Integración de la Analítica de Datos POS y el Machine Learning para la Gestión Estratégica de Categorías en Comercial Nutresa S.A.S: Desarrollo de un Dashboard para Recomendación de Productos y Predicción de Ventas Utilizando Metodología CRISP-DM

Semestre de industria

Luis Fernando Mejía Torres

Trabajo de grado para optar por el título de Ingeniero Industrial

Asesor

Julián Andrés Castillo Grisales, Magíster en Ingeniería con énfasis en Simulación

Universidad de Antioquia
Facultad de ingeniería,
Ingeniería industrial
Medellín
2024

Cita	(Mejía, 2024)
Referencia **Estilo APA 7 (2020)**	Mejía Torres, L. (2024). *Integración de la Analítica de Datos POS y el Machine Learning para la Gestión Estratégica de Categorías en Comercial Nutresa S.A.S: Desarrollo de un Dashboard para Recomendación de Productos y Predicción de Ventas Utilizando Metodología CRISP-DM* [semestre de industria]. Universidad de Antioquia, Medellín.

Créditos a escenario de prácticas, personas, proyectos que aportaron al desarrollo de la práctica (interna y externamente: empresa y área de la empresa, grupo de investigación, proyecto, organización)

Dedicatoria

A mi madre y a mi padre, por siempre motivarme y apoyarme en todos mis procesos. A mis hermanos, por su confianza y apoyo. A mis amigos, por siempre estar en los momentos que los he necesitado.

Agradecimientos

A Comercial Nutresa y en especial al equipo de desarrollo de categorías por haberme permitido acompañarlos como practicante, por sus enseñanzas y acompañamiento durante el proceso. A mi asesor metodológico, por sus recomendaciones en el desarrollo de este proyecto.

Tabla de contenido

Lista de tablas

Lista de figuras

Siglas, acrónimos y abreviaturas

CM	Category management
EAN	Número de artículo europeo
CRISP-DM	Cross-Industry Standard Process for Data Mining
POS	Punto de venta
MAE	Error absolute medio
RMSE	Error cuadrático medio
ML	Machine Learning

Resumen

El objetivo de este proyecto fue construir un Dashboard que integre la analítica de datos POS con algoritmos de Machine Learning para mejorar la toma de decisiones empresariales en Comercial Nutresa S.A.S, una empresa líder en la distribución y venta de productos de consumo masivo. Utilizando la metodología CRISP-DM, se desarrollaron varios objetivos específicos: asegurar la calidad de los datos mediante técnicas de limpieza y preprocesamiento, realizar un análisis exploratorio de los datos POS, desarrollar un algoritmo de recomendación de productos y crear un modelo inicial para la predicción de ventas a nivel de categorías. La metodología incluyó fases de entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue, empleando herramientas como Python, Microsoft Excel y Power BI ®. Los resultados muestran avances significativos en la recomendación de productos, con un modelo que recomendó 837 productos en 19 categorías con alta precisión y evitando duplicaciones. Sin embargo, el modelo de predicción de ventas presentó resultados insatisfactorios, con altos márgenes de error en varias categorías, indicando la necesidad de mejoras. Las conclusiones destacan que la integración efectiva de analítica de datos POS y algoritmos de Machine Learning proporciona herramientas valiosas para la toma de decisiones empresariales. El análisis exploratorio y el modelo de recomendación demostraron ser eficientes y precisos, mientras que el modelo de predicción de ventas requiere ajustes para capturar mejor las tendencias y variabilidades en los datos de ventas históricos. Este proyecto evidencia el potencial de la analítica de negocios para generar valor y optimizar el modelo de negocio de la organización.

Palabras clave: analítica de negocios, análisis de datos, aprendizaje automático, administración de categorías, CRISP-DM.

Abstract

The objective of this project was to build a Dashboard that integrates POS data analytics with Machine Learning algorithms to improve business decision making in Comercial Nutresa S.A.S, a leading company in the distribution and sale of mass consumption products. Using the CRISP-DM methodology, several specific objectives were developed: ensure data quality through cleaning and preprocessing techniques, perform an exploratory analysis of POS data, develop a product recommendation algorithm and create an initial model for sales prediction at the category level. The methodology included phases of business understanding, data understanding, data preparation, modeling, evaluation and deployment, using tools such as Python, Microsoft Excel and Power BI ®. The results show significant advances in product recommendation, with a model that recommended 837 products in 19 categories with high accuracy and avoiding duplication. However, the sales prediction model presented unsatisfactory results, with high error margins in several categories, indicating the need for improvement. The findings highlight that effective integration of POS data analytics and machine learning algorithms provides valuable tools for business decision making. The exploratory analysis and recommendation model proved to be efficient and accurate, while the sales prediction model requires adjustments to better capture trends and variabilities in historical sales data. This project evidences the potential of business analytics to generate value and optimize the organization's business model.

Keywords: business analytics, data analysis, machine learning, category management, CRISP-DM.

Introducción

Este documento presenta un proyecto desarrollado durante la práctica académica para Comercial Nutresa S.A.S, una empresa líder en la distribución y venta de productos de consumo masivo en Colombia, Centroamérica y el Caribe. Comercial Nutresa cuenta con un amplio portafolio de marcas reconocidas en las categorías de alimentos, bebidas, confitería y cuidado personal. Aunque la empresa ya ha implementado una integración entre el análisis de datos POS (Point of Sale) y el Category Management (CM), existe una oportunidad para ampliar esta integración y obtener aún más beneficios en los procesos de gestión de categorías.

El objetivo principal de este proyecto era la construcción de un Dashboard que integre la analítica de datos POS con algoritmos de Machine Learning para mejorar la toma de decisiones empresariales. Para lograrlo, se desarrollaron varios objetivos específicos, incluyendo la limpieza y preprocesamiento de datos, el análisis exploratorio de datos POS, el desarrollo de un algoritmo de recomendación de productos y la creación de un modelo inicial para la predicción de ventas a nivel de categorías.

La metodología seguida se basa en la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), que incluye fases de entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue. Adicionalmente, se utilizaron las herramientas Microsoft Excel y Power BI ® y el lenguaje de programación Python, para el análisis y visualización de datos. Los resultados obtenidos muestran un alto rendimiento del modelo de recomendación de productos, aunque se identificaron áreas de mejora en el modelo de predicción de ventas.

