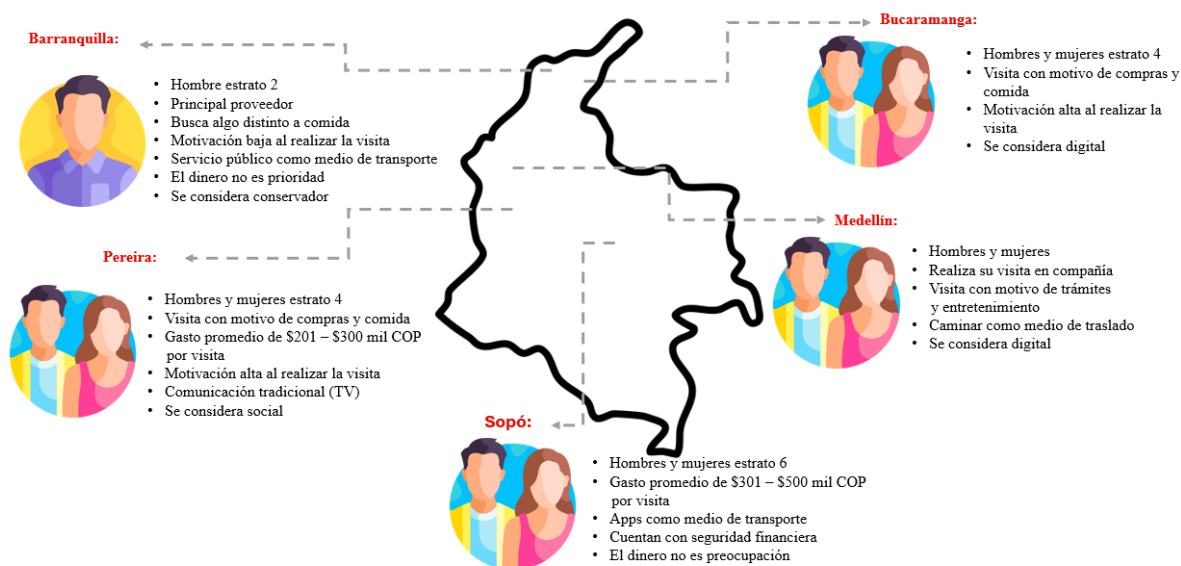
Listado de Variables

- Hombre o mujer estrato 4 y 5
- Edad: 36 a 55 años
- Con hijos
- Independiente con rango de ingresos altos (Mas de 7 millones)
- Usa medio de transporte propio para ir al centro comercial
- Educación profesional y postgrado

13. Recomendaciones para Atracción de Clientes

Teniendo como insumos los resultados de la sección 12.1.1, y 12.2.2, tenemos los resultados del comportamiento y aspectos demográficos de los clientes con mayor probabilidad de compra para cada uno de los CC. A continuación, se presenta un gráfico que agrupa los resultados del listado de variables para posteriormente utilizar esa información y establecer recomendaciones para cada uno de los CC.

Figura 28.
Perfil de Cliente por CC



Nota: Diagrama de perfil con mayor probabilidad de compra por cada Centro Comercial a nivel nacional

Fuente: Elaboración propia

En la Figura 28 está el resultado del perfilamiento para cada uno de los CC.

13.1. Barranquilla

La recomendación para este centro comercial (CC) está dirigida a hombres que buscan mejorar su estado de ánimo actual, sin que el factor económico sea una restricción. Se debe enfocar en las experiencias disponibles en el CC, especialmente en el ámbito del entretenimiento y las marcas asociadas. Es importante considerar que no buscan alimentarse, sino que desean mejorar su estado de ánimo con actividades de entretenimiento ofrecidas en el CC.

Dado que el CC está ubicado en una ciudad de clima cálido y cuenta con sistemas de aire acondicionado, sería ideal disponer de espacios interiores que fomenten el esparcimiento, así como áreas donde los visitantes puedan disfrutar de momentos de soledad, como mesas individuales o espacios de coworking en colaboración con empresas especializadas en este ámbito, como WeWork.

Además, para esta estrategia es necesario tener en cuenta las variables demográficas más relacionadas con el gasto, como se detalló previamente en el literal 12.2.1., es decir, personas de estratos 2 y 3, que se desplazan caminando o por medio de servicios públicos al CC (acá valdría la pena mirar las sinergias con los distintos medios de transporte para facilitar el acceso al CC), y tienen entre 18 y 35 años (vital para entendimiento de gustos de personas en este rango de edades).

13.2. Medellín

En este centro comercial (CC), la estrategia recomendada consiste en promocionarlo como un lugar destinado a reuniones familiares, ya sea con seres queridos, pareja o hijos, y desarrollar tiendas y experiencias que se distingan por su carácter no convencional.

Considerando que los potenciales compradores no son conservadores y tienen afinidad por la tecnología, podemos inferir que están interesados en estar a la vanguardia de los nuevos dispositivos y accesorios, así como en prendas de vestir exclusivas y poco convencionales.

Además, dado que la ciudad en la que se ubica el CC es la segunda más grande e importante del país, la exclusividad puede ser un elemento para considerar en las actividades ofrecidas. Por ejemplo, se podría organizar una exhibición de los últimos modelos de vehículos con la opción de realizar pruebas de manejo para aquellos que adquieran un pase especial.

Adicionalmente, para esta estrategia es crucial considerar las variables demográficas más vinculadas al gasto, tal como se delineó anteriormente en el apartado 12.2.2.

Específicamente, se refiere a individuos pertenecientes al estrato 2, quienes se trasladan al centro comercial (CC) a pie o en taxi, tienen edades comprendidas entre los 36 y 55 años, y poseen un nivel educativo de educación media (técnico/tecnólogo).

13.3. Bucaramanga

Como recomendaciones para este centro comercial (CC), se sugiere presentarlo como un destino ideal para aquellos momentos en los que se experimenta optimismo. Se propone destacar el CC como una recompensa para aquellos estados de ánimo positivos, asegurando que la experiencia en el lugar contribuirá a mantener esa sensación. Además, se recomienda la creación de áreas que fomenten la interacción social sin imponerla. Es decir, aunque no se espera que las personas entablen conversaciones con todos los presentes, se reconoce su aprecio por la presencia humana y el ambiente social. Se sugiere dedicar esfuerzos para

proporcionar espacios que permitan compartir con otros sin interacción directa, como áreas con mesas y sillas estratégicamente ubicadas dentro del CC.

