Modelo de aprendizaje no supervisado para la priorización de inventarios cíclicos

PROYECTO DE GRADO

Ryuma Jonathan Nakano
Edgar Felipe Torres

Asesor
Javier Díaz Cely, Ph. D.

FACULTAD DE INGENIERÍA
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
2021

Modelo de aprendizaje no supervisado para la priorización de inventarios cíclicos

Ryuma Jonathan Nakano
Edgar Felipe Torres

Trabajo de grado para optar al título de
Máster en Ciencia de Datos

Asesor
Javier Díaz Cely, Ph. D.

FACULTAD DE INGENIERÍA
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
2021

CONTENIDO

pág.

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ECUACIONES

LISTA DE ANEXOS

RESUMEN

Es común encontrar discrepancias en los sistemas de información de inventarios, entre las cantidades de los productos que el sistema dice tener y las cantidades físicas en la bodega. Existen múltiples causas para esto, principalmente errores humanos y del propio sistema de información. Estas discrepancias pueden ocasionar serios problemas tanto en las operaciones de bodega como en la relación con los clientes, ya que pueden generar demoras en los despachos, y peor aún, compromisos con clientes imposibles de cumplir. Los conteos cíclicos periódicos son una herramienta útil para corregirlas, pero para que sean efectivos, es de vital importancia identificar las discrepancias lo más temprano posible, y así lograr corregirlas antes de que tengan consecuencias negativas para el negocio. Lo anterior no es nada fácil de lograr cuando se tienen bodegas amplias con un alto número de productos.

El problema que fue abordado y resuelto en este trabajo de grado fue la identificación de las características de los productos a priorizar en los conteos periódicos, utilizando los datos históricos de correcciones de inventario para encontrar patrones que permitieran identificar los productos más propensos a requerir ajustes. Para abordarlo, se propuso la metodología CRISP-DM de seis fases que fueron abordadas siguiendo un enfoque iterativo e incremental.

Para resolver el problema, se propuso la utilización del algoritmo de *k-prototypes*, como una técnica de aprendizaje no supervisado apta para trabajar con datos tanto numéricos como categóricos. El algoritmo permitió identificar *clusters* (agrupaciones) a partir de los patrones encontrados en los datos históricos de correcciones de inventario. Los resultados fueron sometidos a experimentos de validación que permitieron confirmar tanto el número de *clusters*, como evaluar la calidad de los *clusters* resultantes, a partir de la comparación con los resultados obtenidos de aplicar las mismas técnicas, a nuevos conjuntos de datos generados utilizando técnicas de muestreo con reemplazo (*bootstrapping*). Una vez la

solución fue validada siguiendo los experimentos establecidos, se obtuvieron resultados que permitieron avalar la solución como válida para el problema abordado, al confirmar que los *clusters* resultantes eran estables y por lo tanto agrupaciones válidas basadas en patrones y no el simple fruto del azar.

Como resultado de la ejecución del proyecto de grado, se lograron identificar las características principales de los productos que deben ser priorizados en los conteos cíclicos periódicos con el objetivo de corregir discrepancias de inventario de manera oportuna y efectiva.

1. INTRODUCCIÓN

1.1 Contexto y Antecedentes

Para la gran mayoría de las organizaciones, el control de inventarios es un tema crítico. La alta gerencia así lo entiende, y las organizaciones hacen grandes esfuerzos por mantener niveles de inventario adecuados para su operación (Axsäter, 2015).

Toda planeación de inventarios – según la cual se toman decisiones respecto a qué productos se deben ordenar, cuándo, en qué cantidades, y de qué proveedores - parte del nivel de inventario que la empresa cree tener físicamente en sus bodegas. Con la aparición de los sistemas empresariales de negocios hace ya varias décadas, son los sistemas los que tienen la información sobre los niveles de inventario de los diferentes productos: cuantas unidades de cada producto hay en la bodega, y a un nivel mayor de detalle, cuantas unidades de cada producto hay en los diferentes estantes o ubicaciones de bodega. Cualquier imprecisión en los registros se convierte en un verdadero problema, ya que se podría dejar de ordenar un producto del que el sistema “cree” tener cantidades suficientes (aunque no sea así en la realidad), o se podría ordenar un producto del cual hay unidades suficientes en existencia, pero el sistema “cree” que no es así (Vidal, 2010 p. 71). Incluso, en casos en los que el nivel de inventario en el sistema es correcto, pero los estantes o ubicaciones de bodega (o sus balances) no los son, los esfuerzos requeridos para ubicar los productos y el tiempo invertido en investigar las inconsistencias ocasionarán demoras y costos adicionales que entorpecerán y encarecerán la operación.

El profesor Carlos Julio Vidal concluye en su libro Fundamentos de control y gestión de inventarios que “la inexactitud de los inventarios físicos es uno de los grandes obstáculos para la administración de los inventarios en una cadena de abastecimiento” (Vidal, 2010). Ya que, como es bien conocido, “los resultados de

cualquier modelo, por sofisticado que sea, dependen de la información que se le suministre". Asevera también el profesor Vidal, que se hacen necesarios "conteos manuales y corrección de los registros" para que el sistema tenga mayores posibilidades de éxito.

Este proyecto tiene como caso de estudio una empresa de distribución del sureste de los Estados Unidos, en la cual se han ensayado diferentes estrategias para la priorización de productos en la realización de conteos cíclicos periódicos de inventario. Se trata de pequeños conteos manuales diarios de subconjuntos de productos o ubicaciones de bodega según la estrategia implementada en el momento. Lo cíclico, tiene que ver con el número de veces que se debe contar un producto - según políticas corporativas - durante un año calendario. Por ejemplo, en una de las estrategias implementadas, se contaban productos según su clasificación de inventario ABC, y cada uno de los productos "A" debía ser contado doce veces al año (una vez al mes). La empresa hace parte del segmento B2B (*business-to-business* o "negocio-a-negocio" en español, lo cual significa que realiza negocios con otras empresas sin llegar al consumidor o cliente final) y cuenta con proveedores tanto domésticos como internacionales. Ninguna de las estrategias de inventarios cíclicos ensayadas ha sido particularmente exitosa, por lo que este problema sigue siendo un importante tema pendiente por resolver. Esperar hasta la realización del inventario físico anual en el mes de diciembre (en caso de realizarse) para hacer los ajustes requeridos a los niveles de inventario resulta especialmente costoso e inconveniente, y enfrentar problemas de despacho e incumplimiento a clientes por bajos niveles de inventario (o la otra cara de la moneda, exceso de ellos) como resultado de errores en la planeación ocasionados por las discrepancias genera perjuicios a la operación del negocio. Es por todo lo anterior, que urge explorar nuevas opciones para solucionar este problema.

1.2 Planteamiento del Problema

Ante las dificultades de la organización, para encontrar una estrategia de inventarios acorde con la dinámica y operación del negocio, así como la necesidad de encontrar una manera de corregir las discrepancias entre los niveles de inventario en el sistema y las existencias físicas en bodega de manera oportuna, se ha optado por explorar la amplia cantidad de datos de transacciones de inventario disponibles en el sistema en busca de respuestas a esta problemática. El proyecto tiene como objetivo "entablar una conversación" con los datos, a través de técnicas de *machine learning* de aprendizaje no supervisado, para lograr encontrar patrones en los datos históricos de ajustes de inventario, y así identificar los productos a priorizar de acuerdo con los hallazgos. Los productos identificados como los más susceptibles a requerir correcciones en sus niveles de inventario, serán identificados como los elementos con "mayor riesgo", por lo que se considerarán de alta prioridad en la estrategia de inventarios cíclicos de la organización.

1.3 Objetivo General

Formular y evaluar un modelo que permita efectuar la priorización oportuna de los productos para la implementación de una estrategia de conteo cíclico.

1.4 Objetivos Específicos

1. Estructurar la información de transacciones de ajustes de inventario que se encuentra disponible en la organización para permitir efectuar su apropiado análisis.
2. Establecer un conjunto de posibles técnicas que permitan procesar la información y generar el conocimiento necesario para efectuar la priorización.

3. Obtener una priorización de productos que permita apoyar la estrategia de inventarios cíclicos periódicos de la organización.

1.5 Organización del Documento

El documento está estructurado de la siguiente manera: en el capítulo 2 se presenta el marco teórico, tanto en lo relacionado con el área del negocio, como con las técnicas de *machine learning* de aprendizaje no supervisado utilizadas; en el capítulo 3 se describen las fases de la metodología CRISP-DM y su adaptación al proyecto; el capítulo 4 describe el proceso de selección y recolección de información, y presenta el conjunto inicial de datos disponible; el capítulo 5 discute el preprocesamiento de datos, así como las decisiones de ingeniería de datos tomadas para llegar al conjunto final de datos a analizar; el capítulo 6 discute detalles de la implementación de clustering como la técnica de *machine learning* utilizada para resolver el problema; el capítulo 7 describe los experimentos de validación realizados para confirmar tanto el número de *clusters* seleccionados, como la calidad de los clusters resultantes; el capítulo 8 presenta los resultados obtenidos en el proceso, y el capítulo 9, las conclusiones del proyecto.

2. ANTECEDENTES

2.1 Marco Teórico

2.1.1 Conceptos del área del negocio

Desde el punto de vista contable, el **inventario** es un activo corriente que representa una propiedad tangible disponible para la venta como parte de las actividades regulares del negocio. En empresas donde se realizan actividades de producción o transformación, puede tratarse de un artículo en proceso, o materias primas a ser utilizadas como insumos para la fabricación. Para el caso que compete a este proyecto, desarrollado en una empresa del sector de distribución, la definición de inventario se limita a la presentada inicialmente: productos terminados disponibles para la venta. Dados los altos costos de tener inventario en la bodega, así como los problemas que acarrea no tener disponibilidad cuando se presenta una oportunidad de venta, el manejo del nivel de inventario se ha convertido en una medida fundamental del desempeño de la cadena de abastecimiento y todas las actividades logísticas que la componen (Waller, 2014). Es por lo anterior, que "el control de inventarios es uno de los temas más complejos y apasionantes de la Logística (sic) y de la planeación y administración de la cadena de abastecimiento" (Vidal, 2010 p. 15).

Las decisiones finales de manejo y control de inventarios se realizan al nivel de ítems individuales, llamados **SKU** por su nombre en inglés (*Stock Keeping Unit*). Un SKU es una unidad de inventario que se puede diferenciar claramente de otra, y tiene su propio código en el sistema de información. Dependiendo de la naturaleza del negocio, las diferencias entre un SKU y otro pueden ser tan sutiles como el tono de su color – como es el caso de la industria en el que se desempeña la empresa donde se desarrolló este proyecto, aunque en otros casos pueden existir clasificaciones con un mayor nivel de agregación donde un solo SKU puede representar familias de artículos semejantes (Vidal, 2010). La

clasificación **ABC** permite catalogar los distintos SKU de la compañía como A, B, o C en función de su contribución anual a las ventas de la compañía – o alternativamente, con base en cualquier otra fórmula que permita medir la importancia de cada uno de los SKU para el negocio. La fórmula tradicionalmente utilizada multiplica las ventas anuales de cada producto por su costo unitario. Los resultados son organizados de mayor a menor, donde la parte alta de la tabla, tradicionalmente el 80%, corresponde a los productos A, la parte media a los productos B, y la parte baja a los productos C. Los porcentajes utilizados pueden variar según la naturaleza del negocio y las políticas de la compañía (Vidal, 2010).

Los sistemas integrados de información empresarial conocidos como **ERP** por su nombre en inglés (*Enterprise Resource Planning*) son los más poderosos y complejos sistemas de información empresarial disponibles en el mercado. Los sistemas ERP se enfocan principalmente en los procesos internos, integrando los diferentes procesos de negocio de la organización. Los sistemas ERP están tradicionalmente compuestos por múltiples módulos, entre los que se encuentran producción, recursos humanos, finanzas y contabilidad, ventas y distribución (Magal, 2011). Los datos utilizados en este proyecto provienen del archivo histórico de transacciones de inventario del módulo de distribución del ERP de la compañía.

Tal y como se discutió en la sección 1.1, cualquier imprecisión en los registros de inventario se convierte en un problema, ya que podría dejarse de ordenar un producto del cual el sistema ERP “cree” tener una cantidad mayor a la que se tiene físicamente en bodega o, por el contrario, ordenar un producto del cual se tiene suficiente inventario, pero el sistema ERP “cree” tener una cantidad menor (Vidal, 2010). En una situación ideal, las cantidades físicas en bodega y los balances de inventario en el sistema ERP siempre están de acuerdo, y toda actividad física que genera un aumento o una disminución del nivel de inventario de un producto corresponde a una transacción en el sistema que la representa

adecuadamente. Desafortunadamente, hay múltiples situaciones en el mundo real que ocasionan **desbalances** entre la realidad en la bodega y los niveles de inventario en el sistema ERP. Esto incluye una variedad de errores humanos en la recepción y despacho de productos, tales como cantidades erradas, productos equivocados, confusión en el manejo de las unidades de empaque (UoM por su nombre en inglés, *Unit of Measure*), procesos inconsistentes o propensos a errores, e incluso errores en el sistema ERP, sea por fallas de los usuarios, o por errores propios del sistema (Muller, 2003).