1 Objetivos

1.1 Objetivo general

Construir un Dashboard que integre la analítica de datos POS con algoritmos de machine learning para mejorar la toma de decisiones empresariales

1.2 Objetivos específicos

- Aplicar la metodología CRISP-DM en todas las fases del proyecto para asegurar un enfoque estructurado en la integración de la analítica de datos POS y el Machine Learning.
- Desarrollar un algoritmo que permita asegurar la calidad y coherencia de los datos mediante técnicas de limpieza y preprocesamiento.
- Construir un análisis exploratorio que permita identificar y entender las características principales de los datos POS.
- Desarrollar un algoritmo de recomendación de productos clave para cada cliente
- Desarrollar un modelo base para la predicción de ventas a nivel de categorías

2 Marco teórico

En la actualidad, la aplicación de técnicas de análisis de datos en el contexto de los negocios es un proceso que puede generar ventajas competitivas a las organizaciones; como lo menciona (Aguilar, 2019) la Analítica de Negocios (Business Analytics) es un concepto creado por proveedores y consultores de TI que se enfoca en herramientas y técnicas para analizar y entender datos. Estas herramientas incluyen desde el procesamiento analítico en línea (OLAP de sus siglas en inglés Online Analytical Processing), estadísticas, y modelos de datos, hasta tecnologías avanzadas de inteligencia artificial, como el aprendizaje automático y el aprendizaje profundo. El aprendizaje automático (ML) como lo definen (Dzyabura & Yoganarasimhan, 2018) se refiere al estudio de métodos o algoritmos diseñados para aprender los patrones subyacentes en los datos y hacer predicciones basadas en estos patrones.

Con el paso de los años el ML ha jugado un papel fundamental en las industrias que implementan procesos de marketing, ya que posibilita la extracción de grandes conjuntos de datos, brindando a los especialistas en marketing la oportunidad de obtener nuevas visuales sobre el comportamiento del consumidor y optimizar el desempeño de las operaciones de marketing (Cui et al., 2006).

Por su parte, la Gestión de Categorías (Category Management por sus siglas en inglés CM) se define como un enfoque estratégico para el marketing minorista que considera un grupo de productos afines como una única unidad de negocio (categoría). Si bien la CM ha sido un componente crucial para el éxito del marketing minorista desde finales de la década de 1980, el panorama actual exige una evolución en sus prácticas; La integración de Big Data en la CM presenta una gran oportunidad para los minoristas, como lo destaca (Dekimpe, 2020): "El comercio minorista está en el centro de una tormenta de oportunidades y desafíos de Big Data". Esto denota la importancia en la utilización de la analítica de negocios de forma que permita

elevar la participación de las categorías en los diferentes formatos de clientes y estimar cómo sería su comportamiento con el transcurso del tiempo.

3 Metodología

Este proyecto basado en ciencia de datos presenta un enfoque cuantitativo siguiendo la metodología CRISP-DM en el cual se utilizan técnicas de procesamiento de datos, modelación estadística, algoritmos de aprendizaje automático y herramientas de visualización para facilitar el análisis de los resultados obtenidos; todo esto llevado a cabo mediante la integración del lenguaje de programación Python y herramientas como Microsoft Excel y Power BI ®.

En base a (Wirth & Hipp, 2000) la metodología CRISP-DM llevada a cabo en este proyecto abarca las siguientes fases:

1. Entendimiento del negocio: En esta primera fase se realizó la definición y entendimiento de los objetivos del proyecto desde la perspectiva empresarial, se verificaron los requerimientos frente a la información y recursos necesarios para llevarlo como un problema de ciencia datos.
2. Entendimiento de los datos: Siguiendo los pasos y definición de (Meeting & Chapman, s/f), en esta fase se realiza la recopilación de los datos y se realizan actividades que permitan tener un mayor conocimiento de ellos tales como un análisis exploratorio y descriptivo de los mismos.

   De este modo, se recopilaron los datos desde enero 2020 hasta abril de 2024 que posee la compañía en plataformas como SAP BO (SAP BusinessObjects de sus siglas en inglés) y se almacenaron en copias locales en formato CSV(Comma separated values de sus siglas en inglés); estos datos corresponden a las ventas POS, formado un total de 36.970.650 datos, con 1.478.826 de registros (ventas) y 25 variables; luego de esto, se realizó un análisis exploratorio de los datos, haciendo uso del lenguaje de programación Python para conocer el tipo de datos y variables, las tendencias de venta a nivel de clientes con variables como canal, formatos, segmento de cadena, y a nivel de producto descrito por categoría, subcategoría, marca y fabricante.

3. Preparación de los datos: Para (Wirth & Hipp, 2000) esta etapa se basa en todas las actividades que permiten construir un modelo de datos final partiendo de los datos sin procesar que se tienen inicialmente. Es probable que las tareas de preparación de datos se realicen varias veces y no en ningún orden prescrito. Las tareas incluyen selección de tablas, registros y atributos, limpieza de datos, construcción de nuevos atributos y transformación de datos para herramientas de modelado (Wirth & Hipp, 2000).

   De acuerdo con lo anterior, para este caso se inició con la transformación de todas las variables y su contenido a minúsculas para evitar errores al nombrarlas, seguido de un tratamiento de los datos nulos iniciando por variables categóricas como: segmento, subsegmento, presentación, tamaño, macromundo y misión compra, en las cuales los valores nulos fueron reemplazados por ND (no data/sin datos), para la variable tipo oferta, se reemplazaron los valores nulos por sin oferta; paso seguido, se continúa con el tratamiento de nulos en las variables numéricas ventas en pesos y ventas en unidades, para este caso los datos nulos fueron eliminados ya que los productos sin ventas son porque se dejaron de vender o cambiaron de código, estos corresponden a un total de 1.069 registros. A continuación, teniendo los datos limpios se realizó la creación de una nueva variable clave para el estudio, la cual fue nombrada propio/competencia y corresponde al resultado de determinar qué fabricantes pertenecen a Nutresa, cuales pertenecen a marcas propios y cuales son solo marcas de competencia; finalmente, los datos que se utilizaron corresponden a 1.477.757 registros y 26 variables, las cuales son descritas en la **Tabla 1**.

Tabla 1
Descripción de los datos.

Variable	Descripción
Fecha	Fecha del registro
Canal	Canal al cual pertenece el cliente
Segmento cadena	Segmento al cual pertenece el cliente.
Región	Región donde se realiza la venta
Formatos	Nombres de los clientes
EAN	Código EAN del producto
Descripción producto	Nombre del producto

Tipo producto	Describe si el producto es regular u oferta
Tipo oferta	Tipo de oferta del producto
Macrocategoría	Categoría principal/global del producto
Categoría	Categoría del producto dentro de la Macrocategoría
Subcategoría	Subcategoría a la cual pertenece el producto
Segmento	Segmento del producto
Subsegmento	Subsegmento del producto
Fabricante	Fabricante del producto
Marca	Marca del producto
Submarca	Submarca del producto
Presentación	Presentación en la cual viene el producto
Peso	Peso del producto
Tamaño	Tamaño en el cual viene el producto
Macromundo	Corresponde a clasificación del producto
Misión compra	Motivo de consumo del producto
Propio/competencia	Diferencia los productos de fabricación propia y los de la competencia
Numérica	Numero de formatos en los que el producto está presente
Ventas_pesos	Ventas en pesos del producto
Ventas_unidades	Ventas en unidades del producto

Fuente. Elaboración propia.