Además, para esta estrategia es esencial considerar las variables demográficas más vinculadas al gasto, como se expuso previamente en la sección 11.2.3. En particular, se hace referencia a individuos pertenecientes a los estratos 3 y 4, los cuales se desplazan hacia el centro comercial utilizando transporte privado a través de aplicaciones móviles (con potencial para establecer acuerdos con las más utilizadas en la ciudad a fin de favorecer viajes al CC). Estos individuos tienen edades comprendidas entre los 36 y 55 años, cuentan con un nivel educativo de educación media (técnico/tecnólogo), y poseen ingresos que oscilan entre 1 y 4 millones de pesos (esto nos indica que hay la posibilidad de llevar a cabo un análisis de la oferta dentro del CC para atender a personas con este rango de ingresos, lo que podría generar oportunidades para establecer convenios con empresas financieras como Addi).

13.4. Pereira

Para atraer clientes a este centro comercial (CC), se sugiere dirigir los esfuerzos hacia individuos que están motivados y que han experimentado un día, una noticia o un momento positivo, y que buscan recompensarse, ya sea visitando CC o alguno de sus establecimientos (por ejemplo, una heladería). Dada la tendencia de la población a obtener información a través de la televisión, podemos inferir que la demografía que frecuenta este CC tiende a ser de edad más avanzada. Además, dado su interés por la socialización, se plantea la organización de actividades grupales que promuevan intereses poco convencionales y rompan con la rutina, como clases de arte acompañadas de degustaciones

de vino en el último piso del CC con vistas a la calle. Se sugiere que estas actividades fomenten la integración de las personas y generen la emoción adecuada que motive su participación.

Además, para esta estrategia es esencial considerar las variables demográficas más vinculadas al gasto, como se expuso previamente en la sección 11.2.4. En particular, se hace referencia a individuos pertenecientes a los estratos 3 y 4, los cuales se desplazan hacia el centro comercial (CC) utilizando transporte privado a través de aplicaciones móviles (con potencial para establecer acuerdos con las más utilizadas en la región). Estos individuos tienen edades comprendidas entre los 36 y 55 años, cuentan con un nivel educativo de educación media (técnico/tecnólogo), y poseen ingresos que oscilan entre 1 y 4 millones de pesos (con la posibilidad de llevar a cabo un análisis de la oferta dentro del CC para atender a personas con este rango de ingresos, lo que podría generar oportunidades para establecer convenios con empresas financieras como Addi).

13.5. Sopó

Considerando que los visitantes de este centro comercial (CC) no enfrentan preocupaciones financieras y buscan experimentar un sentimiento de orgullo, se puede inferir que están interesados en realizar compras que les brinden una sensación de satisfacción al adquirir artículos de lujo, especialmente dado el carácter de outlets de este CC. Aprovechando los amplios espacios disponibles en el CC, se sugiere impulsar actividades de alta gama que estén asociadas a la realización de una compra. Ejemplos de estas actividades podrían incluir encuentros con pilotos de carreras, competiciones de

drones o desfiles de moda protagonizados por algunas de las marcas representadas en el CC, entre otras posibilidades.

Además, para esta estrategia es esencial considerar las variables demográficas más vinculadas al gasto, como se expuso previamente en la sección 11.2.5. En particular, se hace referencia a individuos pertenecientes a los estratos 4 y 5, los cuales se desplazan hacia el centro comercial (CC) utilizando carro propio. Estos individuos tienen edades comprendidas entre los 36 y 55 años, cuentan con un nivel educativo de educación alta (posgrado), y poseen ingresos que superan los 7 millones de pesos en la mayoría por independientes (esto podría implicar que se pueden realizar actividades en horarios no convencionales, ya que estos individuos pueden tener mayor grado de manejo de tiempo).

14. Conclusiones

El trabajo realizado a lo largo de este documento consistió en identificar los comportamientos, motivaciones, y características de los clientes con mayor probabilidad de compra. Esto se hizo teniendo en cuenta la necesidad de Parque Arauco descritos en la introducción, mejorar su rentabilidad y experiencia del cliente, enfocados en los clientes que están generando valor para sus activos.

Después de analizar los cinco centros comerciales, se evidencia que no existe un cliente estándar que se pueda generalizar para todos los activos de Parque Arauco. Existen numerosos factores que contribuyen a esta variabilidad, siendo las diferencias geográficas y culturales las más influyentes. Sin embargo, esto no implica que no haya factores comunes entre los diferentes perfiles de clientes. De hecho, en la sección 11.1.1, se analizaron los

perfiles de cada uno de los centros comerciales, y se identificó un componente común: el aspecto emocional del cliente. Este aspecto emocional, que puede manifestarse como orgullo, motivación o una baja prioridad al dinero, es consistentemente relevante y debe ser considerado con mayor importancia al diseñar campañas de marketing o promoción de eventos.

Se recomienda, al igual que menciona Lehnert et al. (2020), ampliar aún más estas variables para lograr un mayor entendimiento del componente emocional de los clientes y con este mayor entendimiento tener una mejor comprensión de los mismos. Las variables discutidas en este documento, además de las relacionadas con la emoción, permitirán enfocar los esfuerzos en el tipo de cliente ideal para cada centro comercial. Por lo tanto, entre las acciones sugeridas para la empresa, se incluye el desarrollo de campañas emocionales dirigidas a los perfiles identificados en cada uno de los centros comerciales, tal y como se menciona en el capítulo 12.1.1. de este documento.

Dentro de la información proporcionada, se identifican numerosos estudios que podrían llevarse a cabo, el aquí mencionado y desarrollado es uno de los que se pueden desarrollar. Estos estudios podrían profundizar en el análisis del comportamiento de los clientes en cada uno de los Centros Comerciales, complementándose con información macroeconómica del momento en que fueron obtenidos. Los resultados aquí obtenidos, la implementación de este estudio por parte del Grupo Empresarial en sus diferentes CC, junto con un seguimiento adecuado de métricas, en especial en ventas, facilitaría la evaluación del valor de estas investigaciones.