El método tradicional de corrección ha sido el inventario físico anual, en el cual se hace un conteo exhaustivo de todos los productos en la bodega. Dos grandes inconvenientes de este método son la magnitud de la tarea, y el largo tiempo que puede pasar desde el momento en que se crea un desbalance hasta su corrección. Los **conteos cíclicos periódicos** presentan una solución oportuna a estos dos problemas, ya que permiten implementar un método sistemático para reconciliar los niveles de inventario en el sistema ERP con las cantidades físicas en bodega. Se trata de contar frecuentemente los productos considerados como estadísticamente significativos entre todos los SKU en inventario. Contar frecuentemente un reducido número de SKU – algo muy diferente a la tarea monumental de contar todos los productos en bodega - brinda una mejor oportunidad de identificar y corregir cualquier discrepancia antes de que ocasione mayores perjuicios al negocio (Muller, 2013). Es por tanto de vital importancia utilizar un buen método de priorización de los productos a contar con mayor frecuencia en los conteos cíclicos periódicos. Algunos métodos tradicionales incluyen: conteo de secciones físicas de la bodega, selección completamente aleatoria de productos, conteos por categorías o agrupaciones de productos, y priorización ABC. La implementación de una efectiva estrategia de conteos cíclicos le brinda la opción a la compañía de eliminar la necesidad de realizar un inventario físico anual (Muller, 2013).

Este proyecto propone continuar con la priorización ABC establecida por lineamientos corporativos, pero haciendo un énfasis especial – en cuanto a oportunidad y frecuencia – de las agrupaciones de productos identificadas, a partir del análisis de las transacciones de ajuste de inventario, como las más propensas a sufrir desbalances. Las transacciones de ajustes de inventario utilizadas como insumo para el proceso de analítica de datos de este proyecto son el fruto de las correcciones históricas resultantes de los conteos cíclicos e inventarios físicos de la compañía. Es importante anotar que, aunque las transacciones de ajustes de inventario utilizadas como insumo de datos para este proyecto fueron registradas a nivel SKU, tanto el alto número de SKU existentes en la compañía, como el conocimiento del negocio, llevaron a buscar una priorización a un nivel de agregación mayor. Priorizar un listado de SKU no hubiera sido valioso para el negocio, mientras que el identificar un nivel de agregación mayor a partir de las transacciones de SKU individuales si lo es.

2.1.2 Conceptos relacionados con las técnicas utilizadas

Según la Real Academia de la Lengua Española (RAE), un modelo es una representación de la realidad. Una simplificación “de una realidad compleja que se elabora para facilitar su comprensión y el estudio de su comportamiento” (Real Academia Española, s.f.a).

El *machine learning* nació como respuesta a la curiosidad humana sobre la capacidad de un computador de ir más allá de las instrucciones que le hemos dado, y aprender por si solo a realizar una tarea. El aprendizaje, en el contexto de *machine learning*, describe el proceso automático de búsqueda de mejores representaciones de los datos, a través de transformaciones que les permitan ajustarse mejor a la tarea a realizar (Chollet, 2017 pp. 28-30).

Un modelo de *machine learning* es “entrenado”, en lugar de ser programado como se acostumbra en la computación tradicional. El entrenamiento consiste en presentarle ejemplos relevantes de la tarea a realizar, para que el modelo busque estructura en estos ejemplos (que no son más que las observaciones o datos de entrada mencionados anteriormente), y obtenga un conjunto de reglas que le permita automatizar la tarea (Chollet, 2017 p. 28). En el caso de este proyecto, los ejemplos son transacciones de ajuste de inventario, y la tarea a realizar es su agrupación con fines de priorización, a partir de los patrones encontrados en los datos.

Las tareas de *machine learning* son normalmente categorizadas en tres grandes grupos: 1) **Aprendizaje supervisado**, en el cual el objetivo es generalizar un modelo a partir de datos de entrenamiento con “etiquetas” – donde las etiquetas corresponden a las respuestas conocidas con antelación para los datos de entrenamiento - un ejemplo podría ser la predicción de niveles de inventario para un SKU basada en la demanda histórica del producto, o la clasificación de nuevos productos como A, B, o C según sus características; 2) **Aprendizaje no supervisado**, el cual se ocupa del aprendizaje a partir de datos sin “etiquetas”, por lo que busca descubrir patrones desconocidos a partir de los propios datos que permitan su categorización – un ejemplo de esto es el problema que se enfrenta en este proyecto, donde se busca encontrar grupos de productos a priorizar a partir de datos históricos de transacciones, pero sin conocer previamente las categorías resultantes; y 3) **Aprendizaje por refuerzo**, en el cual se busca mapear situaciones a acciones que produzcan una máxima recompensa - un ejemplo ampliamente conocido de aprendizaje por refuerzo, es el de los programas que controlan los vehículos autónomos, los cuales deben interactuar permanentemente con condiciones dinámicas y ajustar su estado para lograr maximizar el logro de sus objetivos (Swamynathan, 2017).

Dentro del aprendizaje no supervisado, se reconocen tres áreas principales de aplicación: 1) ***Clustering*****,** donde el objetivo es dividir el conjunto de datos en grupos lógicos, relacionados entre sí, pero no conocidos con anterioridad; 2) **Reducción de dimensionalidad**, donde se pretende simplificar un conjunto de datos de múltiples dimensiones, en un espacio con una dimensionalidad menor; y 3) **Detección de anomalías**, donde se busca detectar aquellas observaciones que no se ajustan al patrón esperado (Swamynathan, 2017).

El proyecto que se presenta en este documento resuelve un problema de *clustering*, que tal y como se mencionó en el párrafo anterior, pertenece a la categoría de *machine learning* de aprendizaje no supervisado. El objetivo del *clustering* es separar un conjunto de datos finito en un número finito y discreto de estructuras de datos naturales denominadas *clusters*. La razón principal para hacer *clustering* es la necesidad de explorar la naturaleza desconocida de datos (Xu, 2009). Ese es el caso de este proyecto, donde se busca obtener respuestas de priorización de productos a partir de transacciones de ajustes de inventario que no proveen información explicita previa sobre cómo deben ser agrupadas.

Xu (2009) cita, en su libro *Clustering*, los principales objetivos del *clustering* presentados previamente por Aldenderfer y Bashi en 1984:

- Desarrollar una categorización o segmentación.
- Investigación de esquemas conceptuales útiles para agrupar entidades.
- Generación de hipótesis a partir de la exploración de los datos.
- Prueba de hipótesis o el intento de determinar si los tipos definidos están realmente presentes en los datos.

El *clustering* “divide un grupo de objetos en un número más o menos homogéneo de subgrupos con base en una medida subjetiva de similitud – escogida subjetivamente con base en su habilidad para crear *clusters* interesantes, de manera que la similitud entre los objetos dentro del mismo subgrupo sea mayor

que la similitud entre objetos pertenecientes a otros *clusters*" (Xu, 2009). Cabe mencionar, que la selección de un criterio o algoritmo diferente de *clustering* puede arrojar resultados completamente diferentes. Lo anterior, describe claramente el elemento subjetivo siempre presente en un proceso de *clustering* (Xu, 2009).

Un concepto fundamental en *clustering* es el de la medida de proximidad, o que tan similar o diferente es considerada una observación de otra, ya que esta medida determina que observaciones comparten el mismo *cluster* y cuales terminan en *clusters* diferentes. La media de proximidad a utilizar para comparar características o variables de un conjunto de datos depende en gran medida del tipo de datos a comparar. Existen tres tipos principales de atributos o variables en un conjunto de datos: 1) **variables continuas**, que son aquellas que pueden adoptar un número infinito de valores, 2) **variables discretas**, que son aquellas que solo pueden tomar un numero finito de valores, y 3) **variables binarias** o dicotómicas, que pueden tomar exactamente dos valores. La distancia euclidiana (ver Ecuación 1) – también conocida como la norma L_2 – es la más comúnmente utilizada para comparar variables continuas. Igualmente, existen medidas de proximidad comúnmente utilizadas para comparar variables discretas y binarias, pero se omite profundizar en el tema, ya que el conjunto de datos del proyecto es considerado un conjunto de datos mixto, dada la presencia de variables tanto continuas como discretas o categóricas en cada una de las observaciones, lo cual lleva a aplicar una medida de proximidad particular que incluye el cálculo de la distancia euclidiana para la parte continua de cada observación y un cálculo de disimilitud para la parte discreta o categórica, según lo propuesto por Huang en su artículo publicado en 1998 (Huang, 1998) y discutido un poco más adelante en esta sección. Tal y como lo puntualiza Xu en su libro, corresponde a los investigadores seleccionar una medida de proximidad y los atributos a utilizar, como parte indispensable del marco de trabajo del ejercicio (Xu, 2009).

Ecuación 1. Cálculo de la distancia euclidiana entre dos puntos i, j del conjunto de datos X de d dimensiones.

$$D(X_i, X_j) = \left(\sum_{k=1}^{d} \sqrt{|X_{ik} - X_{jk}|} \right)^2$$

Fuente: Clustering (Xu, 2009).

Las técnicas de *clustering* más conocidas y comúnmente utilizadas son: 1) el ***clustering* particional**, el cual divide el conjunto de puntos o datos en un número previamente seleccionado de *clusters* – sin ningún tipo de estructura jerárquica, y 2) el ***clustering* jerárquico**, el cual agrupa los datos en una serie de particiones anidadas que pueden ir desde un solo *cluster* para todos los datos u observaciones, hasta cada observación en su propio *cluster* – y viceversa. Adicionalmente, existen otras técnicas menos populares o de aplicaciones más específicas como pueden ser el *clustering* basado en densidad para datos geográficos o posicionales, el *clustering* basado en redes neuronales, el *clustering* basado en *kernels*, el *clustering* de datos secuenciales (aplicable a series de tiempo), y el *clustering* para datos a gran escala (Xu, 2009).

La técnica de *clustering* seleccionada para este proyecto, dada la naturaleza de sus datos, fue la de *clustering* particional, de la cual *k-means* es el algoritmo más reconocido y comúnmente aplicado (Xu, 2009), y es la base para el algoritmo propuesto por Huang que fue seleccionado para resolver el problema en cuestión. "*k-means* busca crear una partición óptima de los datos a partir de minimizar la suma de los errores cuadrados a través de un procedimiento de optimización iterativo" (Xu, 2009) para un número preestablecido de k *clusters* así:

- Inicializa las k particiones.
- Asigna cada observación al *cluster* más cercano.
- Recalcula los *clusters* basado en las nuevas particiones.
- Repite hasta que los *clusters* no sufran cambio alguno.

k-means es el algoritmo más popular de *clustering* particional, pero está limitado a trabajar con variables continuas (Xu, 2009). Es por esto, que el conjunto de datos mixtos – combinación de variables continuas y discretas o categóricas – de este proyecto requiere de una alternativa diferente.

Zhexue Huang "presenta dos nuevos algoritmos que utilizan el paradigma de *k-means* para *clustering* de datos categóricos" (Huang, 1998) en su artículo "*Extensiones al algoritmo k-Means para Clustering de grandes volúmenes de datos con variables categóricas*". Para Huang, *k-means* es un algoritmo eficiente, pero "su limitación a trabajar solo con valores numéricos le impide ser utilizado en el mundo real con valores categóricos". Huang también considera, que el método tradicional de convertir variables categóricas en numéricas no puede ser generalizado, por lo que se hace necesario plantear una nueva alternativa.

El algoritmo ***k-modes*** permite extender *k-means* para ser utilizado con variables categóricas, a partir del cálculo de una medida de disimilitud que permita comparar observaciones categóricas (ver Ecuación 2), y la utilización de modas en lugar de medias para calcular los *clusters*. Por motivos de eficiencia computacional, el algoritmo *k-modes* sigue estos pasos para un número prestablecido de k *clusters*:

- Selecciona k modas, una para cada cluster.
- Asigna cada observación al cluster con la moda más cercana según la medida de disimilitud (Ver Ecuación 2).
- Actualiza la moda de cada cluster después de cada asignación.
- Recalcula la medida de disimilitud después de que todas las observaciones han sido asignadas, y procede a reubicar aquellas más cercanas a otro cluster que al propio.
- Recalcula las modas para los clusters afectados.
- Repite los dos últimos pasos hasta que ninguna observación cambie de cluster después de un ciclo completo para todo el conjunto de datos.