4. Modelamiento: Según (Wirth & Hipp, 2000) En esta fase, se seleccionan y aplican diversas técnicas de modelado, y se calibran sus parámetros a valores óptimos. Típicamente, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas requieren formatos específicos de datos y es por esto por lo que en ocasiones se debe volver a la fase de preparación de los datos antes de aplicar algunas técnicas.

Para los objetivos de este proyecto, se realizaron dos algoritmos con tres fuentes de información diferentes, dos obtenidas desde la base de datos original y otra base adicional que contenía la validación de los códigos EAN con los códigos SAP de los

productos Nutresa. El primer modelo que se creó constaba de un algoritmo que permitiera la recomendación de productos que tuvieran en los últimos tres meses una venta promedio en pesos y unidades igual o superior al tercer cuartil de los datos; para este modelo se utilizó una fuente de datos extraída de la fuente original con las variables canal, formato, categoría, ean, propio o competencia, ventas en pesos y ventas en unidades; adicionalmente, se utilizó para este modelo una fuente de datos que contenía el código SAP correspondiente a los códigos EAN de los productos de la compañía, esto debido a que pueden haber productos que aunque son el mismo vienen en presentación diferente (individual, caja, paquete) y por ende tienen código EAN diferente mientras que para la compañía son el mismo y por eso los identifica con código SAP (no contempla las unidades individuales), de esta forma se evitaría recomendar una porción individual y solo recomendar productos que realmente el cliente no tenga, y con la variable propio/competencia se ajusta el modelo para que no recomiende productos considerados como marca propia ya que son marcas que se consideran exclusivas de los clientes.

El segundo modelo desarrollado fue una predicción para el cual se utilizó una fuente de datos tomada de la original con las variables fecha, categoría, ventas en pesos y ventas en unidades, en este modelo se utilizaron técnicas estadísticas y de ML para buscar un primer acercamiento a lo que sería un modelo predictivo de ventas a nivel de categoría que se planea desarrollar en el área de desarrollo de categorías, y es por esto que al tratarse de un primer acercamiento no se hace mucho énfasis en múltiples técnicas de modelo de series de tiempo debido a que esto sería un trabajo de varias fases las cuales tendrían una extensión y un alcance mayor al de la práctica académica. Para este modelo, se utilizó el lenguaje de programación Python y se utilizan librerías especializadas en pronósticos como sklearn (Pedregosa et al., 2011), skforecast y funciones de modelado y ajuste como ForecasterAutoreg[1] y RandomForestRegressor[2]; el modelo se realizó con un

[1] Es una biblioteca de Python que facilita el uso de regresores de scikit-learn como pronosticadores de un solo paso y de múltiples pasos.

[2] Es un metaestimador que se ajusta a una serie de árboles de decisión regresores en varias submuestras del conjunto de datos y utiliza el promedio para mejorar la precisión predictiva y controlar el sobreajuste

autoajuste de hiperparámetros para un *n_estimators = 100*, un *random_state = 42* y *12 lags* y se iteró sobre cada categoría para realizar los pronósticos en conjunto.

5. Evaluación. En esta fase como lo define (Wirth & Hipp, 2000) ya se ha construido uno o más modelos que parecen tener alta calidad, desde una perspectiva de análisis de datos. Sin embargo, antes de proceder con la implementación final del modelo, es importante evaluar más a fondo el modelo y revisar los pasos ejecutados para construirlo, para asegurarse de que logra correctamente los objetivos comerciales.

   Buscando aplicar lo anterior, se definieron diferentes formas de evaluar los resultados de cada modelo construido; para el caso del modelo de recomendación de productos se evalúa el modelo mediante la verificación de que los productos recomendados no hagan parte del cliente al cual se le están recomendando y que no se recomiende el mismo producto con presentaciones diferentes. Para el modelo de forecasting se toman solo las métricas de error absoluto medio (MAE Mean Absolute Error de sus siglas en inglés) y el error cuadrático medio (RMSE Root Mean Squared Error de sus siglas en inglés) los cuales permiten conocer en este caso a cuanto a se encuentra la venta pronosticada de la real.

6. Despliegue: En esta fase final, tal y como lo de denota (Wirth & Hipp, 2000) la creación del modelo generalmente no es el final del proyecto. Por lo general, los conocimientos adquiridos deberán ser organizados y presentados de manera que el cliente pueda utilizarlos. Dependiendo de los requisitos, la fase de implementación puede ser tan simple como generar un informe o tan compleja como implementar un proceso repetible de minería de datos. Para continuar aplicando la metodología en este proyecto, los resultados de la exploración de los datos y del modelo de recomendación de productos se exportaron a formato *CSV* para ser llevados finalmente a un tablero de Power BI ® en el cual las personas las personas del área pueden realizar un seguimiento constante de la información; para la creación del tablero se inició se tuvo en cuenta la información que era solicitada por área como por ejemplo: un resumen de los datos, la tendencia de ventas del top cinco de marcas y productos, la tendencia general de las ventas y por último una tabla en la cual se pudieran visualizar todos los productos recomendados; los resultados

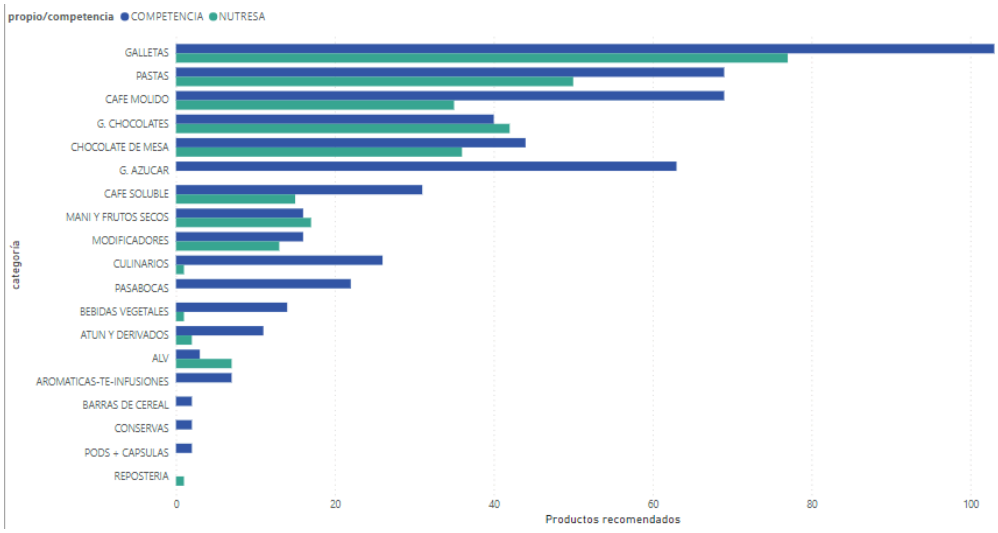
del modelo de forecasting fueron presentados como informe en un comité cerrado ya que es un proceso que se planea seguir mejorando para poder hacer un despliegue completo de las predicciones.