Referencias bibliográficas

- Acuña, L. K. Q. (2023). Aplicación de Machine Learning a un modelo tradicional de Prevención y detección de fraude en entidad financiera proyectado a periodos trimestrales. *Universidad de la Salle*.
- Análisis de Regresión Múltiple*. (s. f.). Recuperado 15 de junio de 2024, de <https://rua.ua.es/dspace/bitstream/10045/8143/1/Regresion%20MUTIPLE.pdf>
- Analitik, V. (2022, mayo 26). Centros comerciales: Así cambiaron los hábitos de consumo. *Valora Analitik*. <https://www.valoraanalitik.com/2022/05/26/centros-comerciales-habitos-consumo-colombianos/>
- Argote, C. A. (2021, enero 29). *Parque Arauco se expandirá tras negocio con Credicorp Capital por \$588.000 millones*. Diario La República. <https://www.larepublica.co/empresas/parque-arauco-se-expandira-tras-negocio-con-credicorp-capital-por-588-000-millones-3117632>
- Bejarano, J. M. L. (2023, septiembre 7). *Parque Arauco suma compras por US\$51 millones en proyectos de Bogotá y Medellín*. Diario La República. <https://www.larepublica.co/empresas/parque-arauco-suma-compras-por-us-51-millones-en-proyectos-de-bogota-y-medellin-3698067>
- Carletta, J. (2012). *Squibs and Discussions: Assessing Agreement on Classification Tasks: The Kappa Statistic*. <https://web.archive.org/web/20120526213354/http://acl.ldc.upenn.edu/J/J96/J96-2004.pdf>
- Cerda, J., Vera, C., & Rada, G. (2013). *Odds ratio: Aspectos teóricos y prácticos*. 141.
- Cue Mancera, A. (2014). *Economía Internacional*. Grupo Editorial Patria.

Fayad, V. (2023, mayo 12). “*Tenemos alrededor de 1.050 millones de visitas al año a los centros comerciales*”. Diario La República.

<https://www.larepublica.co/especiales/vamos-al-centro-comercial/tenemos-alrededor-de-1-050-millones-de-visitas-al-ano-3610358>

Gomez Guasca, C. (2023, agosto 21). *La transformación del consumo, la cuota del retail y de las marcas* [Periódico]. Portafolio.co.

<https://www.portafolio.co/negocios/empresas/consumo-en-colombia-transformacion-cuota-del-retail-y-cambio-en-las-marcas-portafolio-30-anos-587646>

Greenacre, M. J. (with Internet Archive). (1984). *Theory and applications of correspondence analysis*. London ; Orlando : Academic Press.

<http://archive.org/details/theoryapplicatio0000gree>

Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013, marzo 22). *Applied Logistic Regression / Wiley Series in Probability and Statistics*.

<https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387>

Hu, B., Shao, J., & Palta, M. (2006). PSEUDO-R² in logistic regression model. *Statistica Sinica*, 16, 847-860.

Jiménez, M. V. G., Izquierdo, J. M. A., & Blanco, A. J. (2000). *La Predicción Del Rendimiento Académico: Regresión Lineal Versus Regresión Logística*. 248-252.

Junta Ordinaria de Accionistas 2024. (s. f.). Recuperado 14 de junio de 2024, de

https://parquearauco.modyocdn.com/uploads/d18c52d2-c514-4df2-bc9e-3f10683c1831/original/JOA_2024.pdf

Kliemann, L., & Sanders, P. (2016). *Algorithm Engineering: Selected Results and Surveys*. Springer.

Kmeans. (s. f.). Recuperado 11 de diciembre de 2023, de

https://www.unioviado.es/compnum/laboratorios_py/kmeans/kmeans.html

Mailund, T. (2017). *Beginning Data Science in R*. APRESS.

Migrant & Seasonal Head Start. (2006). *Introduction to Data Analysis Handbook*. AED.

Morales, D. A., Hernández, F. T., & Sierra, M. C. (2018). *Inteligencia de Negocios con Excel y Power BI* (1.^a ed.). Bii Soluciones & Excel Free Blog.

Olaya, S. (2023, febrero 27). <https://raddar.net/la-inflacion-el-gran-reto-para-las-marcas-y-los-consumidores-en-el-2023/>. <https://raddar.net/la-inflacion-el-gran-reto-para-las-marcas-y-los-consumidores-en-el-2023/>

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85-117. <https://doi.org/10.1016/j.neunet.2014.09.003>

Sheskin, D. J. (2020). *Handbook of Parametric and Nonparametric Statistical Procedures, Fifth Edition* (5.^a ed.). Chapman and Hall/CRC.
<https://doi.org/10.1201/9780429186196>

Vidal, R., Ma, Y., & Sastry, S. S. (2010). *Generalized Principal Component Analysis* (Vol. 40). Springer.

Anexo I – Encuesta Realizada

Sólo continúe si el entrevistado indica expresamente “SÍ”. Debe decir la palabra “SÍ”.

A. Marque el Centro Comercial de la encuesta:

- | | | |
|-------------|-------------|-------------|
| 1. Parque 1 | 3. Parque 3 | 5. Outlet 6 |
| 2. Parque 2 | 4. Parque 4 | |

B. ¡Bueno comencemos! ¿Por cuál de las siguientes razones vino a visitar el centro comercial hoy? **ENC: Mencione una a una las opciones**

1. A hacer un trámite o diligencia para alguien más (**Esta opción no aplica para Outlet**)
2. A trabajar o realizar un trabajo que el centro comercial o uno de las tiendas, almacenes o restaurantes le encomendó
3. A pasear, comer, comprar o hacer su propia diligencia

Algún otro? _____

ENC: Si menciona opciones 1 o 2, agradezca, termine Y marque en CNE. Preg B

C. ¿Diría que en este centro comercial usted ha permanecido el tiempo suficiente como para tener una buena impresión acerca de él, o todavía no?

1. Sí
2. No.....**ENC: Agradezca, Termine y marque en CNE C**

D. ¿Me puede confirmar que ningún familiar o acompañante suyo ha respondido esta misma encuesta hoy o durante esta semana, al menos que usted sepa?

1. Lo confirma
2. NO lo confirma..... **ENC: Agradezca y termine y marque en CNE D**

E. ¿Qué estrato aparece en la factura de energía de su hogar? (**ENC: Tenga en cuenta las cuotas. Estrato 1 sólo aplica para Parque Alegria y Fabricato. Si en otro C.C la persona se identifica como de estrato 1, termine y marque en CNE E).**

- | | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|

F. ¿En cuál de los siguientes rangos de edad se encuentra usted hoy? (**ENC: Mencione Las Opciones Una Por Una. Si es menor de 18 años o mayor de 75, agradezca y termine y marque en CNE F. Tenga en cuenta cuotas).**

- | | | |
|---------------|---------------|---------------|
| 1. De 18 a 25 | 3. De 36 a 45 | 5. De 56 a 65 |
| 2. De 26 a 35 | 4. De 46 a 55 | 6. De 66 a 75 |

G. Género del entrevistado (**ENC: Si tiene dudas, pregunte. Tenga en cuenta cuotas):**

1. Hombre
2. Mujer
- Otro: _____

*Quiero recordarle que esta es una entrevista para conocer mejor lo que nuestros visitantes piensan y quieren, así que no hay respuestas correctas, y tampoco piense que debería complacerme respondiendo algo en particular. Yo no pertenezco al centro comercial así que le agradezco que me hable con toda la honestidad del mundo. Gracias por recibirme aquí en (**ENC: Mencione la zona en la que se encuentra haciendo la encuesta -salida peatonal o de parqueadero- para que quede registrado en grabación según muestreo).***

1. Con base en su experiencia, ¿con qué probabilidad recomendaría a este centro comercial a un amigo, colega o familiar? Siendo cero “nada probable” y diez “muy probable”.