Ecuación 2. Cálculo de la medida de disimilitud (*dissimilarity*) entre datos categóricos X y Y.

$$d_1(X,Y) = \sum_{i=1}^{m} \delta(x_j, y_j)$$

donde

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$$

Fuente: Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values.

El algoritmo ***k-prototypes*** integra *k-means* y *k-modes* para realizar *clustering* con datos mixtos (i.e. observaciones con valores tanto numéricos como categóricos). El algoritmo calcula una medida de disimilitud que tiene en cuenta tanto los valores numéricos – calculando la distancia euclidiana cuadrada como primer término, como los valores categóricos – utilizando como segundo término, la misma medida de disimilitud del algoritmo *k-modes*, multiplicada por un peso gamma (γ) que intenta evitar que se favorezca alguno de los dos términos de la ecuación (ver Ecuación 3).

Ecuación 3. Cálculo de la medida de disimilitud (*dissimilarity*) entre dos puntos de datos mixtos X y Y.

$$d_2(X,Y) = \sum_{\mathrm{j}=1}^{p} (x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j)$$

Fuente: Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values.

Los pasos del algoritmo no cambian, ya que en palabras de Huang “nada ha cambiado”, excepto el cálculo de la medida de disimilitud. Huang también considera, que el algoritmo *k-prototypes* es más útil que *k-modes*, por su aplicación en el mundo real, donde es común encontrar observaciones que incluyen atributos tanto continuos como discretos o categóricos (Huang, 1998).

Huang plantea dos tipos de inicialización diferentes, tanto en la publicación original en la que introduce el algoritmo de *k-modes* (Huang, 1997), como en el artículo discutido en los párrafos anteriores, en el que presenta *k-prototypes* como una extensión para trabajar con conjuntos de datos mixtos (Huang, 1998). El primer tipo de inicialización simplemente selecciona las primeras *k* observaciones diferentes del conjunto de datos como los centroides iniciales; mientras que el segundo método, más elaborado, organiza de mayor a menor frecuencia los valores únicos de cada una de las variables categóricas, y luego utiliza las listas resultantes para formar nuevas filas de valores que asemejan nuevas observaciones, y procede a buscar y asignar como centroides iniciales, los puntos del conjunto de datos más cercanos a estas nuevas filas según la medida de disimilitud de Huang (Huang, 1998).

La implementación de *k-protoypes* del paquete *k-modes* en el lenguaje *Python* utiliza el segundo método con una variación, al organizar los valores de cada variable categórica de menor a mayor, pero manteniendo todos los valores – sin limitarlos a valores únicos – para luego escoger de manera aleatoria aquellos que conforman las nuevas filas que posteriormente se comparan con el conjunto de datos para seleccionar los centroides iniciales. El autor del paquete justifica el componente aleatorio con un comentario en el código, en el que menciona que, al repetirse los valores según su número de ocurrencias en el conjunto de datos, aumenta la probabilidad de que tengan presencia en las nuevas filas a utilizar para seleccionar los centroides iniciales (de Vos, 2021). Adicional a la inicialización de Huang, la implementación en *Python* ofrece como alternativa la inicialización de Cao, basada en el artículo de Fuyuan Cao (Cao, 2009), la cual sugiere calcular la densidad promedio del objeto, para luego utilizar tanto la distancia como la densidad en la selección de los centroides iniciales.

El **coeficiente silueta**, fue la métrica elegida para seleccionar el número de *clusters* resultantes de la ejecución del algoritmo *k-prototypes* sobre el conjunto de datos de transacciones de ajustes de inventario del proyecto. Esta métrica se distingue por valorar tanto la cohesión intra *cluster* como la separación entre *clusters*. El mejor valor posible del coeficiente silueta es 1 e indica *clusters* completamente definidos y separados entre sí, mientras que su peor valor es -1 y representa problemas en la asignación de los *clusters*, siendo 0 un valor intermedio que indica que los clusters se traslapan, y por tanto el ejercicio de *clustering* resulta particularmente inefectivo (scikit-learn versión 0.24.2, 2021).

Es importante antes de cerrar esta sección, hacer énfasis en la importancia de los experimentos de validación después de un ejercicio de *clustering*, tal y como lo menciona Xu en su libro, cuando dice que “es indispensable justificar los resultados del *clustering*, y asegurarse de entender correctamente la estructura intrínseca de los datos” (Xu, 2009 p. 277). Xu hace particular énfasis, en la importancia de incluir la confirmación del número de *clusters* seleccionado como uno de los experimentos de validación. Cabe anotar, que los métodos de muestreo con reemplazo o ***bootstrapping*** por su nombre en inglés, resultan muy útiles cuando se trata de implementar enfoques heurísticos de validación como los utilizados en este proyecto. *Bootstrapping* es una poderosa herramienta estadística, que permite obtener nuevos conjuntos de datos, a partir del muestreo repetitivo de observaciones del conjunto de datos original (James, 2013).

2.2 Estado del arte / trabajos relacionados

2.2.1 Trabajos seleccionados

Existe una amplia variedad de investigaciones y artículos tanto sobre inventarios, y en particular para el tema de este proyecto, su clasificación, y los problemas que acarrean las discrepancias entre la realidad física en bodega y los balances en los

sistemas de información, como sobre técnicas de clustering y su utilización en procesos de priorización. A pesar de ello, no se logró encontrar literatura publicada en revistas académicas que combinara los dos temas, y contara con un buen número de citaciones en otros trabajos. Por lo anterior, se tomó la decisión de seleccionar trabajos sobre los dos temas, aunque fuera por separado. Los siguientes trabajos académicos, desarrollados para abordar problemáticas similares, han sido seleccionados para evaluar la relevancia del proyecto "Modelo de aprendizaje no supervisado para la priorización de inventarios cíclicos" presentado en este documento:

Martin (2007), propuso una metodología para estimar el máximo número rentable de rotaciones para un sistema de clasificación ABC (*A methodology for estimating the maximum profitable turns for an ABC inventory classification system*). En este artículo publicado por Warren Martin y Robert E. Stanford en la revista IMA Journal of Management Mathematics, los autores definieron una serie de pasos para calcular la cantidad de rotaciones y optimizar la rentabilidad en la adquisición de inventarios, a través de la clasificación de los productos, y de la identificación de los costos de almacenamiento en adición al valor de venta. Estos pasos incluyen:

- Identificación de las características de las clases del inventario.
- Determinación de los costos de mantenimiento.
- Cálculo de número de turnos para las clases de inventario de la compañía.

El estudio concluye que la implementación de esta metodología permitiría a los administradores de inventarios definir los tamaños de las órdenes para cada uno de los tipos de productos, así como validar la predicción de las rotaciones, y compararlas con las almacenadas en el sistema, con la finalidad de resaltar aquellas que requieran una mayor atención.

Kang (2005), muestra el impacto que puede tener en los procesos de una compañía, cualquier diferencia entre la información del sistema y la información real en las bodegas, pudiendo incluso generar situaciones severas de falta de

inventario que podrían resultar en una fuerte reducción de los ingresos, y poner en riesgo la continuidad del negocio. En el artículo “Discrepancias en sistemas de inventario: pérdidas e inventarios agotados” (*Information inaccuracy in inventory systems: stock loss and stockout*) publicado por Yun Kang y Stanley B. Gershwin en la revista IIE Transactions, los autores identifican como la modernización de los procesos de gestión de inventarios, que ha impulsado la automatización y la toma de decisiones basada en la información del sistema, trae como consecuencia que la calidad, precisión, y actualización de los datos tomen mayor relevancia y su inadecuado manejo se convierta en un importante riesgo para la organización.

Las discrepancias de inventario son muy comunes dentro de las compañías, y son causadas principalmente por pérdidas de inventario, errores en las transacciones, inventarios inaccesibles, o la identificación incorrecta de productos. Dichas causas podrían presentarse de manera intencional o fortuita, tanto al interior como de manera externa de los procesos, sistemas e infraestructura de la organización.
Para tratar de disminuir la materialización del riesgo de las discrepancias de inventarios, los autores establecen que es de vital importancia definir políticas internas de acuerdo con las siguientes estrategias:

- Inventario de seguridad: definición de una estrategia para la adquisición de un inventario pequeño basado en el conocimiento sobre la demanda de los productos en la organización.
- Verificación manual del inventario: realización de conteos periódicos de todos los productos del inventario, y actualización de la información en el sistema.
- Reinicio manual de los registros de inventario: búsqueda manual de patrones en la información del sistema por parte de su administrador, para poder balancear las posibles diferencias.
- Reducción constante del registro de inventario: realizar compras más pequeñas durante cada transacción, de manera que el posible error vaya disminuyendo.

- Auto identificación: utilización de sistemas de software y hardware que permitan estimar la ubicación y cantidad de los inventarios sin la intervención humana, como: RFID, básculas, bandas inteligentes, entre otras.

Boylan (2008), define un marco de trabajo conceptual que permite categorizar los productos de acuerdo con la frecuencia de compra. En el artículo “Clasificación para predicción y control de inventario: un caso de estudio” **(***Classification for forecasting and stock control: a case study*) publicado por John Boylan, Aris Syntetos y George Karakostas en la revista Journal of the Operational Research Society, los autores utilizan los datos de una empresa para identificar falencias en la clasificación de los productos con fines de predicción de los niveles de inventario, utilizando evidencia empírica de la demanda a partir de categorías y patrones no visibles. Los autores buscan determinar si los productos se adquieren de manera intermitente o irregular, lo cual no permitiría encontrar una distribución normal de su demanda – fenómeno que denominan “no normales” (*non-normals*).

En el artículo también se describen las siguientes clasificaciones de la demanda:

- Movimiento lento (slow moving): productos cuya demanda promedio por periodo es baja.
- Errática (erratic): productos cuya variabilidad en la demanda es muy volátil.
- Abultada (lumpy): productos intermitentes con una volatilidad muy alta.
- Agrupada (clumped) : productos intermitentes con una volatilidad constante.

De acuerdo con el marco de trabajo, los autores plantean dos procesos para abordar el problema. El primero, es la investigación empírica para la predicción de inventarios, tomando en cuenta sus características, y efectuando procesos de limpieza y validación pertinentes como el MSE (*mean squared error* o error cuadrático medio). Esta implementación se divide en 3 pasos:

- Inicialización: inicia los valores del bloque de datos con técnicas recursivas.

- Calibración: identificación de variables de suavizamiento basados en el MSE .
- Medición de rendimiento: utiliza las variables de suavizamiento para realizar la predicción.

El segundo proceso es la investigación empírica para el control de inventarios, y consiste en hacer uso del primer proceso para lograr disminuir las diferencias en la información de los inventarios, implementando los siguientes pasos:

- Diseño de simulación y criterios de rendimiento.
- Implicaciones del control de inventario para predicciones de productos lentos.
- Implicaciones del control de inventario para predicciones de productos con demanda abultada.

Es importante aclarar que, a pesar de contar con múltiples productos, sólo fue utilizado un grupo (piezas de motores) que tenía tendencia a comportarse de manera estable.

DeHoratius (2008), plantea que la falta de precisión entre los sistemas de inventarios y las bodegas de almacenamiento genera costos adicionales de operación, al sobre almacenar productos que se encontraban en existencia, y de igual manera, ocasionar perdidas de oportunidades de venta al no visibilizar productos que se encontraban disponibles en inventario. En el artículo “Manejo de inventarios de venta minoristas cuando los registros de información son imprecisos” (*Retail inventory management when records are inaccurate*) publicado por Nicole DeHoratius, Adam Mersereau y Linus Schrage en la revista Manufacturing & Service Operations Management, los autores identifican tres formas diferentes de corregir este tipo de falencias:

- Prevención: mejorar los procesos para la reducción o eliminación de las causas principales de discrepancias.

- Corrección: identificar y corregir los registros con diferencias
- Integración: utilizar herramientas lo suficientemente robustas para la toma de decisiones y planeación de los inventarios, que permitan reconocer la presencia de registros incorrectos.

En esta investigación, los autores se enfocan en el proceso de integración utilizando el “registro de inventario Bayesiano”, el cual definen cómo la capacidad de intuir la necesidad de validar la veracidad del inventario de un producto a través de un análisis Bayesiano de las ventas, a partir de la actualización constante de la distribución de probabilidad del inventario sujeta a disparadores estocásticos de discrepancias IRI (*Inventory Record Inaccuracy*). Basándose en los registros de inventarios Bayesianos, plantean implementar un conjunto de tareas para realizar la restauración de los registros identificados. La primera tarea consiste en definir políticas óptimas de restauración a través de programación dinámica, apoyándose en POMDP (siglas en inglés del Proceso de Decisión de Markov Parcialmente Observable), el cual asume el inventario físico como desconocido. La segunda tarea es la auditoría, la cual consiste en identificar un producto en particular basándose en los registros Bayesianos para realizar la conciliación de manera manual. La tercera tarea es la estimación de la demanda visible e invisible, en la que el algoritmo Bayesiano asume que se conoce la demanda diaria, y utiliza una distribución de demanda binomial negativa para la estimación.