4 Resultados

Para el modelo de recomendación de productos se encontraron en total 4.267 registros, correspondientes a 837 productos recomendados en 19 categorías, de las cuales las categorías con mayor número de productos recomendados son galletas, pastas y café molido (**Figura 1**.); de los productos recomendados el 35,48% pertenecen a Nutresa, mientras que el 64,52% pertenecen a la competencia (**Figura 2**.)

Figura 1

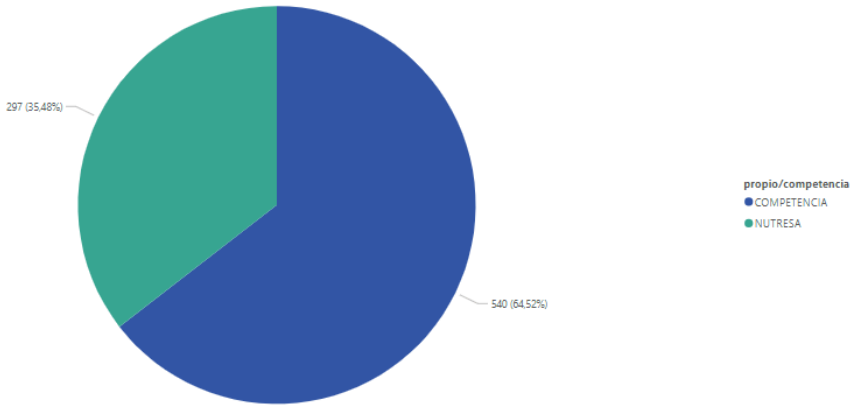
Productos recomendados por categoría y fabricante.



Fuente. Elaboración propia.

Figura 2

Porcentaje de productos recomendados por fabricante.



Fuente. Elaboración propia.

De los productos Nutresa recomendados (297) solo 30 correspondientes aproximadamente al 10% fueron recomendados en una presentación diferente a la que ya existe en el formato, esto quiere decir que el margen de error del modelo es bajo teniendo en cuenta que a ningún formato se le recomendó un producto igual al que ya tiene en su portafolio.

En el caso del modelo de forecasting se encontraron resultados poco satisfactorios debido a que, en diferentes categorías pronosticadas, el error es bastante alto lo cual denota un bajo aprendizaje del modelo frente a los datos que se le entregaron (**Tabla 2**).

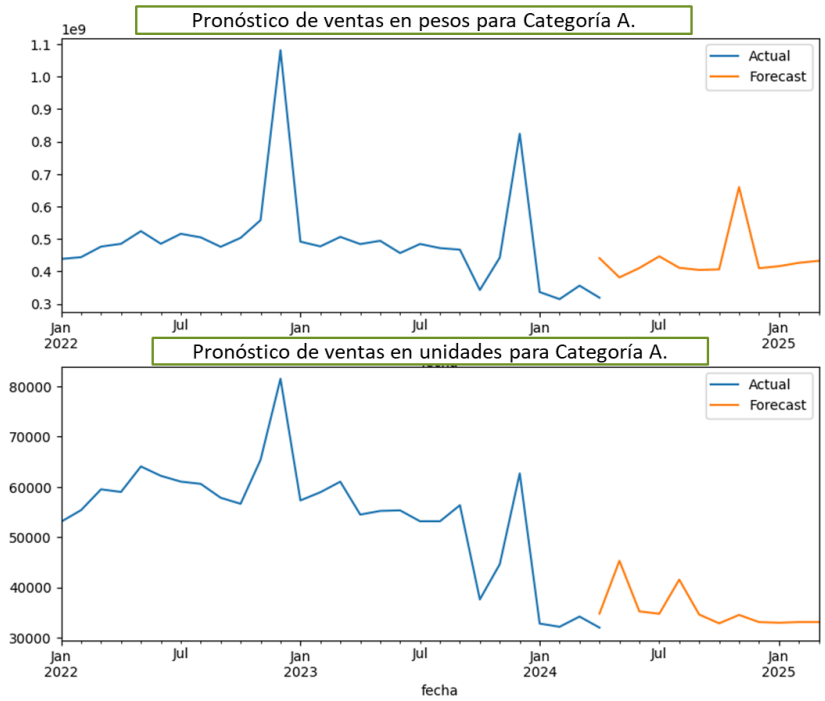
Tabla 2
Resultados del modelo forecasting.

Categoría	MAE ($)	RMSE ($)	MAE (Und.)	RMSE (Und.)
Categoría A	75488219.20	83452996.96	10654.59	13967.87
Categoría B	827049083.59	1043888441.85	69792.46	84225.82
Categoría C	24597873.79	27476946.79	1820.49	2229.95
Categoría D	60688103.12	73565413.67	3490.61	4359,29
Categoría E	280491625.92	311370835.12	8076.93	10044.43

Fuente. Elaboración propia.