2. ¿Con qué frecuencia diría que visita este centro comercial? **(ENC: Lea una a una las opciones y marque sólo una)**

- | | | |
|-------------------------------|-------------------------|-----|
| 1. Dos o más veces por semana | 4. Una vez al mes | 7. |
| Primera vez que viene | | |
| 2. Una vez por semana | 5. Algunas veces al año | ... |
| 3. Entre 2 y 3 veces al mes | 6. Muy esporádicamente | |

3. ¿Tiene algún día de la semana en el que generalmente acostumbre a venir a este centro comercial? **(ENC: No lea las opciones. Puede marcar varias siempre que sea lógico).**

- | | | | | |
|--------------------|---------------|------------------|----------------|-----|
| 1. Entre semana | 3. Los lunes | 5. Los miércoles | 7. Los viernes | 9. |
| Los domingos | | | | |
| 2. Fines de semana | 4. Los martes | 6. Los jueves | 8. Los sábados | 10. |
| No tengo día | | | | |

5. Cuénteme si la mayoría de veces que visita este centro comercial viene acompañado o solo? ¿Con quién o quiénes viene la mayoría de veces? **(ENC: Puede marcar todas las que apliquen).**

- | | | |
|--|---------------------------------|------------|
| 1. Solo(a) | 4. Compañeros trabajo / estudio | 7. Con su |
| pareja | | |
| 2. Con amigo/a(s) | 5. Colega / compañero / cliente | 8. Mascota |
| 3. Familiar / familia (diferentes a pareja o hijo/nieto) | 6. Hijo(a) / nieto(a) | Otro: |

6. En promedio, las últimas veces que ha venido, ¿cuánto tiempo en total ha permanecido en el centro comercial?

Total Hras _____ Minutos _____

7. ¿Cuál diría que es la **principal** motivación por la cual generalmente visita este centro comercial? **(ENC: NO leer las opciones. Debe mencionar una sola)**

¿Cuál es la **segunda** motivación más importante generalmente? **(ENC: Debe mencionar una sola y Debe ser diferente a la principal motivación).**

¿Cuál es la **tercera** más importante? **(ENC: Debe mencionar una sola y Debe ser diferente a la principal y la segunda motivación).**

PRIMERA	SEGUNDA	TERCERA
1. A comer / almorzar	1. A comer / almorzar	1. A comer / almorzar
2. Comer helado	2. Comer helado	2. Comer helado
3. A comprar algo	3. A comprar algo	3. A comprar algo
4. A pasear / pasar el rato	4. A pasear / pasar el rato	4. A pasear / pasar el rato
5. A tomar café	5. A tomar café	5. A tomar café
6. A distraerse, cambiar de ambiente	6. A distraerse, cambiar de ambiente	6. A distraerse, cambiar de ambiente
7. Por servicios estéticos	7. Por servicios estéticos	7. Por servicios estéticos
8. Por servicios bancarios	8. Por servicios bancarios	8. Por servicios bancarios
9. Encontrarse con alguien	9. Encontrarse con alguien	9. Encontrarse con alguien
10. Entrar a cine	10. Entrar a cine	10. Entrar a cine
11. Estudiar / Trabajar	11. Estudiar / Trabajar	11. Estudiar / Trabajar
12. A relajarse / estar tranquilo(a)	12. A relajarse / estar tranquilo(a)	12. A relajarse / estar tranquilo(a)
13. A traer a hijo/niño a que se entretenga	13. A traer a hijo/niño a que se entretenga	13. A traer a hijo/niño a que se entretenga
14. Asistir a algún evento	14. Asistir a algún evento	14. Asistir a algún evento
15. Bolos	15. Bolos	15. Bolos
	16. Ningún otro	16. Ningún otro
Otro:	Otro:	Otro:

8. Perfecto. Ahora hablemos de HOY. Sólo piense en hoy, y específicamente en cómo se sentía justo antes de tomar la decisión de venir al centro comercial. **NO** cuando ya decidió venir, sino justo **ANTES** de **DECIDIR** venir. Piense en dónde estaba y qué estaba haciendo. Entonces le voy a mostrar estas opciones (**ENC: Presente tarjeta # 1**) y me va a decir cuál o cuáles son las dos o tres que mejor describen la forma como usted se sentía justo antes de decidir venir.

- | | | |
|----------------------|-----------------|-------------------|
| 1. Preocupado(a) | 9. Triste | 15. Satisfecho(a) |
| 2. Estresado(a) | 10. Rabia | 16. Descansado(a) |
| 3. Asustado(a) | 11. Ansioso(a) | 17. Libre |
| 4. Solitario(a) | 12. Aburrido(a) | 18. Enamorado(a) |
| 5. Confudido(a) | 13. Cansado(a) | 19. Optimista |
| 6. Inseguro(a) | | 20. Motivado(a) |
| 7. Desesperanzado(a) | 14. Normal | 21. Feliz |
| 8. Decepcionado(a) | | 22. Orgullosa(a) |

No está en la lista? Indique: _____

9. Pensando en su visita de hoy, diría que después de haber estado en este centro comercial, su estado de ánimo: (**ENC: Lea una a una y marque sólo una**).

- | | | |
|------------------------|-------------------------|-------------------------|
| 1. Ha mejorado mucho | 3. Ha permanecido igual | 4. Ha empeorado un poco |
| 2. Ha mejorado un poco | | 5. Ha empeorado mucho |

10. Mucha gente nos ha dicho que va a un centro comercial no sólo a comprar algo, a comer, o a hacer una vuelta, sino también por propósitos personales o emocionales. Así entonces, de las siguientes, cuáles son las dos o tres que mejor describen lo que usted la mayoría de veces viene a buscar al centro comercial: (**ENC: Presente tarjeta # 2 y pregunte por mínimo dos; acepte máximo 4; las más determinantes para el entrevistado**).