Los autores concluyen que el algoritmo permite trabajar con los registros imprecisos, y brindar un apoyo importante tanto en las políticas como en la auditoría de los productos.

Ochella et al. (2021), combina técnicas de minería de datos y *k-Means clustering*, con una herramienta de mantenimiento RCM (*Reliability-Centered Maintenance* o mantenimiento basado centrado en confiabilidad) llamada curva de falla potencial P-F (Potential-Failure). En el artículo “Adopción de machine learning y curvas de

monitoreo de condiciones de falla potencial (P-F) para determinar y priorizar activos de alto valor para la extensión de su vida útil" (*Adopting machine learning and condition monitoring P-F curves in determining and prioritizing high-value assets for life extension*) publicado por Sunday Ochella, Mahmood Shafiee, y Chris Sansoma en la revista Expert Systems with Applications, los autores utilizan técnicas de RCM para identificar equipos esenciales a monitorear, y generar curvas P-F de las condiciones de monitoreo, para posteriormente desarrollar índices que sirvan de indicadores de la salud a partir de las condiciones de monitoreo y los datos operacionales de los equipos. Luego utilizan modelos de regresión, y *k-Means* clustering para agrupar los equipos con características similares. El *clustering* se encarga de identificar y priorizar los equipos más aptos para la extensión de su vida útil. Los resultados sirven de apoyo a los procesos de planeación, toma de decisiones, y mantenimiento de recursos.

2.2.2 Comparación

Con el fin de utilizar los trabajos seleccionados para evaluar la relevancia del proyecto, se ha preparado un cuadro comparativo utilizando el tipo de trabajo, su área de estudio, y su propósito como criterios de comparación. Los resultados han sido registrados en la Tabla 1, referenciando los trabajos del estado del arte por sus autores, e identificando el "Modelo de aprendizaje no supervisado para la priorización de inventarios cíclicos" simplemente como "Proyecto" (ver Tabla 1)

2.2.3 Conclusiones

Aunque los cinco trabajos discutidos en esta sección son todos relevantes, y tienen algunos elementos en común con el proyecto, también difieren en múltiples aspectos, y por lo tanto no ponen en tela de juicio la originalidad o la relevancia del proyecto presentado en este documento.

Tabla 1. Cuadro comparativo Estado del arte vs. Proyecto

Trabajo	Principales áreas del conocimiento	Métodos o técnicas	Propósito	Aplicable al caso de estudio
Martin (2007)	Ingeniería industrial, analítica de datos, y matemáticas	Métodos matemáticos	Cálculo de la rotación de inventarios con máxima rentabilidad	No
Kang (2005)	Ingeniería industrial, analítica de datos, y estadística	Métodos de simulación y compensación	Compensación de discrepancias de inventario	Si, pero no disminuye la necesidad de priorización
Boylan (2008)	Ingeniería industrial, e investigación de operaciones	Técnicas de predicción y categorización	Categorización de inventario en el sistema siguiendo patrones de la demanda	Complementaria al proyecto
DeHoratius (2008)	Ingeniería industrial, investigación de operaciones, y matemáticas	Modelo matemático y analítica prescriptiva	Impacto de las discrepancias de inventario en ventas al detal	Complementaria al proyecto
Ochella et al. (2021)	Ingeniería industrial y ciencia de datos	Modelo de aprendizaje no supervisado (*clustering* de datos continuos, *k-Means*)	Priorización de activos para la extensión de la vida útil	Técnicas y solución similares para un problema diferente en otra área del negocio
Proyecto	Ingeniería industrial y ciencia de datos	Modelo de aprendizaje no supervisado (*clustering de datos mixtos, k-prototypes*)	Priorización de productos en inventarios cíclicos periódicos	Caso de estudio (solución propuesta)

Fuente: Propia.

2.3 Estado de la práctica

A pesar de que este documento incluye una sección de Estado del arte, no sobra mencionar el estado de la práctica en la empresa donde se desarrolla el proyecto, ya que permite enfatizar la necesidad de encontrar nuevos caminos para apoyar el proceso con técnicas innovadoras que no han sido siquiera consideradas con anterioridad.

Durante años se han ensayado diversos enfoques con resultados mixtos, pero ninguno particularmente exitoso. Entre las estrategias utilizadas en los últimos 15 años se incluyen:

- Secuencia alfabética de productos.
- Secuencia alfabética de ubicaciones de bodega.
- Zonas de la bodega previamente asignadas y ordenadas.
- Priorización automática ABC siguiendo lineamientos corporativos de frecuencia.
- Combinaciones de las anteriores.

La única verdadera estrategia de priorización utilizada en esta larga ventana de tiempo ha sido la priorización automática ABC, donde los estándares corporativos indicaban que los productos A debían contarse doce veces al año, los B seis veces, y los C solo una – esta última frecuencia fue reducida a cero después de una actualización (i.e. omitir contar productos C). Implementarla exitosamente autorizaba a la empresa a omitir el desgastante y complejo proceso de realizar un inventario físico anual – en este caso, “éxito” está definido simplemente como cumplir con las frecuencias establecidas, y no tiene ninguna relación con la efectividad o inefectividad de la estrategia. Esta priorización se mantiene en algún grado, pero no como una “camisa de fuerza”, y se siguen buscando alternativas para lograr identificar y corregir discrepancias de manera oportuna y efectiva.

3. METODOLOGÍA

3.1 CRISP-DM

Cross-Industry Standard Process for Data Mining (**CRISP-DM**), el cual se puede traducir como proceso estándar para la minería de datos para todo tipo de industria, es considerado un método probado para llevar a cabo proyectos de datos (IBM, s.f.). Las fases de la metodología CRISP-DM son:

1. **Comprensión del negocio**: en esta fase se exploran las expectativas de la organización, se recopila información sobre la situación actual con el propósito de entender el problema, y se establecen los objetivos, el alcance, y los criterios de evaluación del proyecto. El resultado final de esta fase será la elaboración del plan del proyecto. Esta fase había sido cubierta prácticamente desde el inicio del proyecto, dada la experiencia de más de 15 años de uno de los miembros del equipo trabajando directamente con la organización y sus procesos de negocios. El entregable fue la definición del proyecto y un plan de trabajo.
2. **Comprensión de los datos**: En este paso se estudian más de cerca los datos disponibles para el proyecto. Es un paso clave para prevenir problemas inesperados en la fase de preparación de los datos. Esta fase implica tener acceso a los datos a través de un esfuerzo inicial de recopilación, para luego explorarlos con la ayuda de gráficas y tablas con el objetivo de lograr un mayor entendimiento sobre su origen, calidad, cantidad, y características principales, y así planear su transformación (si es necesario), y proceder a preparar un informe de descripción de los datos. Igual que sucedió con la comprensión del negocio, la fase de comprensión de datos se encontraba bastante adelantada desde el inicio, dada la experiencia de más de 15 años de uno de los miembros del equipo trabajando con los datos utilizados como insumos del proyecto. A pesar de esto, fue necesario realizar varias iteraciones para corregir errores en la extracción de los datos, e identificar y eliminar observaciones anómalas que

la experiencia previa con los datos no permitió prevenir ni corregir con anterioridad. El entregable de esta fase fue la información estructurada y lista para preprocesamiento en Python.

3. **Preparación de los datos**: Tradicionalmente se estima que esta fase puede tomar entre el 50-70% del tiempo y esfuerzo de un proyecto. El objetivo final de esta fase es llevar los datos al estado requerido para la fase de modelado. Algunas actividades típicas de la fase de preparación de los datos son: la fusión de datos, la selección de muestras, la generación de nuevos atributos o características, la clasificación de datos para el modelado, la eliminación o sustitución de datos en blanco o perdidos, y la separación de los conjuntos de datos de entrenamiento y evaluación. Para el caso de estudio, esta fase requirió varias iteraciones ante el descubrimiento de errores, anomalías, e incluso decisiones cuestionables durante la etapa de preparación de los datos, que llevaron a regresar – aunque fuera parcialmente - a la etapa de comprensión de los datos, para luego desembocar nuevamente en la fase de preparación de datos, con el objeto de realizar ajustes y continuar el proceso. El entregable de esta etapa fue el conjunto de datos estructurado y listo para el modelaje.
4. **Modelado**: Un paso crítico de esta fase es la decisión sobre el tipo de modelo a utilizar. Algunos factores claves para la toma de esta decisión son el tipo de datos, el problema a resolver, los objetivos del proyecto, y los requisitos de cada tipo de modelado. Luego de definido el tipo de modelo, se procede a seleccionar las métricas de evaluación y los datos a utilizar para este propósito. La generación de modelos es un proceso iterativo, por lo cual se acostumbra a experimentar con diferentes modelos y diversas configuraciones, y a documentar sus resultados, para luego compararlos y determinar tanto el modelo final a utilizar, como la configuración de los parámetros que produce mejores resultados. Para el caso de estudio, esta fase involucró la exploración de varios algoritmos de *clustering*, aunque la experiencia y las directrices del tutor del proyecto llevaron rápidamente a

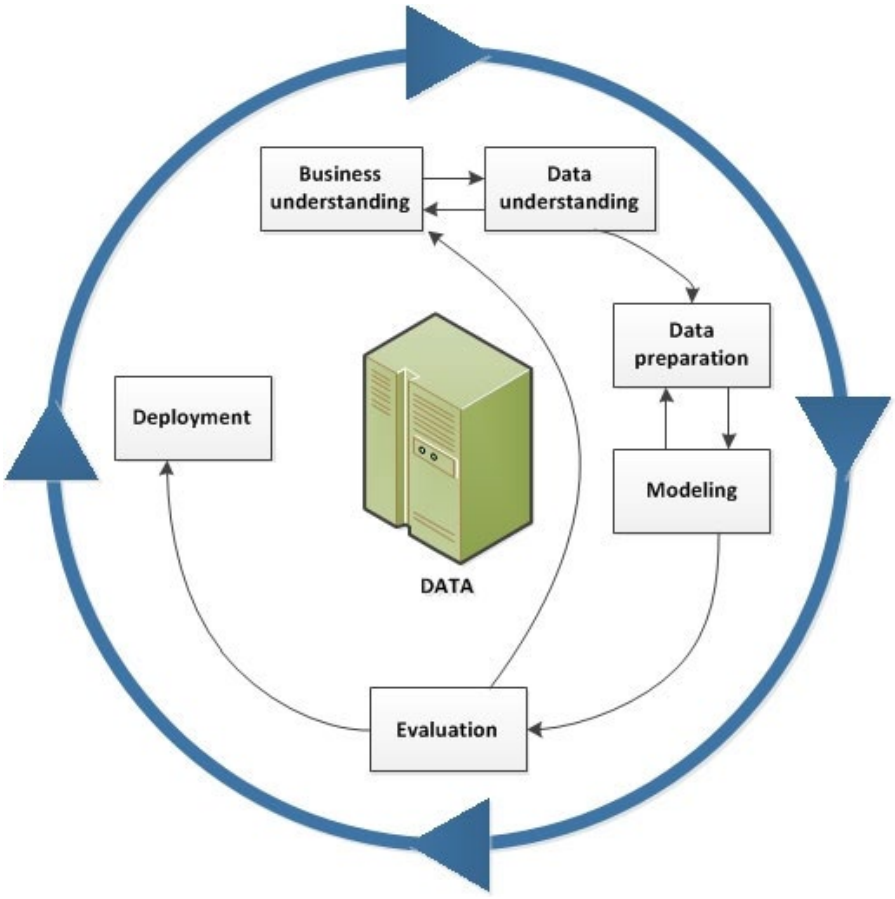
concentrar esfuerzos en la técnica de *clustering* especializada en datos mixtos y por tanto acorde con la naturaleza de los datos del proyecto. Fueron requeridas varias iteraciones mientras se afinaba el conjunto de datos, de manera que se pudiera lograr un buen ejercicio de *clustering*. La elección del algoritmo de *clustering k-prototypes*, la elección del número de *clusters* ideal para el conjunto de datos del proyecto, y las respuestas a la pregunta del negocio que los *clusters* resultantes brindan, fueron los entregables de esta fase.