Además, dentro de los resultados también se visualizaron las gráficas de los pronósticos de cada una de las categorías los cuales permitieron observar la poca precisión del modelo. Los gráficos se muestran se muestran a continuación:

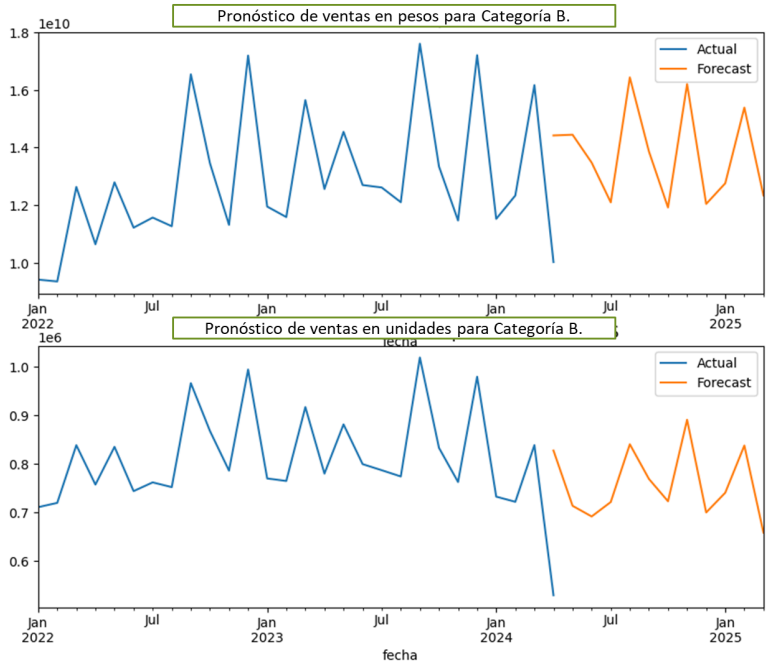
Figura 3
Grafica de pronósticos para categoría A.



Fuente. Elaboración propia.

Figura 4

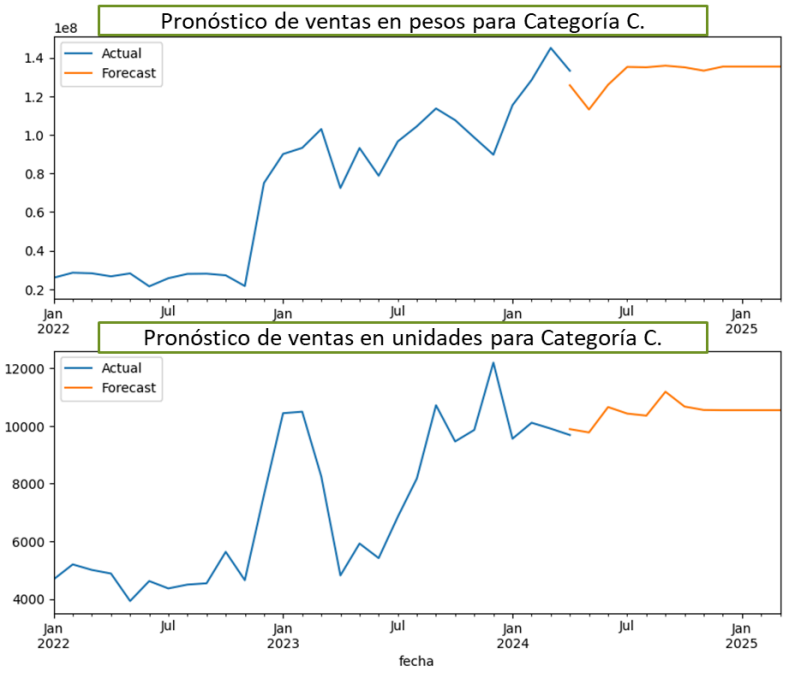
Grafica de pronósticos para categoría B.



Fuente. Elaboración propia.

Figura 5

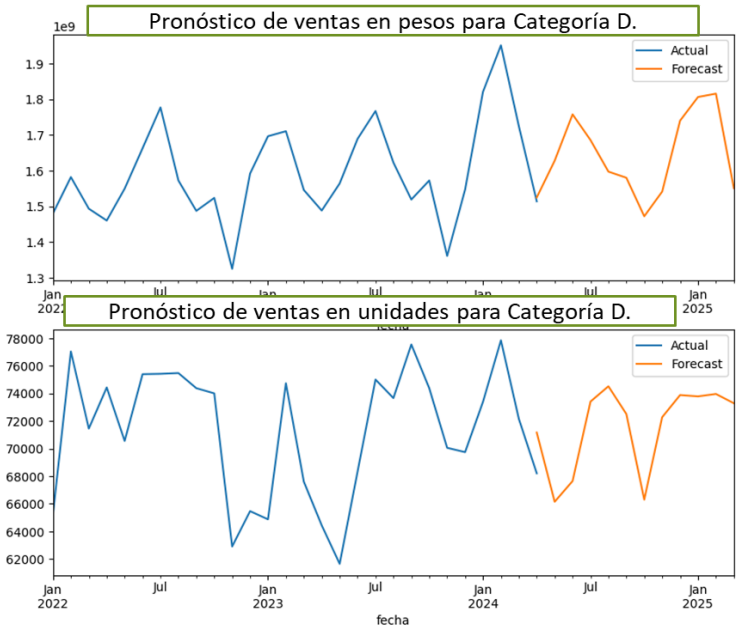
Grafica de pronósticos para categoría C.



Fuente. Elaboración propia.

Figura 6

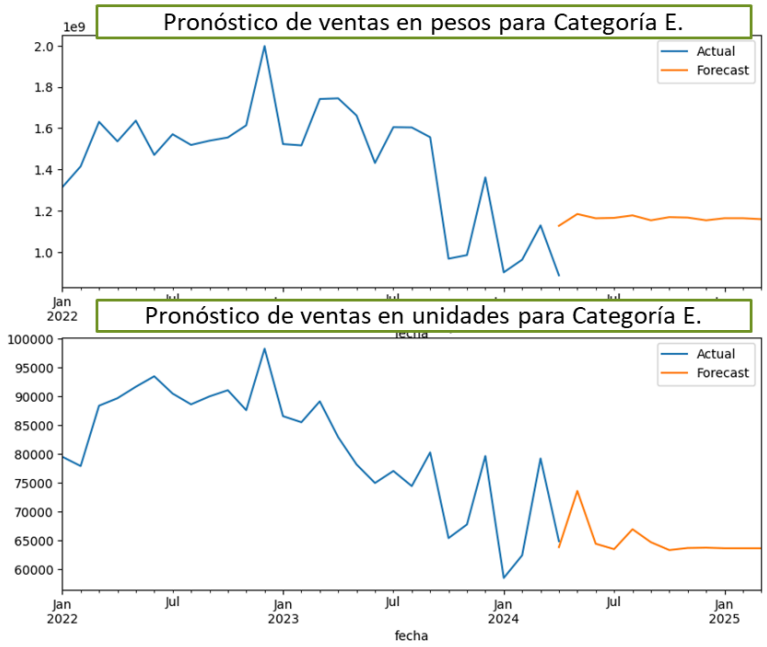
Grafica de pronósticos para categoría D.



Fuente. Elaboración propia.

Figura 7

Grafica de pronósticos para categoría E.



Fuente. Elaboración propia.