1. Pasar tiempo entretenido con mi familia / familiar
2. Pasar tiempo entretenido con amigo/a(s)
3. Pasar tiempo con mi pareja
4. Buscar un ambiente que me ayude a descansar
5. Buscar un ambiente que me ayude a desestresarme / calmarme
6. Buscar un ambiente que me despeje para pensar con mayor claridad
7. Buscar un ambiente donde ver gente y movimiento
8. Buscar un ambiente estimulante de emociones y experiencias diferentes
9. Me merecía un tiempo de esparcimiento o disfrute
10. Ninguna de las anteriores ya que vine a algo muy concreto 1
11. Ninguna de las anteriores ya que vine a algo muy concreto 2

11. ¿Qué diría que es lo que a usted más lo/la desmotiva de venir a este Centro Comercial? (**ENC: NO lea las opciones. Marque o escriba en Otros todas las que apliquen**) Indique por favor 2 o 3 cosas al menos. ¿Qué más? ¿Qué otra cosa? (**ENC: ¡Sondee sin mencionar opciones!**)

- | | |
|---|--|
| 1. Muy lleno de gente | 8. El desplazamiento hasta el centro comercial / lejos |
| 2. Las filas en restaurantes / tiendas | 9. La falta de tiempo |
| 3. Los precios altos | 10. Parquear / Pagar parqueo |
| 4. Que siempre es lo mismo / falta innovación | 11. Vías de acceso |
| 5. Que no encuentre lo que busco | 12. Falta de beneficios o incentivos al cliente |
| 6. Que toca caminar mucho | 13. Nada me desmotiva |
| 7. Mala asesoría / servicio | |

Otros: _____

12. Ahora indíqueme: de estos espacios del centro comercial que ve aquí en esta tarjeta, ¿cuáles son los que la mayoría de veces sí visita? (**ENC: Presente tarjeta # 3. Marque todos los que apliquen. Encuestador debe tener conocimiento de estas zonas**).

1. Plazoleta de comida
2. Espacios de Entretenimiento (cine, zona de niños, espacios de adultos, gimnasios, boleras)
3. Eventos
4. Social (restaurantes a mantel) **(Sólo en Alegra y Caracolí)**
5. Parqueaderos
6. Tienda(s) o establecimientos de servicio
7. Cafés - Cafeterías
8. Plaza del gusto **(Sólo en Fabricato)**
9. Calle del gusto **(Sólo en Fabricato)**
10. Parque de las Estrellas **(Sólo en Fabricato)**

13. ¿El día de hoy compró o gastará en algo dentro del centro comercial o no? ¿Dígame todo en lo que terminará gastando o comprando? **(ENC: Sondee ¡SIN LEER LAS OPCIONES! Acomode la respuesta a una de las opciones. Si definitivamente considera que no cabe en alguna, escríbala en OTROS).**

1. Vestuario	10. Productos de cuidado personal
2. Accesorios (bolsos, joyas y perfumes)	11. Tecnología / productos tecnológicos (incluso si son para el hogar)
3. Zapatos/calzado	12. Electrodomésticos
4. Restaurante (no cafés)	13. Juguetes / productos para niños
5. Helados / Dulces / Postres	14. Productos/Servicios de belleza
6. Atracciones para niños	15. Servicios bancarios y cambio de dólares
7. Cine	16. Otros servicios (viajes, operadores móviles, y otros)
8. Productos para mascotas	17. Café - cafetería
9. Productos o muebles para el hogar	18. Bolos
Otros:	

999. No compró / gastó en nada..... **ENC: Salte a pregunta 19**

14. ¿Dónde compró cada uno de esos productos o servicios? **(ENC: Leas las opciones una a una, y para cada una espere respuesta de si "sí" o "no", y marque las que si).**

- | | |
|---|--|
| 1. En uno de los módulos ubicados en los pasillos | 5. En una tienda / almacén |
| 2. En un restaurante de la plazoleta de comidas Fabricato) | 6. En la Calle del Gusto (Sólo Fabricato) |
| 3. En zona social (Sólo en Alegra y Caracolí Fabricato) | 7. En Plaza del gusto (Sólo Fabricato) |
| 4. En puestos de café o cafetería | |

15. ¿Quién o quiénes serán las personas que finalmente usarán o consumirán todos los productos o servicios que compró? **(ENC: ¡Sondee! Marque todas las que apliquen).** ¿Alguien más? ¿Alguna otra persona?

- | | | |
|-------------------------|-----------------------|-----------------------------------|
| 1. Ella / él de trabajo | 4. Padres / abuelos | 7. Oficina / empresa / compañeros |
| 2. Su pareja | 5. Familia en general | |
| 3. Hijo/Hija | 6. Amiga(o) | Otro: |

16. Le voy a leer unos rangos para que me indique cuánto sumará el gasto total que hará HOY en este centro comercial, entre todo lo que ha comprado y lo que piensa comprar:

1. Menos de 50 mil pesos
2. Entre 50 y 100 mil
3. Entre 101 mil y 200 mil
4. Entre 201 mil y 300 mil
5. Entre 301 mil y 500 mil
6. Entre 501 mil y 1 millón
7. Más de 1 millón
8. No da información (ENC: Evitar esta respuesta)

19. De las tres afirmaciones que le leeré, escoja la que usted honestamente cree que ocurrirá de ahora en adelante:

1. Voy a venir a este centro comercial con **menor** frecuencia a la acostumbrada
2. Vendré **más o menos con la misma** frecuencia a la acostumbrada
3. Vendré con **más** frecuencia a la acostumbrada

20. Quisiera que me diga si está de acuerdo o en desacuerdo con cada una de las siguientes afirmaciones que le voy a leer: **(ENC: Después de que mencione “de acuerdo” aclare si está “Totalmente de acuerdo” o “Algo de acuerdo; si menciona que está “en desacuerdo”, aclare si está “Algo en desacuerdo” o “Totalmente en desacuerdo”.**