5. **Evaluación**: En esta fase se procede a evaluar los resultados del proyecto de acuerdo con los criterios de evaluación establecidos en la fase inicial de compresión del negocio. Es importante dedicar tiempo para reflexionar sobre los aciertos y errores del proceso, y aprender de las experiencias propias en beneficio de futuros proyectos. Es aquí cuando se debe tomar la decisión sobre el paso siguiente a tomar: desplegar el modelo - en caso de considerarse listo – y producir un informe final, o volver a una fase previa, particularmente a la fase de comprensión del negocio, para realizar ajustes e iniciar una nueva iteración. Para el caso de estudio, esta fase implicó la preparación y ejecución de experimentos de validación que confirmaron los resultados del proyecto. Esta confirmación dio paso al análisis detallado de los resultados, así como al análisis retrospectivo del proceso del proyecto en sí, y su principal entregable es el capítulo de conclusiones que se encuentra al final de este documento.
6. **Despliegue**: Esta fase implica utilizar los nuevos conocimientos obtenidos para alcanzar mejoras en la organización donde se lleva a cabo el proyecto. El despliegue puede implicar la utilización de nuevos sistemas, la integración de nuevas soluciones o componentes con sistemas existentes, o la utilización de nuevo conocimiento en la planificación y en la toma decisiones de negocios. Las tres actividades principales de esta fase son: la planificación y control del despliegue de los resultados, la presentación de un informe final, y la revisión del proyecto. Por tratarse de un caso

académico, no hay una verdadera fase de despliegue. El completar este documento se considera el cierre del proyecto. Es importante aclarar, que existe la posibilidad de implantar el proyecto en la compañía del caso de estudio, lo que implicaría una verdadera fase de despliegue futura, ya por fuera del ámbito académico y del alcance de este documento.

Es importante mencionar, que la secuencia de las fases no es estricta. Es común saltar de una fase a otra según sea necesario (ver Figura 1). La metodología CRISP-DM es flexible y se puede personalizar fácilmente. Que fases toman mayor relevancia o consumen mayores recursos depende, en buena medida, tanto del tipo de proyecto como de la organización en que se lleva cabo (IBM, s.f.).

Figura 1. Fases de la metodología CRISP-DM.



Fuente: IBM

3.2 Cronograma del Proyecto.

Tabla 2. Cronograma del proyecto.

CRISP-DM	Actividad	Entregable	Inicio	Fin
Comprensión del negocio	Discusión y formalización del nuevo proyecto	Plan del proyecto	07-MAR-2021	09-MAR-2021
Comprensión de los datos	Extracción y refinamiento del conjunto de datos	Información estructurada y lista para preprocesamiento en Python	09-MAR-2021	21-ABR-2021
Preparación de los datos	Limpieza y preparación de los datos. Corrección de errores y anomalías (múltiples iteraciones)	Información estructurada y lista para modelaje en Python	10-MAR-2021	21-ABR-2021
Modelado	Diseño de técnica de clustering (múltiples ejecuciones y ensayos)	Implementación original del algoritmo de *clustering*	24-MAR-2021	21-ABR-2021
Modelado	Selección del número de *clusters* (k)	Modelo de *clustering k-prototypes* para priorización de inventarios cíclicos	21-ABR-2021	03-MAY-2021
Evaluación	Diseño, ejecución, y análisis de experimentos de validación	Resultados de pruebas *clusterboot* y silueta *bootstrap*	04-MAY-2021	08-MAY-2021
Despliegue	Documentación y revisión del proyecto	Informe final del proyecto	08-MAY-2021	05-JUN-2021

Fuente: Propia.

4. CONJUNTO DE DATOS INICIAL DEL PROYECTO

La primera decisión importante del proyecto fue qué datos utilizar. Se tenía a disposición 20 años de historia, la base de datos completa de la compañía, y autorización de parte de las directivas para realizar el análisis de los datos. Descartar los primeros seis años fue una decisión fácil, ya que a finales del verano del 2005 se realizó una migración a – lo que en ese momento era - una nueva versión del sistema de información ERP. El salto fue amplio y la confianza en los datos previos a la migración no era la misma a la que se tenía en los datos procesados a partir de ese momento. La clave para la decisión en cuanto a la fecha de corte de los datos estuvo en los acontecimientos claves ocurridos a finales del año 2016: el grupo empresarial al que pertenece la empresa en que se desarrolló el proyecto adquirió una nueva compañía, y esta adquisición fue asignada a la empresa como una nueva división. La nueva división trajo consigo una línea de productos completamente novedosa, con nuevos tipos de clientes, y un crecimiento en la operación de tal magnitud que obligó a un cambio de sede, a una nueva bodega del doble del tamaño de la anterior. El cambio de sede, el aumento en el volumen, y la novedad en los productos requerían de un período de transición para recuperar la estabilidad en la operación, y por esta razón se tomó la decisión de utilizar datos a partir de enero de 2018. Se renunció a un muy alto volumen de datos – 20 o al menos 15 años de historia – por asegurar la calidad, y por ende la confiabilidad de los datos utilizados. Al final de cuentas, se tomó la decisión de utilizar los últimos 3 años de historia, correspondientes al período **2018-2020**. Ya con la decisión clara respecto al período de tiempo, no fue difícil decidir de qué tablas o archivos extraer la información. Se quería analizar qué tan propensos eran los productos a presentar discrepancias de inventario entre la realidad física en bodega y los balances en el sistema, y por tanto qué tan frecuentemente requerían transacciones de ajustes para realizar correcciones. Por lo anterior, resultaba fácil seleccionar el archivo de transacciones de inventario

como la fuente principal de información, y se seleccionaron **17.007** transacciones de inventario como conjunto de datos inicial del proyecto.

A los datos de las transacciones de inventario resultantes tanto de los inventarios físicos anuales, como de los conteos cíclicos periódicos, se añadió una gran variedad de información proveniente de los archivos maestros de productos, y el porcentaje de contribución de cada uno de los productos en la última clasificación anual ABC (ver Tabla 3). Adicionalmente, en un esfuerzo por anonimizar y minimizar el grado de exposición de la compañía, se procedió a mapear los identificadores que pudieran permitir su identificación a identificadores genéricos secuenciales. Los SKU fueron mapeados a identificadores del tipo SKU1, SKU2, y así hasta el último, y los mismo se hizo con las demás variables de categorización IG, IPG, HTS, y CAT. El costo de las transacciones (COSTX) fue multiplicado por un factor secreto para mantener la privacidad de los datos.

Tabla 3. Conjunto de datos inicial (25 columnas).

Columna	Descripción
TDATE	Fecha en la que se realizó la corrección al inventario
SKU	Identificador de producto
LOC	Ubicación en la bodega compuesta por tres dimensiones (pasillo, columna y estantería)
QTY	Corrección (cantidad), menor cero = unidades faltantes, mayor a cero = unidades adicionales no contabilizadas
AISLE	Número del pasillo de la ubicación del producto
COLUMN	Columna en del pasillo en la que se encuentra el producto
SHELF	Número de la estantería en la que se encuentra el producto
TYPE	Tipo de ubicación
IG	Agrupación a la que pertenece el producto
IPG	Agrupación de precio a la que pertenece el producto
HTS	Código de importación del producto
CLASS	Indicador de inventario, 1 = producto de inventario, 2 = compra bajo orden
PLANNING	Indicador de planeación de inventario, 1 = planeado, 0 = no planeado
ABC	Clasificación ABC de inventario

Columna	Descripción
UOM	Unidad de medida del producto
CREATION	Fecha de creación del producto en el sistema
CAT	Categoría de agrupación del producto
COSTX	Costo extendido de la transacción, costo x cantidad
QTYBEFORE	Número de unidades antes de la corrección
QTYVAR	Porcentaje de variación (QTY/QTYBEFORE)
WEIGHT	Peso del producto
LENGTH	Largo del producto
HEIGHT	Altura del producto
WIDTH	Ancho del producto
CONTRPERC	Porcentaje de contribución del producto (ABC)
HEIGHT	Altura del producto
WIDTH	Ancho del producto
CONTRPERC	Porcentaje de contribución del producto (ABC)

Fuente: Propia.

5. ANÁLISIS Y PREPROCESAMIENTO DE LOS DATOS

5.1 Columnas descartadas

Después del análisis tanto desde el punto de vista del negocio, como desde el punto de vista técnico de datos, se decidió conservar 9 de las 25 columnas del conjunto de datos inicial (ver Tabla 4). A continuación, se explican las razones para la eliminación de las 16 columnas descartadas.

- Granularidad:
    - SKU: por lógica del negocio se toma la decisión de realizar la priorización al nivel de agrupación de productos (CAT, IG, HTS, ABC, UOM).
    - CREATION: fecha de creación del producto en el sistema de información, queda descartada al eliminar el nivel SKU.

- Redundancia:
    - LOC, COLUMN, SHELF: la columna AISLE (pasillo de bodega) provee suficiente información para priorización respecto a la ubicación física en bodega
    - QTY, QYBEFORE: la columna QTYVAR (variación porcentual de cantidad) provee suficiente información sobre el impacto de la transacción de ajuste de inventario en términos de cantidad o tamaño de la discrepancia.
    - CLASS: clase del producto, se integra con la variable PLANNING siguiendo la práctica del negocio – detalles en siguiente sección.
    - CONTRPERC: porcentaje de contribución, su valor está implícito en el código ABC – detalles en siguiente sección.

- Valor para el proyecto:
    - TDATE: la fecha de la transacción no resulta relevante para el caso de estudio.

- TYPE: el tipo de ubicación de bodega fue creada como una variable temporal para validar las ubicaciones, pero no aporta ningún valor después de realizada esta validación.

- Valores nulos:
  - IPG: grupo o categoría de precio muy utilizada por el área comercial del negocio, pero más de la mitad de las transacciones (8.626) tienen valores nulos (productos sin IPG en el sistema).
  - WEIGHT: peso del producto, 1.019 valores nulos (transacciones de productos sin peso registrado en el sistema).
  - HEIGHT, LENGTH, WIDTH: dimensiones del producto, 1.298 valores nulos para cada uno (transacciones de productos sin dimensiones en el sistema).

Tabla 4. Conjunto de datos final (9 columnas).

Columna	Descripción
AISLE	Número del pasillo de la ubicación del producto
IG	Agrupación a la que pertenece el producto
HTS	Código de importación del producto
PLANNING	Combinación de CLASS/PLANNING (1/0,1/1,2/0)
ABC	Clasificación ABC de inventario
UOM	Unidad de medida del producto
CAT	Categoría de agrupación del producto
COSTX	Costo extendido de la transacción, costo x cantidad
QTYBEFORE	Número de unidades antes de la corrección
QTYVAR	Porcentaje de variación (QTY/QTYBEFORE)

Fuente: Propia.

5.2 Integración de columnas

PLANNING: es común escuchar a los funcionarios de las diferentes áreas del negocio referirse a los productos en términos de su planeación de inventarios como 1/0, 1/1, y 2/0, a pesar de que la primera parte de este código se refiera a la

clase del producto (CLASS en el conjunto de datos inicial) y la segunda al método de planeación (PLANNING en el conjunto de datos inicial), y se trate de dos valores muy relacionados pero almacenados de manera independiente en el sistema de información. Por esta razón, se tomó la decisión de integrarlos en la variable PLANNING del conjunto de datos final. En la sección *Variables categóricas* (5.5.1) se explican en detalle los códigos de PLANNING resultantes de la combinación de las variables originales CLASS y PLANNING aquí descrita.

CONTRPERC: el porcentaje de contribución es el porcentaje de ventas del producto, calculado a partir de las ventas totales anuales de la compañía dentro del ejercicio de clasificación ABC realizado al final del período. Esta columna fue incluida en el conjunto de datos inicial como una variable continua que podría resultar de interés, pero después de varias iteraciones completas de comprensión de datos, preparación de datos, y modelado, se descubrió que producía un importante sesgo en el *clustering*, ya que el desbalance entre el porcentaje de contribución de los primeros cinco productos de la clasificación ABC y los demás era grande. Esto hizo necesario eliminar el porcentaje de contribución del conjunto de datos, con el objeto de evitar que dicho porcentaje se convirtiera en la gran fuerza detrás del *clustering*. Al final de cuentas, los códigos ABC son asignados según el porcentaje de contribución, por lo cual tenerlo como una variable aparte resultaba redundante.

5.3 Corrección de errores y anomalías

A pesar del amplio conocimiento de los datos, del sistema de información e incluso del negocio en sí, se presentaron observaciones duplicadas en el conjunto de datos inicial, en parte por las dificultades que implicaba armar el conjunto de datos del proyecto a partir de diferentes tablas del sistema ERP. Estas observaciones fueron eliminadas, dejando el número total de filas en **16.333**.