5 Análisis

El modelo de recomendación de productos demostró ser eficiente y preciso en su desempeño. Se analizaron un total de 4,267 registros, que corresponden a 837 productos distribuidos en 19 categorías. Las categorías con mayor número de productos recomendados fueron galletas, pastas y café molido. De los productos recomendados, el 35.48% pertenecen a Nutresa, mientras que el 64.52% pertenecen a la competencia; esta distribución sugiere que el modelo tiene una inclinación equilibrada al considerar tanto productos propios como de la competencia, lo que puede ser beneficioso para la diversidad del portafolio.

Un aspecto destacable es la precisión del modelo en cuanto a la presentación de productos recomendados, ya que solo el 10% de los productos Nutresa recomendados se presentaron en una variante diferente a la ya existente, lo que implica un bajo margen de error. Es de resaltar que el modelo no recomendó ningún producto que ya existiera en el portafolio actual de cada formato, evitando duplicaciones y demostrando una alta precisión en la recomendación de nuevos productos; este desempeño sugiere que el modelo está bien ajustado para identificar oportunidades de mercado y expandir el portafolio de productos sin redundancias.

El análisis del modelo de forecasting revela resultados menos satisfactorios debido a altos márgenes de error en varias categorías. **La Tabla 2** presenta los errores absolutos medios (MAE) y las raíces del error cuadrático medio (RMSE) en pesos y unidades para diferentes categorías. Por ejemplo, en la Categoría A, el MAE es de $75,488,219.20 y el RMSE es de $83,452,996.96, con errores en unidades de 10,654.59 y 13,967.87, respectivamente; estos valores indican una considerable variabilidad en las predicciones. La Categoría B muestra los errores más altos, con un MAE de $827,049,083.59 y un RMSE de $1,043,888,441.85, y errores en unidades de 69,792.46 y 84,225.82. Estos errores extremos sugieren que el modelo tiene dificultades significativas para predecir con precisión en esta categoría; otras categorías, como la C y la D, también presentan errores elevados, aunque en menor medida que la Categoría B. Además, al analizar los gráficos de cada una de las categorías se logra complementar lo dicho anteriormente; En la Categoría A (**Figura 3**), las ventas actuales presentan picos significativos que el pronóstico

no captura, manteniéndose relativamente plano. Esto indica que el modelo no está capturando adecuadamente las fluctuaciones en las ventas, resultando en un alto error de predicción. Las ventas en unidades muestran un patrón similar, con picos y caídas que no se reflejan en el pronóstico, subestimando consistentemente las ventas reales. Para la Categoría B (**Figura 4**), las ventas actuales muestran una variabilidad considerable con picos y valles, pero el pronóstico no sigue esta tendencia y se mantiene relativamente constante. Esta falta de seguimiento de la variabilidad en las ventas actuales indica un bajo rendimiento del modelo para esta categoría; por su parte, las ventas en unidades presentan un comportamiento similar al de las ventas en pesos, con una gran variabilidad en las ventas actuales que no se refleja en el pronóstico, resultando en errores significativos.

En la Categoría C (**Figura 5**), el gráfico muestra una tendencia creciente en las ventas actuales, pero el pronóstico se desvía ligeramente de esta tendencia, lo que indica que el modelo no está capturando completamente la tendencia creciente de las ventas. Aunque el pronóstico sigue parcialmente la tendencia creciente de las ventas en unidades, hay una desviación que indica que el modelo no está capturando con precisión las fluctuaciones. La Categoría D (**Figura 6**) presenta una variabilidad significativa en las ventas actuales, pero el pronóstico no refleja esta variabilidad y se mantiene relativamente constante. Esto sugiere que el modelo no está aprendiendo adecuadamente de los datos históricos. Las ventas en unidades siguen un patrón similar con un pronóstico que no captura la variabilidad de las ventas actuales, resultando en una subestimación constante. Finalmente, para la Categoría E (**Figura 7**), las ventas actuales presentan una tendencia a la baja con picos y valles pronunciados. El pronóstico no captura adecuadamente estas variaciones y se mantiene plano, subestimando las ventas; las ventas en unidades también están subestimadas consistentemente, mostrando una gran variabilidad y una tendencia decreciente que el modelo no está capturando correctamente.

El análisis de los gráficos y los resultados del modelo de forecasting revela varias deficiencias. El modelo muestra una incapacidad para capturar la variabilidad y las tendencias en los datos de ventas históricos, lo cual se refleja en las altas métricas de error (MAE y RMSE) discutidas anteriormente. Además, intentar predecir las ventas de diferentes categorías de

productos simultáneamente puede introducir errores, ya que modelar todas las series de tiempo como si fueran iguales puede ignorar comportamientos específicos y distintas estacionalidades de cada categoría.

6 Conclusiones

El aprovechamiento de la analítica de negocios puede brindar herramientas que generen valor al modelo de negocio de la organización, iniciando desde técnicas de exploración hasta la implementación de modelos heurísticos y de aprendizaje automático. En este proyecto, el desarrollo del Dashboard ha permitido la integración efectiva de la analítica de datos POS y algoritmos de ML, brindando una herramienta poderosa para la toma de decisiones empresariales. Los resultados muestran avances significativos en la recomendación de productos y áreas de mejora en la predicción de ventas. Se logró implementar un algoritmo robusto para la limpieza y preprocesamiento de datos, asegurando la calidad y coherencia de la información utilizada; esto se reflejó en la eficiencia del modelo de recomendación de productos, que demostró un bajo margen de error y alta precisión en sus sugerencias.

El análisis exploratorio proporcionó una comprensión profunda de las características principales de los datos POS, identificando patrones clave en las ventas y preferencias de productos. Este análisis fue fundamental para el desarrollo de los modelos de recomendación y forecasting. Por su parte, el algoritmo de recomendación de productos mostró un rendimiento sobresaliente, recomendando 837 productos en 19 categorías con una precisión notable. Se destacó la capacidad del modelo para evitar duplicaciones, ya que solo el 10% de los productos Nutresa recomendados se presentaron en una variante diferente a la existente en el portafolio, demostrando un bajo margen de error y una alta eficiencia en la identificación de nuevas oportunidades de mercado. Por último, el modelo de predicción de ventas presentó resultados insatisfactorios, con altos márgenes de error en varias categorías. Esto sugiere una necesidad de revisión y mejora del modelo para capturar mejor las tendencias y variabilidades en los datos de ventas históricos.

7 Recomendaciones

Para el modelo de recomendación de productos, se recomienda continuar utilizando y refinando el modelo, dado su buen desempeño. Adicionalmente, se recomienda mantener una base actualizada con los códigos únicos de la compañía para cada producto, ya que de esta forma se aumentaría la precisión del modelo al no recomendar un mismo producto en diferente presentación.