	Totalmente de acuerdo	Algo de acuerdo	Algo en desacuerdo	Totalmente en desacuerdo
A. En esta etapa de mi vida estoy más enfocado(a) en producir ingresos que en mi bienestar físico o emocional	1	2	3	4
B. En vez de viajar prefiero comprar productos	1	2	3	4
C. Como te ves te juzgan	1	2	3	4
D. Dicen que la plata no es felicidad, pero la verdad es que sí es muy importante para ser feliz	1	2	3	4
E. Prefiero construirlo o arreglarlo, que comprarlo nuevo	1	2	3	4
F. Suelo antojarme de experiencias excitantes	1	2	3	4
G. Disfruto salir a distraerme o entretenerme yo solo(a), sin la compañía de otras personas	1	2	3	4
H. Cuando me siento bajo(a) de ánimo o cuando algo me intranquiliza, prefiero quedarme en casa en vez de salir	1	2	3	4
I. Compartir con mis amigo(a) es demasiado importante para mi felicidad y estabilidad emocional	1	2	3	4
J. Me considero más intelectual que social	1	2	3	4
K. Es sano evitar estar todo el tiempo con la familia	1	2	3	4
L. Me considero una persona introvertida	1	2	3	4
M. Me tienen sin cuidado la moda y las tendencias	1	2	3	4
N. Me gusta mucho experimentar cosas nuevas	1	2	3	4
O. Me considero de mentalidad conservadora	1	2	3	4
P. Mi perfil es más del lado humano y de las artes que del lado analítico y numérico	1	2	3	4
Q. Me gustaría vivir en otros países	1	2	3	4
R. Me gusta estar siempre en actividad, preferiblemente fuera de mi casa	1	2	3	4
S. Soy más práctico que filosófico	1	2	3	4
T. Las redes sociales, en general, son negativas para la sociedad	1	2	3	4
U. Me enfoco más en precio que en calidad o marca	1	2	3	4
V. En este momento de mi vida, la prioridad es mi familia	1	2	3	4
W. Casi siempre lo mejor es lo más caro	1	2	3	4
X. Cuando debo contactar al servicio al cliente de una empresa, prefiero que sea en persona o, en su defecto, por teléfono	1	2	3	4
Y. Me considero una persona digital	1	2	3	4

CONSUMO DE TECNOLOGÍA, REDES Y CANALES

21. ¿Cuál es el canal principal por dónde se entera usted de las noticias de su ciudad, el país o el mundo? **(ENC: No lea las opciones- Puede marcar hasta dos)**

- | | | |
|------------------------------|-----------------------|-------|
| 1. Prensa o medios impresos | 4. La televisión | Otro. |
| Cuál? _____ | | |
| 2. Prensa o medios digitales | 5. La radio | |
| 3. En redes sociales | 6. Por otras personas | |

22. ¿Consulta usted redes sociales al menos ocasionalmente?

1. Sí.....¿Cuáles redes sociales consulta regularmente? **(ENC: No lea las opciones- Puede marcar hasta dos)**

- | | | |
|--------------|-------------|-------------|
| 1. Facebook | 4. Youtube | |
| 2. Tik Tok | 5. LinkedIn | Otra: _____ |
| 3. Instagram | 6. Twitter | |

2. No

23. En qué barrio vive? _____

24. En un día de tráfico normal, y suponiendo que viniera en carro particular, cuántos minutos se demora de su casa a este centro comercial? **(ENC: Pida un promedio; no se deben poner rangos)**

_____ minutos

25. ¿Qué medio de transporte es el que usualmente utiliza para venir a este centro comercial?

- | | | |
|--------------|---------------------------------|----------------------------|
| 1. Caminando | 4. Taxi | 7. Patines / patineta |
| 9. Metro | | |
| 2. Bicicleta | 5. Uber/beat... (carros de App) | 8. Carro propio/particular |
| 3. Moto | 6. Bus /o tipo Transmilenio | Otro. |
| Cuál? _____ | | |

26. ¿Tiene hijos?

1. Sí..... **(ENC: Marque los que apliquen)**

1. Tiene menores de 10 años
2. Tiene entre 10 y 18 años
3. Tiene mayores de 18 años

2. No

27. ¿En su hogar tiene mascota? ¿Cuáles?

- | | | |
|----------|------------------------|---------------|
| 1. Perro | 3. Pez | 5. Pájaro/ave |
| 2. Gato | 4. Mini Pigs (cerdito) | 6. No tiene |
| _____ | | Otro: |

28. **(ENC: Sólo aplica para Parque Alegria):** Honestamente, usted diría que, en general, ¿prefiere comprar en tiendas en la calle o en tiendas que están dentro de un centro comercial?

1. En la calle

2. En centro comercial

.....¿Qué tipo de productos prefiere comprar en centros comerciales? **(ENC: Sondee)** _____

.....¿Y por qué esos productos prefiere comprarlos en centro comercial?

3. Depende.....¿De qué depende? **(ENC: Si depende de qué producto sea, indicar el producto)**

-
29. En total, ¿cuántas personas viven en su hogar incluyéndose a usted? _____....
30. Los ingresos mensuales totales de su hogar en promedio están en: **(ENC: Mencione TODAS Las Opciones. Una sola Respuesta).**
- | | | |
|---|-----------------------------------|---------|
| 1. Menos de un millón de pesos | 4. Entre siete y diez millones | 999. No |
| 2. Entre un millón y cuatro millones | 5. Entre diez y trece millones | |
| 3. Entre cuatro millones y siete millones | 6. Más de trece millones de pesos | |
- 31.Cuál de las siguientes corresponde a su actividad actual principal: **(ENC: Mencione Las Opciones. Acepte Una Sola Respuesta)**
- | | | |
|-------------------------------------|----------------|----------------|
| 1. Independiente con empresa propia | 4. Ama de casa | 7. Desempleado |
| 2. Independiente sin empresa propia | 5. Estudiante | |
| 3. Empleado | 6. Pensionado | Otro? |
-
32. ¿Qué nivel de estudios máximo ha completado usted? **(ENC: Mencione Las Opciones. Acepte Una Sola Resp.)**
- | | | |
|-------------|------------------------|------------------------|
| 1. Ninguno | 3. Bachillerato | 5. Universitario |
| 2. Primaria | 4. Técnico / tecnólogo | 6. Postgrado/Doctorado |
33. ¿Cómo es la composición del hogar donde vive? **(ENC: Mencione TODAS Las Opciones una a una)**
- | | |
|---------------------------------|--|
| 1. Padres con hijos | 5. Padres con hijos y otro(s) familiar(es) |
| 2. Pareja únicamente | 6. Más de una familia |
| 3. Madre o padre solo con hijos | 7. Vive solo |
| 4. Amigos | Otro. ¿Cuál? |
-
34. En el hogar en el que vive, ¿qué relación tiene usted con el jefe de hogar? **(ENC: Mencione TODAS Las Opciones una a una)**
- | | | |
|-------------------------------------|-------------------|----------|
| 1. Es el jefe de hogar | 3. Es hijo | 5. Amigo |
| 2. Es la/el esposa(o) o ama de casa | 4. Es un familiar | |
| Otro. _____ | | |