Igualmente, después de varias iteraciones completas de comprensión de datos, preparación de datos, y modelado, se descubrieron cinco observaciones anómalas que, por lo diferentes a todas las demás, impactaban de manera importante el resultado del *clustering*. Es curioso cómo, el mismo conocimiento de los datos, jugó una mala pasada en esta ocasión, al prevenir la validación de la existencia de una combinación que era considerada “imposible”, como lo es una variación de cantidad negativa (se encontró en la ubicación de bodega una cantidad menor a la que indicaba el sistema) acompañada de un costo positivo, y lo contrario, una variación de cantidad positiva (se encontró en la ubicación de bodega una cantidad mayor a la que indicaba el sistema) acompañada de un costo negativo. Ninguna de estas dos situaciones es posible en el mundo real: una variación negativa siempre tiene un costo negativo, y una variación positiva siempre tiene un costo positivo. Pero errores en un campo del sistema de información previo a las transacciones en cuestión las hicieron posibles en el sistema. La revisión de información estadística de los *clusters* obtenidos, cuando ya se creía tener una respuesta final a la pregunta del negocio, fue lo que permitió al equipo de trabajo identificar una anomalía y validar la existencia de una situación que se consideraba inicialmente imposible. Las 5 transacciones anómalas fueron eliminadas, dejando el número total de filas en **16.328**, y obligando a iniciar una nueva iteración prácticamente completa del ciclo CRISP-DM.

Antes de la integración de las columnas CLASS y PLANNING, se eliminaron 2 observaciones asociadas con productos con código de planeación = 9, el cual está reservado para productos que no están listos para ser lanzados al mercado, por lo que las observaciones fueron eliminadas, para un nuevo total de filas de **16.326**.

5.4 Corrección de tipo de datos e imputación de valores

Se realizaron las tareas tradicionales de corrección de tipo de datos, en particular aquellas variables categóricas que para la máquina parecen numéricas por estar

conformadas solo por dígitos. Lo anterior incluye las variables AISLE (o número de pasillo de la bodega), los códigos de CLASS y PLANNING previos a su integración en una sola columna, las fechas (TDATE y CREATION) que fueron luego eliminadas.

Las tareas de imputación de valores fueron realmente menores. Solo fue necesario hacerlo para la columna HTS (o código de importación) que tenía 10 valores nulos que fueron reemplazados por el código “No HTS”; y la columna AISLE (o número de pasillo de la bodega) que era el resultado de dividir el código de la ubicación original LOC (por *location*, su nombre en inglés) en AISLE(pasillo)-COLUMN(columna dentro del pasillo)-SHELF(estante en esa columna, dentro de ese pasillo) en sus diferentes componentes, resultando como valor nulo en aquellos casos en que la transacción correspondía a un área genérica de la bodega como podría ser por ejemplo 99REC o área de recepción de productos, para la cual no es posible descomponer su código de ubicación de bodega en tres partes. La solución fue simple, optando por el nombre completo de la ubicación como si se tratara del pasillo, en el caso de LOC = 99REC, AISLE – antes nulo – sería simplemente 99REC.

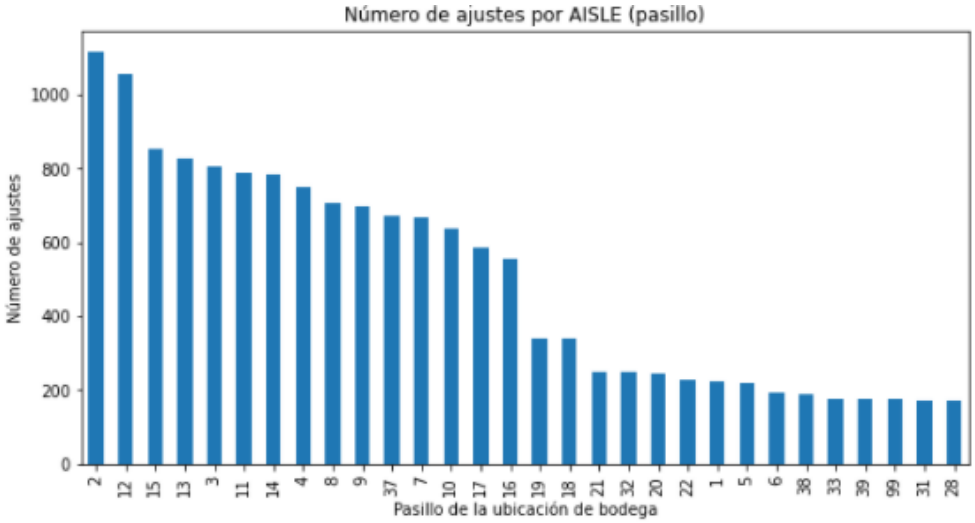
5.5 Análisis descriptivo del conjunto de datos final

5.5.1 Variables categóricas

Las siguientes son las variables categóricas del conjunto de datos (ver Tabla 5):

AISLE: número del pasillo de la bodega donde se encuentra la ubicación en la que se realizó el ajuste (ver Figura 2) .

Figura 2. Variable categórica AISLE.

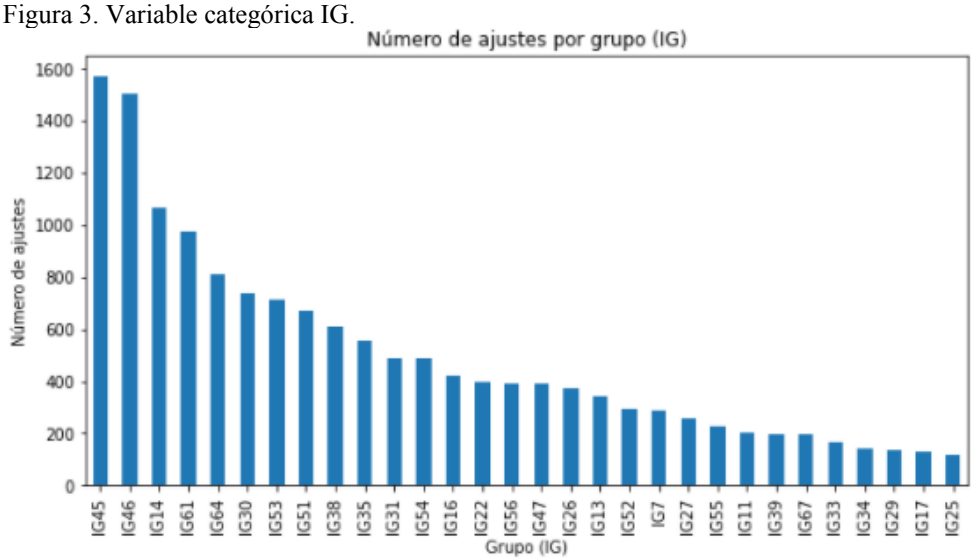


Nota: La gráfica muestra los treinta pasillos con mayor número de ajustes de un total de cuarenta y tres pasillos con transacciones en el conjunto final de datos.

Fuente: Propia

IG: grupo de productos (*Item Group* por su nombre en inglés), utilizado para reportes, precios, restricciones de ventas, cálculo de costos, y tareas contables (ver Figura 3).

Figura 3. Variable categórica IG.

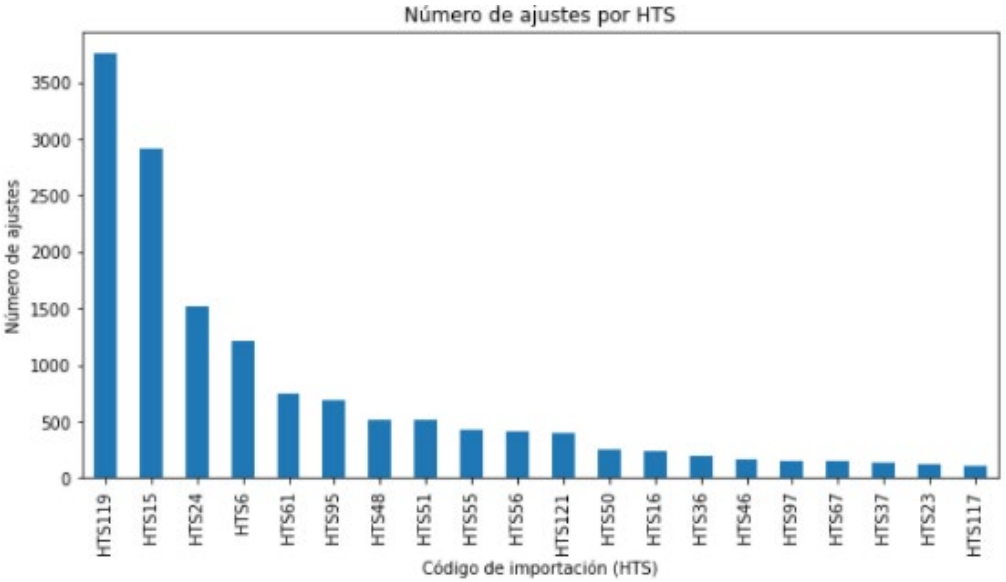


Nota: La gráfica muestra los treinta grupos con mayor número de ajustes de un total de sesenta y cinco grupos con productos en el conjunto final de datos.

Fuente: Propia

HTS: código de importación (*Harmonized Tax Schedule code* por su nombre en inglés), utilizado para reportes, y cálculo de costos (ver Figura 4).

Figura 4. Variable categórica HTS.



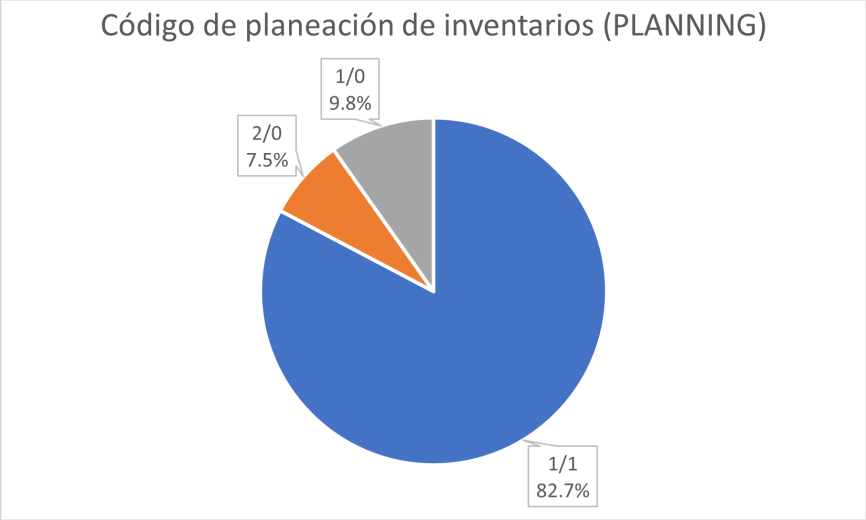
Nota: La gráfica muestra los veinte códigos HTS con mayor número de ajustes de un total de ciento treinta y dos códigos con productos en el conjunto final de datos.

Fuente: Propia

PLANNING: los códigos PLANNING están conformados por la unión del tipo de producto (conocido en el sistema de información como “clase”) y el método de planeación. El tipo de producto “1” es aquel que se almacena en inventario, mientras que el tipo “2” es reservado para aquellos productos que solo se compran cuando hay una orden de un cliente y son conocidos como BTO por su sigla en inglés (*Buy To Order*). Por tanto, los productos tipo “2” nunca son planeados y solo hay unidades físicas en inventario si el cliente canceló la orden que ocasionó su compra, o si las unidades en inventario existían antes de que el producto fuera catalogado como BTO. El método de planeación es “1” para los productos que se ordenan de proveedores después de un ejercicio de planeación que involucra el análisis de la demanda histórica del producto, y “0” para los que no se realiza

ejercicio de planeación alguno. Los productos tipo “1” con método de planeación “0” son aquellos que se encuentran en proceso de ser descontinuados, y solo se espera lograr vender las unidades restantes o en su defecto destruirlas. La regla del negocio determina entonces que los productos en que se debe enfocar la atención son aquellos con código PLANNING = 1/1, o sea tipo “1” (almacenados en inventario) con método de planeación “1” (niveles de inventarios planeados con base en su demanda histórica), ya que en ellos se concentran de manera absoluta los esfuerzos de ventas y manejo de inventarios de la compañía (ver Figura 5).

Figura 5. Variable categórica PLANNING.