Para mejorar el modelo de forecasting, se recomienda realizar un análisis de calidad de los datos para asegurarse de que los datos históricos sean representativos y no contengan inconsistencias. Probar diferentes configuraciones de hiperparámetros puede mejorar la capacidad del modelo de captar las fluctuaciones y tendencias en los datos; incluir variables adicionales que puedan afectar las ventas, como promociones, eventos estacionales, o factores económicos, podría mejorar la precisión del modelo. Finalmente, dado que cada categoría de producto puede tener comportamientos y estacionalidades diferentes, desarrollar modelos separados para cada categoría permitiría capturar las dinámicas específicas de cada una, mejorando la precisión de las predicciones.

Referencias

Aguilar, L. J. (2019). *Inteligencia de negocios y analítica de datos: Una visión global de Business Intelligence & Analytics*. Alpha Editorial. https://books.google.es/books?hl=es&lr=&id=ifR5EAAAQBAJ&oi=fnd&pg=PR7&dq=anal%C3%ADtica+de+datos&ots=bCfbQJqDYi&sig=ADKGR32jr4mE-jHdLr7Gi5CRkXo#v=onepage&q&f=false

Cui, G., Wong, M. L., & Lui, H. K. (2006). Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. *https://doi.org/10.1287/mnsc.1060.0514*, *52*(4), 597–612. https://doi.org/10.1287/MNSC.1060.0514

Dekimpe, M. G. (2020). Retailing and retailing research in the age of big data analytics. *International Journal of Research in Marketing*, *37*(1), 3–14. https://doi.org/10.1016/j.ijresmar.2019.09.001

Dzyabura, D., & Yoganarasimhan, H. (2018). Machine learning and marketing. En *Handbook of Marketing Analytics* (pp. 255–279). Edward Elgar Publishing. https://doi.org/10.4337/9781784716752.00023

Meeting, B. S., & Chapman, P. (s/f). *The CRISP-DM User Guide*.

Pedregosa, F., Varoquaux, G. ¨el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *The Journal of machine Learning research*, *12*, 2825–2830.

skforecast · PyPI. (s/f). Recuperado el 13 de julio de 2024, de https://pypi.org/project/skforecast/

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, *1*, 29–39.

Anexos

Anexo 1. Imágenes del Dashboard desarrollado

Figura 8

Pestaña principal del Dashboard.

Figura 9

Pestaña resumen del Dashboard.

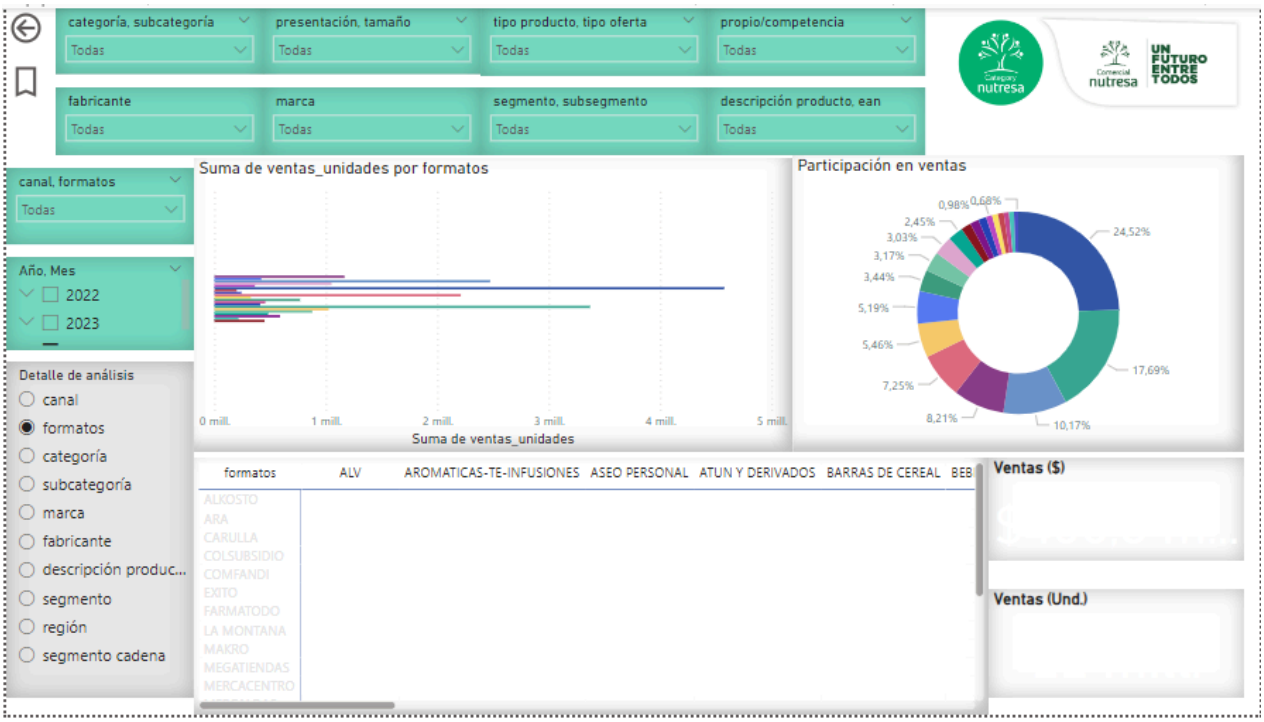


Figura 10

Pestaña top 5 por marcas y productos con tendencia.