Anexo II – Regresión Logística por Centro Comercial

Modelo de Regresión Logística Centro Comercial 1

```

Confusion Matrix and Statistics

          Actual
Predicted 0  1
          0 21 12
          1  6 13

          Accuracy : 0.6538
          95% CI : (0.5091, 0.7803)
          No Information Rate : 0.5192
          P-Value [Acc > NIR] : 0.03477

          Kappa : 0.3004

          Mcnemar's Test P-Value : 0.23859

          Sensitivity : 0.7778
          Specificity : 0.5200
          Pos Pred Value : 0.6364
          Neg Pred Value : 0.6842
          Prevalence : 0.5192
          Detection Rate : 0.4038
          Detection Prevalence : 0.6346
          Balanced Accuracy : 0.6489

          'Positive' Class : 0

```

Matriz de Confusión: El modelo clasificó correctamente 21 muestras de la clase 0 como 0 y 13 muestras de la clase 1 como 1. Sin embargo, clasificó incorrectamente 12 muestras de la clase 0 como 1 y 6 muestras de la clase 1 como 0.

Accuracy (Exactitud): La exactitud del modelo es del 65.38%. Esto significa que aproximadamente dos de cada tres predicciones son correctas.

Intervalo de Confianza (CI): Con un 95% de confianza, la verdadera precisión del modelo probablemente esté entre el 50.91% y el 78.03%.

Kappa: El coeficiente Kappa es de 0.3004, lo que indica una concordancia moderada entre las predicciones del modelo y las clases reales.

Sensibilidad y Especificidad: La sensibilidad del modelo es del 77.78%, lo que significa que puede detectar correctamente el 77.78% de las muestras positivas. La especificidad del modelo es del 52%, lo que indica que puede identificar correctamente el 52% de las muestras negativas.

Valores Predictivos: El valor predictivo positivo es del 63.64%, lo que significa que el 63.64% de las predicciones positivas del modelo son correctas. El valor predictivo negativo es del 68.42%, lo que indica que el 68.42% de las predicciones negativas del modelo son correctas.

Prevalencia y Tasa de Detección: La prevalencia de la clase positiva en el conjunto de datos es del 51.92%. La tasa de detección del modelo es del 40.38%, lo que significa que puede detectar correctamente el 40.38% de todas las muestras positivas.

Balanced Accuracy: La precisión balanceada del modelo, que es la media aritmética de la sensibilidad y la especificidad, es del 64.89%.

El modelo específico para el Centro Comercial 1 arroja resultados que sugieren que tiene capacidad de clasificación, pero hay margen para mejorar, especialmente en cuanto a sensibilidad y especificidad para lograr una mejor identificación de ambas clases.

Modelo de Regresión Logística Centro Comercial 2

Confusion Matrix and Statistics

	Actual	
Predicted	0	1
0	23	5
1	7	22

Accuracy : 0.7895
 95% CI : (0.6611, 0.8862)
 No Information Rate : 0.5263
 P-Value [Acc > NIR] : 3.578e-05

 Kappa : 0.5793

 McNemar's Test P-Value : 0.7728

 Sensitivity : 0.7667
 Specificity : 0.8148
 Pos Pred Value : 0.8214
 Neg Pred Value : 0.7586
 Prevalence : 0.5263
 Detection Rate : 0.4035
 Detection Prevalence : 0.4912
 Balanced Accuracy : 0.7907

 'Positive' Class : 0

Matriz de Confusión: El modelo clasificó correctamente 23 muestras de la clase 0 como 0 y 22 muestras de la clase 1 como 1. Sin embargo, clasificó incorrectamente 5 muestras de la clase 0 como 1 y 7 muestras de la clase 1 como 0.

Accuracy (Exactitud): La exactitud del modelo es del 78.95%. Esto significa que aproximadamente ocho de cada diez predicciones son correctas.

Intervalo de Confianza (CI): Con un 95% de confianza, la verdadera precisión del modelo probablemente esté entre el 66.11% y el 88.62%.

Kappa: El coeficiente Kappa es de 0.5793, lo que indica una concordancia sustancial entre las predicciones del modelo y las clases reales.

Sensibilidad y Especificidad: La sensibilidad del modelo es del 76.67%, lo que significa que puede detectar correctamente el 76.67% de las muestras positivas. La especificidad del

modelo es del 81.48%, lo que indica que puede identificar correctamente el 81.48% de las muestras negativas.

Valores Predictivos: El valor predictivo positivo es del 82.14%, lo que significa que el 82.14% de las predicciones positivas del modelo son correctas. El valor predictivo negativo es del 75.86%, lo que indica que el 75.86% de las predicciones negativas del modelo son correctas.

Prevalencia y Tasa de Detección: La prevalencia de la clase positiva en el conjunto de datos es del 52.63%. La tasa de detección del modelo es del 40.35%, lo que significa que puede detectar correctamente el 40.35% de todas las muestras positivas.

Balanced Accuracy: La precisión balanceada del modelo es del 79.07%, que es bastante alta y muestra un buen equilibrio entre sensibilidad y especificidad.

En resumen, estos resultados indican que el modelo tiene un buen rendimiento, con una alta precisión y una capacidad sólida para detectar tanto las muestras positivas como las negativas.

Modelo de Regresión Logística Centro Comercial 3

```

Confusion Matrix and Statistics

          Actual
Predicted 0  1
          0 12  9
          1 13 38

          Accuracy : 0.6944
          95% CI : (0.5747, 0.7976)
          No Information Rate : 0.6528
          P-Value [Acc > NIR] : 0.2708

          Kappa : 0.2997

          Mcnemar's Test P-Value : 0.5224

          Sensitivity : 0.4800
          Specificity : 0.8085
          Pos Pred Value : 0.5714
          Neg Pred Value : 0.7451
          Prevalence : 0.3472
          Detection Rate : 0.1667
          Detection Prevalence : 0.2917
          Balanced Accuracy : 0.6443

          'Positive' Class : 0

```

Matriz de Confusión: El modelo clasificó correctamente 12 muestras de la clase 0 como 0 y 38 muestras de la clase 1 como 1. Sin embargo, clasificó incorrectamente 9 muestras de la clase 0 como 1 y 13 muestras de la clase 1 como 0.

Accuracy (Exactitud): La exactitud del modelo es del 69.44%. Esto significa que aproximadamente siete de cada diez predicciones son correctas.

Intervalo de Confianza (CI): Con un 95% de confianza, la verdadera precisión del modelo probablemente esté entre el 57.47% y el 79.76%.