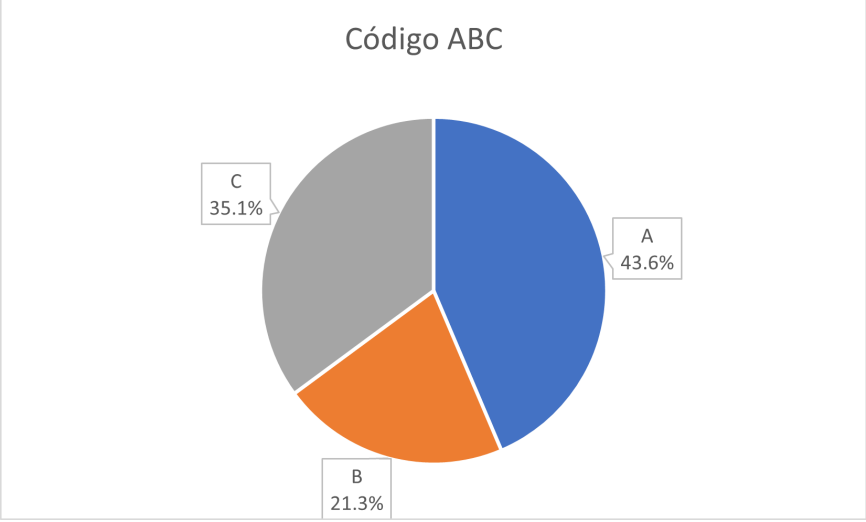


Fuente: Propia

ABC: los códigos ABC provienen del ejercicio anual de clasificación de inventarios ABC, donde los productos reciben uno de los tres códigos según su porcentaje de contribución a las ventas totales de la compañía en el período. Siendo el código A reservado para los productos de mayor contribución, B para el siguiente nivel, y C para los productos de menor contribución. Los porcentajes que determinan si un producto es A, B, o C pueden variar según el negocio, la industria, e incluso lineamientos corporativos en el caso de grupos empresariales (Vidal, 2010). En el caso particular de la compañía donde fueron recolectados los datos, el código A

es asignado a aquellos productos con una contribución del 90% a las ventas anuales, B al siguiente 7%, y C al restante 3%. Vale la pena aclarar que muchos de los productos C no realizan ninguna contribución al no haber sido vendida ni una solo unidad en el año, y son C simplemente por el hecho de permanecer activos en el archivo maestro de productos. Los productos nuevos son creados como B, y tienen la oportunidad de subir de categoría al final del período si logran un nivel de contribución a las ventas de la compañía que los haga ser considerados como A en el período siguiente, o en caso contrario, mantenerse como B, o ser degradados a C si su contribución en el período no logra siquiera el umbral requerido para mantenerse como B. Es por todo lo anterior, que los esfuerzos de ventas y manejo de inventarios se concentran principalmente en los productos A, y en segundo término en los productos B, mientras que no se presta mayor atención a los productos C.

Figura 6. Variable categórica ABC.

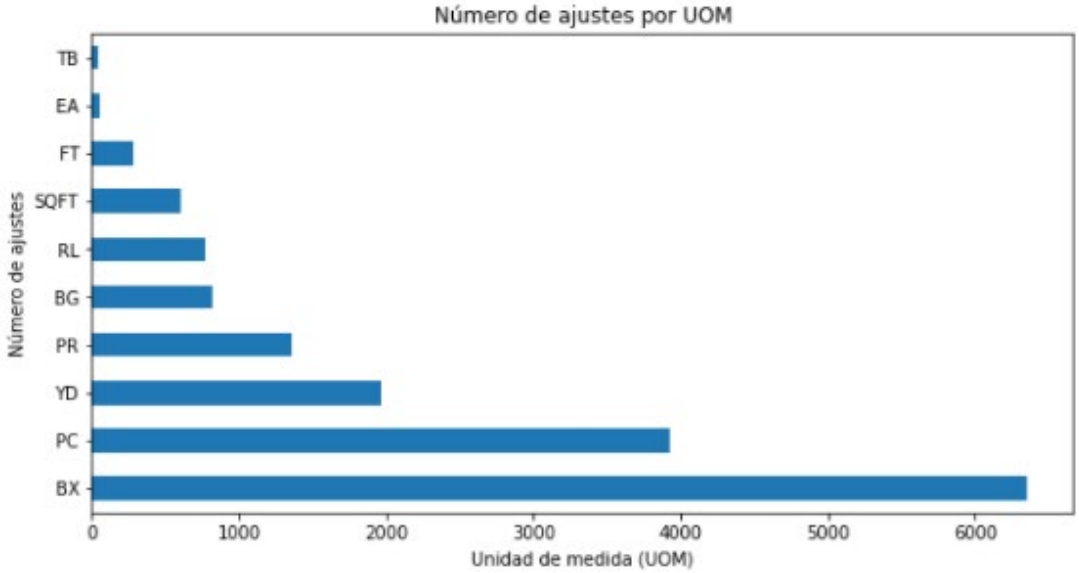


Fuente: Propia

UOM: unidad de media (*Unit of Measure* por su nombre en inglés), todas las transacciones de inventario son realizadas en la unidad de inventario, aunque los productos se pueden vender en otras unidades de venta según la preferencia de

los clientes, o comprar en otras unidades de compra según la negociación con el proveedor (ver Figura 7).

Figura 7. Variable categórica UOM.



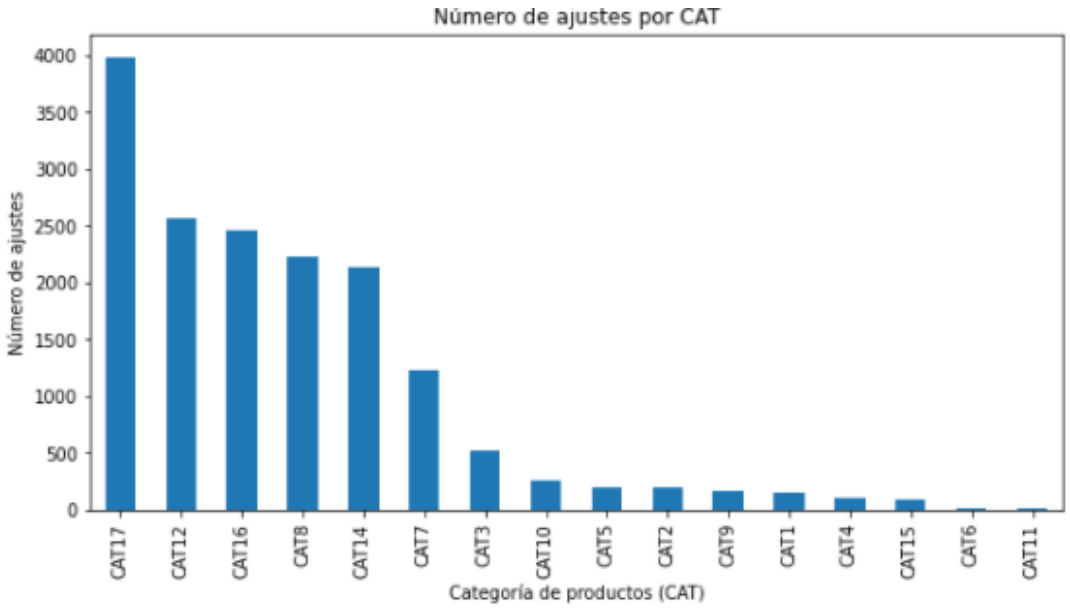
Nota: La gráfica muestra las diez unidades de medida con mayor número de ajustes de un total de catorce unidades de inventario con transacciones en el conjunto final de datos.
Fuente: Propia

CAT: categoría del producto, es la principal agrupación de productos en el sistema de información, y se utiliza para reportes, precios, restricciones de ventas, y tareas contables (ver Figura 8).

Es importante anotar que, aunque en las gráficas anteriores se pueden apreciar categorías con un mayor número de transacciones, estas no representan necesariamente los productos o categorías a priorizar, ya que podría simplemente tratarse de los productos con mayor número de transacciones en el conjunto de datos. Las figuras permiten apreciar el predominio de algunos pasillos (AISLE), y algunas agrupaciones de productos (HTS,IG, y CAT) en el total de las transacciones realizadas en el período estudiado, así como el hecho de que los ajustes tuvieron lugar principalmente en la unidad BX (cajas) – la cual es la principal unidad de inventario de los productos tradicionalmente vendidos por la

compañía, y PC (piezas) – que es la principal unidad de inventario de los productos de la línea de negocios introducida a finales del año 2016.

Figura 8. Variable categórica CAT.



Fuente: Propia

Las figuras 5 y 6 confirman lo esperado según las reglas del negocio, con un amplio predominio de los productos planeados (PLANNING = 1/1 con un 82.7%), y los productos con códigos A y B (64.9%).

La Tabla 5 muestra el número de valores diferentes para cada una de las variables estudiadas en esta sección (*unique*), la moda o valor más común (*top*), y la frecuencia de la moda (*freq*), que permiten terminar de hacerse una idea sobre las variables categóricas del conjunto de datos.

Tabla 5. Variables categóricas.

	AISLE	IG	HTS	PLANNING	ABC	UOM	CAT
count	16239	16239	16239	16239	16239	16239	16239
unique	43	65	133	3	3	14	16
top	2	IG45	HTS119	1/1	A	BX	CAT17
freq	1117	1572	3763	13428	7073	6351	3986

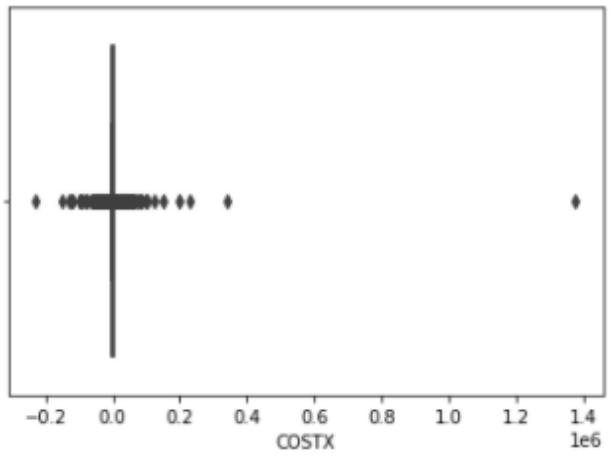
Fuente: Propia.

5.5.2 Variables continuas

Las siguientes son las variables continuas del conjunto de datos (ver Tabla 6):

COSTX: costo total o extendido de la transacción de ajuste de inventario, donde COSTX = cantidad o tamaño del ajuste multiplicado por el costo estándar unitario del producto. Un primer intento de visualizar un diagrama de cajas (*boxplot*) arrojó la gráfica presentada a continuación, donde se pueden apreciar los valores atípicos (ver Figura 9).

Figura 9. Boxplot de COSTX previo a la depuración.



Fuente: Propia.

Se tomó entonces la decisión de utilizar el estadístico *z-score*, que calcula el número de desviaciones estándar por debajo o por encima de la media (Geher, 2014), para eliminar los valores atípicos de COSTX, conservando solo aquellos valores con un valor absoluto menor a 3. La tabla 6 permite apreciar como el máximo valor de *z-score* estaba por encima de 100 (ver Tabla 6).

Tabla 6. z-score (ZcostX) para COSTX previo a la depuración.

	COSTX	ZcostX
count	16,326	16,326
mean	-66	0
std	13,326	1
min	-229,816	0
0.25	-737	0
50%	-33	0
0.75	618	0
max	1,374,824	103

Fuente: Propia.

Los siguientes son los valores de *z-score* antes de la depuración:

- Dentro de 1 desviación estándar: 15719 (96.28 %)
- Entre 1 y 2 desviaciones estándar: 411
- Entre 2 y 3 desviaciones estándar: 109
- Más de 3 desviaciones estándar: 87
- Más de 4 desviaciones estándar: 51
- Más de 10 desviaciones estándar: 7

La tabla 7 permite apreciar el nuevo número de observaciones en el conjunto de datos (**16.239**) después de eliminar aquellas por encima de 3 desviaciones estándar, de las cuales 7 se encontraban por encima de 10 desviaciones estándar. La tabla también presenta un nuevo *z-score* calculado para el conjunto de datos resultantes, con un máximo de 8.42 desviaciones estándar (ver Tabla 7). El boxplot presentado en la figura 10, y correspondiente a COSTX después de la depuración, presenta visualmente una mejor distribución de los valores en comparación con el boxplot de la misma variable presentado en la figura 9. Es

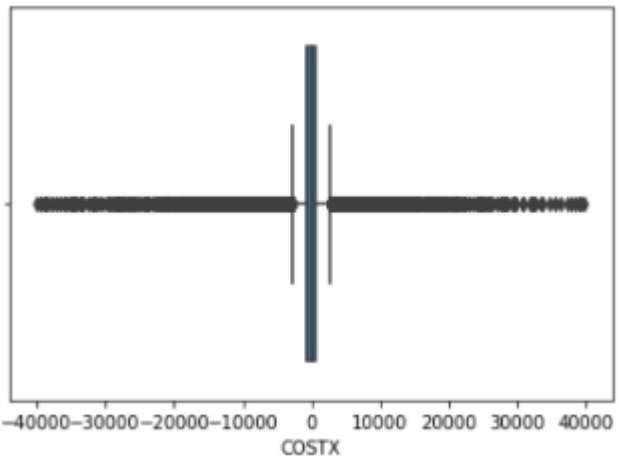
importante anotar que en la figura 10 se pueden apreciar los bigotes de la gráfica, correspondientes al mínimo y máximo excluyendo valores atípicos, gracias a la mejora significativa del rango inter-cuartil (IRQ) que le da el ancho a la caja. Esta mejora le permite pasar de prácticamente una línea en la figura 9, a una caja angosta pero bien formada en la figura 10 (ver Figura 10). El proceso de depuración podría haber continuado con un nuevo ciclo de eliminación de observaciones y cálculo de nuevos valores de z-score, pero debido al interés de trabajar con el mayor número de transacciones del período, eliminando solo aquellas que fuera estrictamente necesario para poder realizar un análisis, se tomó la decisión de no continuar el proceso de depuración y trabajar con los datos resultantes del primer esfuerzo de limpieza de valores atípicos de COSTX.