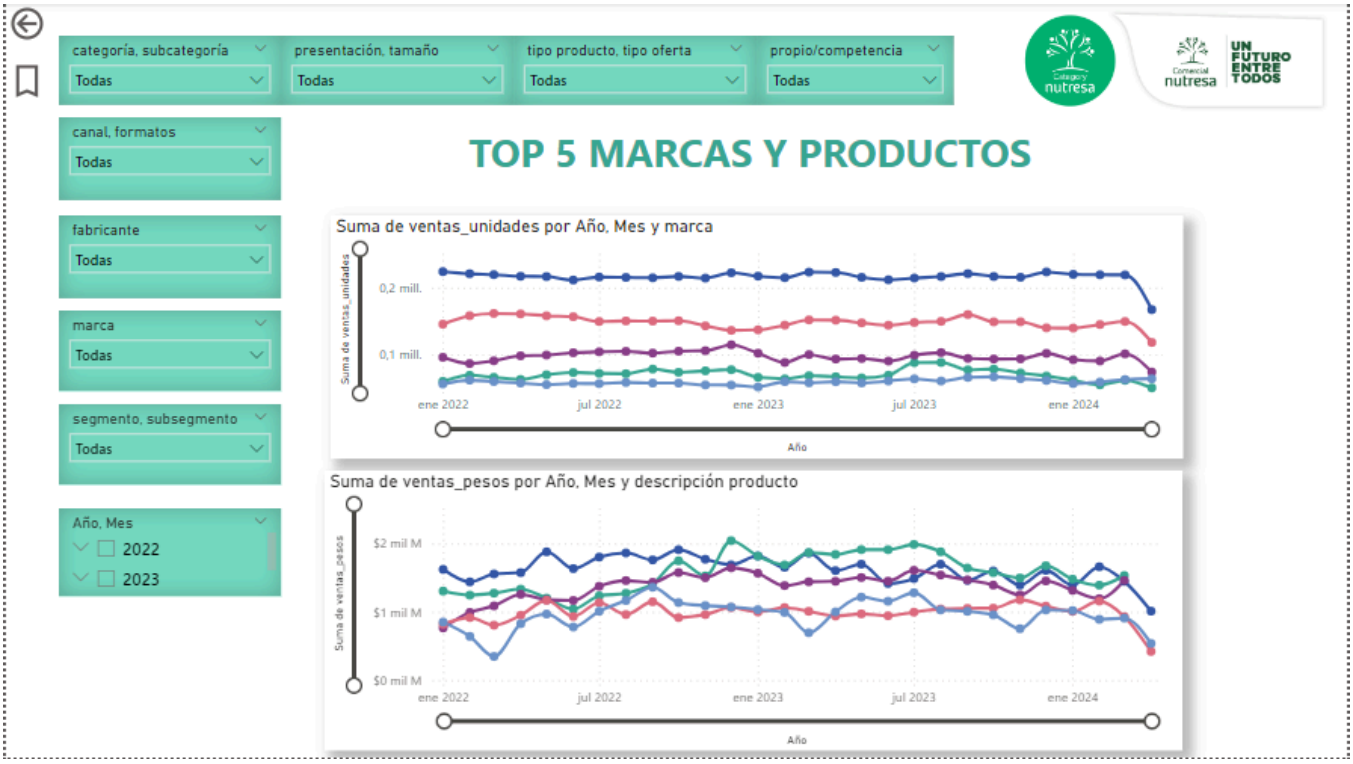


Figura 11

Pestaña de tendencia de ventas.

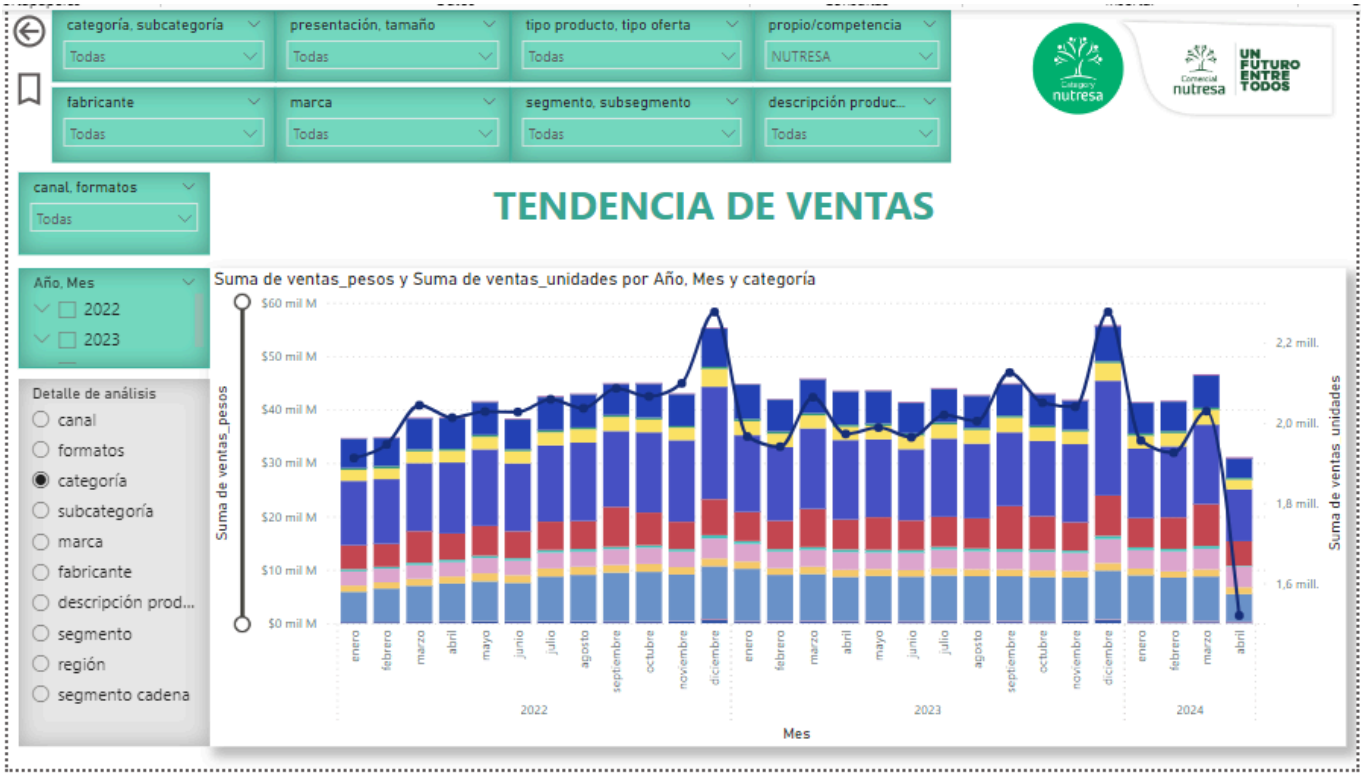


Figura 12

Pestaña oportunidades de codificación (recomendaciones).


canal, formato
Todas
categoría, ean
Todas
marca
Todas
propio/competencia
NUTRESA
nutresa
Comercial nutresa
UN FUTURO ENTRE TODOS
OPORTUNIDADES DE CODIFICACIÓN
Ventas($)
Ventas (Und.)
Categorías recomendadas
canal
Suma de prom_ventas_pesos
Suma de prom_ventas_unidades
AU $1.486.107.663 134477
GRANDES CADENAS $5.835.562.613 204679
ALKOSTO $532.546.026 23043
ARA $526.071.408 21445
CARULLA $541.621.846 18065
COLSUBSIDIO $759.884.514 27118
EXITO $427.837.508 12200
MAKRO $476.837.015 18847
OLIMPICA $605.019.328 17909
OXXO $978.006.461 34298
SUPERINTER $626.206.053 16950
SURTIMAX $361.532.451 14805
Total $7.321.670.276 339156