Kappa: El coeficiente Kappa es de 0.2997, lo que indica una concordancia mínima entre las predicciones del modelo y las clases reales.

Sensibilidad y Especificidad: La sensibilidad del modelo es del 48.00%, lo que significa que puede detectar correctamente el 48.00% de las muestras positivas. La especificidad del modelo es del 80.85%, lo que indica que puede identificar correctamente el 80.85% de las muestras negativas.

Valores Predictivos: El valor predictivo positivo es del 57.14%, lo que significa que el 57.14% de las predicciones positivas del modelo son correctas. El valor predictivo negativo es del 74.51%, lo que indica que el 74.51% de las predicciones negativas del modelo son correctas.

Prevalencia y Tasa de Detección: La prevalencia de la clase positiva en el conjunto de datos es del 34.72%. La tasa de detección del modelo es del 16.67%, lo que significa que puede detectar correctamente el 16.67% de todas las muestras positivas.

Balanced Accuracy: La precisión balanceada del modelo es del 64.43%, lo que muestra un equilibrio moderado entre sensibilidad y especificidad.

En resumen, estos resultados indican que el modelo tiene un rendimiento moderado, con una precisión aceptable pero una sensibilidad relativamente baja. Hay margen para mejorar la capacidad del modelo para detectar correctamente las muestras positivas.

Modelo de Regresión Logística Centro Comercial 4

Confusion Matrix and Statistics

```

          Actual
Predicted 0  1
          0 11  8
          1  4 18

```

```

          Accuracy : 0.7073
          95% CI : (0.5446, 0.8387)
          No Information Rate : 0.6341
          P-value [Acc > NIR] : 0.2104

```

```

          Kappa : 0.4029

```

```

          McNemar's Test P-Value : 0.3865

```

```

          Sensitivity : 0.7333
          Specificity : 0.6923
          Pos Pred Value : 0.5789
          Neg Pred Value : 0.8182
          Prevalence : 0.3659
          Detection Rate : 0.2683
          Detection Prevalence : 0.4634
          Balanced Accuracy : 0.7128

```

```

          'Positive' Class : 0

```

Matriz de Confusión: El modelo clasificó correctamente 11 muestras de la clase 0 como 0 y 18 muestras de la clase 1 como 1. Sin embargo, clasificó incorrectamente 8 muestras de la clase 0 como 1 y 4 muestras de la clase 1 como 0.

Accuracy (Exactitud): La exactitud del modelo es del 70.73%. Esto significa que aproximadamente siete de cada diez predicciones son correctas.

Intervalo de Confianza (CI): Con un 95% de confianza, la verdadera precisión del modelo probablemente esté entre el 54.46% y el 83.87%.

Kappa: El coeficiente Kappa es de 0.4029, lo que indica una concordancia moderada entre las predicciones del modelo y las clases reales.

Sensibilidad y Especificidad: La sensibilidad del modelo es del 73.33%, lo que significa que puede detectar correctamente el 73.33% de las muestras positivas. La especificidad del modelo es del 69.23%, lo que indica que puede identificar correctamente el 69.23% de las muestras negativas.

Valores Predictivos: El valor predictivo positivo es del 57.89%, lo que significa que el 57.89% de las predicciones positivas del modelo son correctas. El valor predictivo negativo es del 81.82%, lo que indica que el 81.82% de las predicciones negativas del modelo son correctas.

Prevalencia y Tasa de Detección: La prevalencia de la clase positiva en el conjunto de datos es del 36.59%. La tasa de detección del modelo es del 26.83%, lo que significa que puede detectar correctamente el 26.83% de todas las muestras positivas.

Balanced Accuracy: La precisión balanceada del modelo es del 71.28%, lo que indica un buen equilibrio entre sensibilidad y especificidad.

En resumen, estos resultados sugieren que el modelo tiene un rendimiento moderado, con una precisión aceptable y un equilibrio razonable entre sensibilidad y especificidad.

Modelo de Regresión Logística Centro Comercial 5

```

Confusion Matrix and Statistics

          Actual
Predicted 0  1
          0 11  6
          1  2 12

          Accuracy : 0.7419
          95% CI : (0.5539, 0.8814)
          No Information Rate : 0.5806
          P-Value [Acc > NIR] : 0.04809

          Kappa : 0.4918

          Mcnemar's Test P-Value : 0.28884

          Sensitivity : 0.8462
          Specificity : 0.6667
          Pos Pred Value : 0.6471
          Neg Pred Value : 0.8571
          Prevalence : 0.4194
          Detection Rate : 0.3548
          Detection Prevalence : 0.5484
          Balanced Accuracy : 0.7564

          'Positive' Class : 0

```

Matriz de Confusión: El modelo clasificó correctamente 11 muestras de la clase 0 como 0 y 12 muestras de la clase 1 como 1. Sin embargo, clasificó incorrectamente 6 muestras de la clase 0 como 1 y 2 muestras de la clase 1 como 0.

Accuracy (Exactitud): La exactitud del modelo es del 74.19%. Esto significa que aproximadamente tres de cada cuatro predicciones son correctas.

Intervalo de Confianza (CI): Con un 95% de confianza, la verdadera precisión del modelo probablemente esté entre el 55.39% y el 88.14%.

Kappa: El coeficiente Kappa es de 0.4918, lo que indica una concordancia moderada a sustancial entre las predicciones del modelo y las clases reales.

Sensibilidad y Especificidad: La sensibilidad del modelo es del 84.62%, lo que significa que puede detectar correctamente el 84.62% de las muestras positivas. La especificidad del modelo es del 66.67%, lo que indica que puede identificar correctamente el 66.67% de las muestras negativas.

Valores Predictivos: El valor predictivo positivo es del 64.71%, lo que significa que el 64.71% de las predicciones positivas del modelo son correctas. El valor predictivo negativo es del 85.71%, lo que indica que el 85.71% de las predicciones negativas del modelo son correctas.

Prevalencia y Tasa de Detección: La prevalencia de la clase positiva en el conjunto de datos es del 41.94%. La tasa de detección del modelo es del 35.48%, lo que significa que puede detectar correctamente el 35.48% de todas las muestras positivas.

Balanced Accuracy: La precisión balanceada del modelo es del 75.64%, lo que indica un buen equilibrio entre sensibilidad y especificidad.

En resumen, estos resultados sugieren que el modelo tiene un rendimiento sólido, con una precisión razonable y una buena capacidad para detectar tanto las muestras positivas como las negativas.