Tabla 7. z-score (ZcostX) para COSTX después de la depuración.

	COSTX	ZcostX
count	16,239	16,239
mean	-148	0
std	4,732	1
min	-40,014	0
0.25	-727	0
50%	-33	0
0.75	614	0
max	39,710	8

Fuente: Propia.

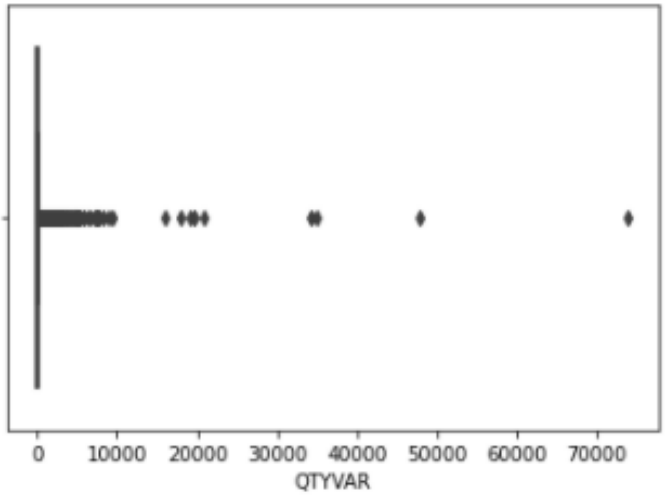
Figura 10. Boxplot de COSTX después de la depuración.



Fuente: Propia.

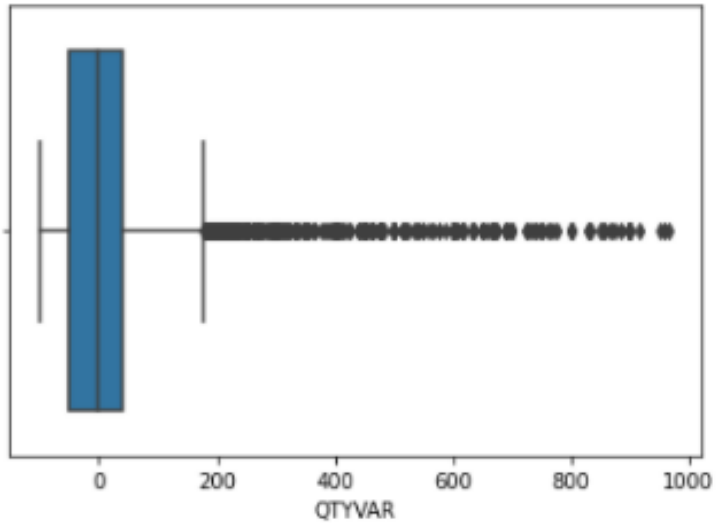
QTYVAR: variación de la cantidad como resultado de la transacción de ajuste de inventario, donde QTYVAR = cantidad o tamaño del ajuste dividido por el balance del producto en la ubicación de bodega previo al ajuste. El diagrama de cajas para QTYVAR también muestra valores atípicos extremos (ver Figuras 11A y 11B), pero se tomó la decisión de convivir con ellos, después de haber realizado la depuración basada en COSTX. Al final de cuentas, una alta variación de cantidad puede ser parte de la actividad rutinaria del negocio – especialmente en el caso de productos inventariados en pies (FT), yardas (YD), o piezas (PC) – mientras el costo de la transacción (COSTX) no resulte en valores atípicos extremos. Un buen ejemplo de alta variación QTYVAR como transacción rutinaria es el valor máximo 73.950% - a primera vista muy elevado, pero correspondiente a encontrar cinco rollos en una ubicación de inventario donde no había ninguno según los balances del sistema de información ERP. Diferencias de redondeo por conversión de unidades dejaron 2 SQFT en la ubicación – lo cual es imposible, ya que correspondería al 0.75% de un rollo estándar de 267 SQFT – y causaron un alto valor de QTYVAR al realizar el ajuste, mientras que el valor para el mismo evento hubiese sido del 100%, en caso de no encontrarse ese pequeño valor de 2 SQFT como balance del producto en el sistema (ver Figura 12).

Figura 11A. Boxplot de QTYVAR (conjunto de datos completo).



Fuente: Propia.

Figura 11B. Boxplot de QTYVAR (acercamiento).



Nota: Variaciones por debajo de 1000%

Fuente: Propia.

Figura 12. Observación correspondiente al valor máximo de QTYVAR.

AISLE	IG	HTS	PLANNING	ABC	UOM	CAT	COSTX	QTYVAR
2	IG38	HTS61	1/1	A	SQFT	CAT8	4704.755	73950

Fuente: Propia.

La tabla 8 permite apreciar un resumen estadístico de las variables continuas COSTX y QTYVAR después del proceso de depuración (ver Tabla 8).

Tabla 8. Variables continuas.

	COSTX	QTYVAR
count	16239.000000	16239.000000
mean	-148.030008	46.006304
std	4732.204184	913.386854
min	-40013.935000	-100.000000
25%	-727.162600	-50.000000
50%	-33.051600	-1.180000
75%	614.167300	46.935000
max	39709.661400	73950.000000

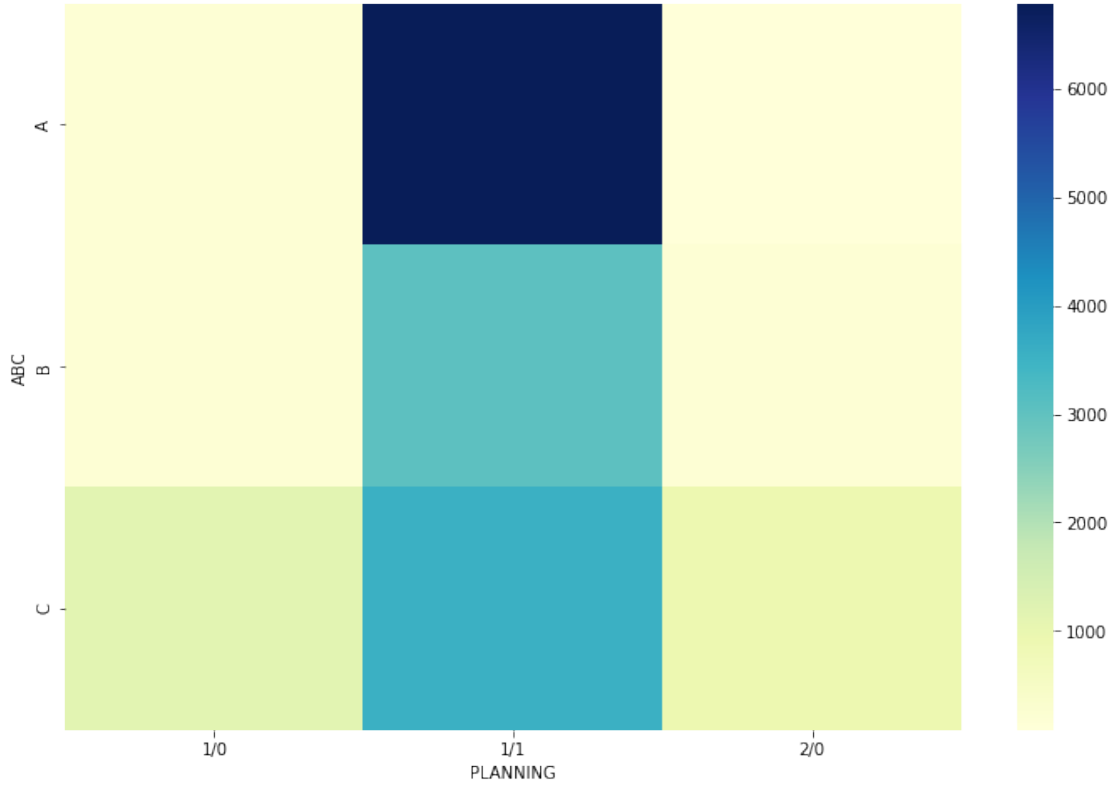
Fuente: Propia.

5.5.3 Análisis bivariado

En esta sección se presentan los resultados de realizar el análisis de los datos con dos variables de manera simultánea. Para esto, se procedió a seleccionar tres parejas en particular, con el propósito de explorar las relaciones entre ellas, y en un caso particular, de confirmar que las reglas teóricas del negocio se cumplen en la práctica, y se ven representadas en las observaciones del conjunto de datos.

La Figura 13 cruza ABC y PLANNING, y confirma las reglas del negocio una vez más, con una fuerte combinación de productos planeados de inventario con código A (ver Figura 13).

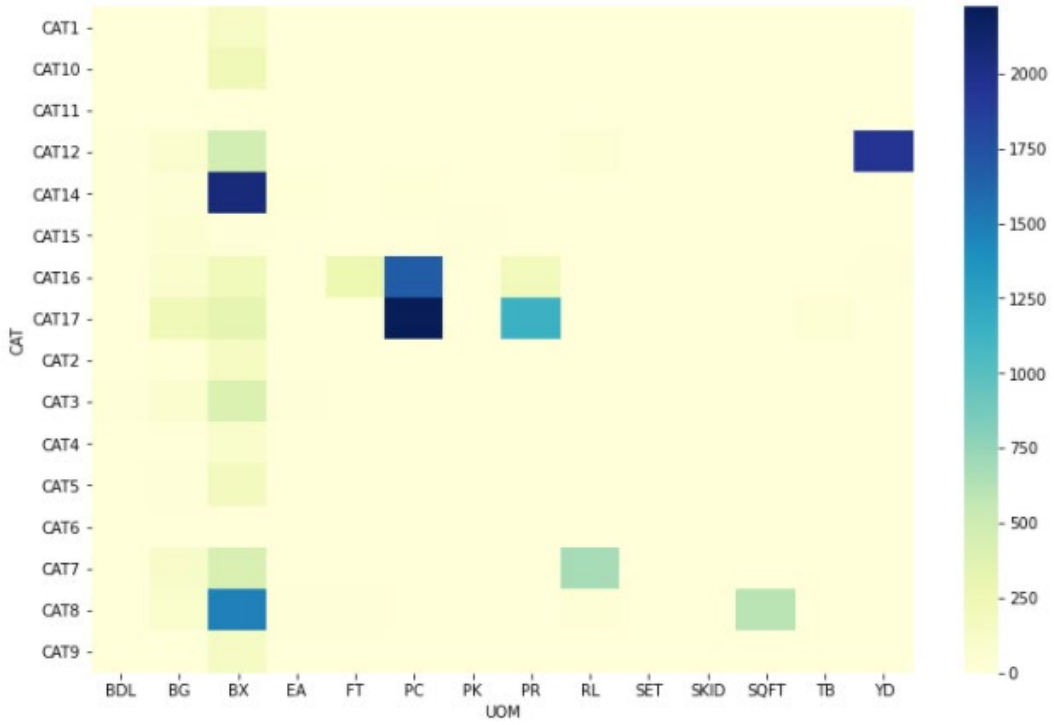
Figura 13. ABC vs PLANNING.



Fuente: Propia.

La Figura 14 cruza CAT y UOM, y también confirma la práctica del negocio, con la tradicional unidad de medida BX (caja) con presencia a través de múltiples categorías y especial notoriedad en las categorías CAT8 y CAT14; la unidad de medida PC (pieza) – tan importante en la línea de negocios introducida a finales del año 2016 – con fuerte presencia en las categorías CAT16 y CAT17; y la unidad de medida RL (rollos) – también de tradicional importancia en el negocio – con presencia fuerte en la categoría CAT12 (ver Figura 14).

Figura 14. CAT vs UOM.

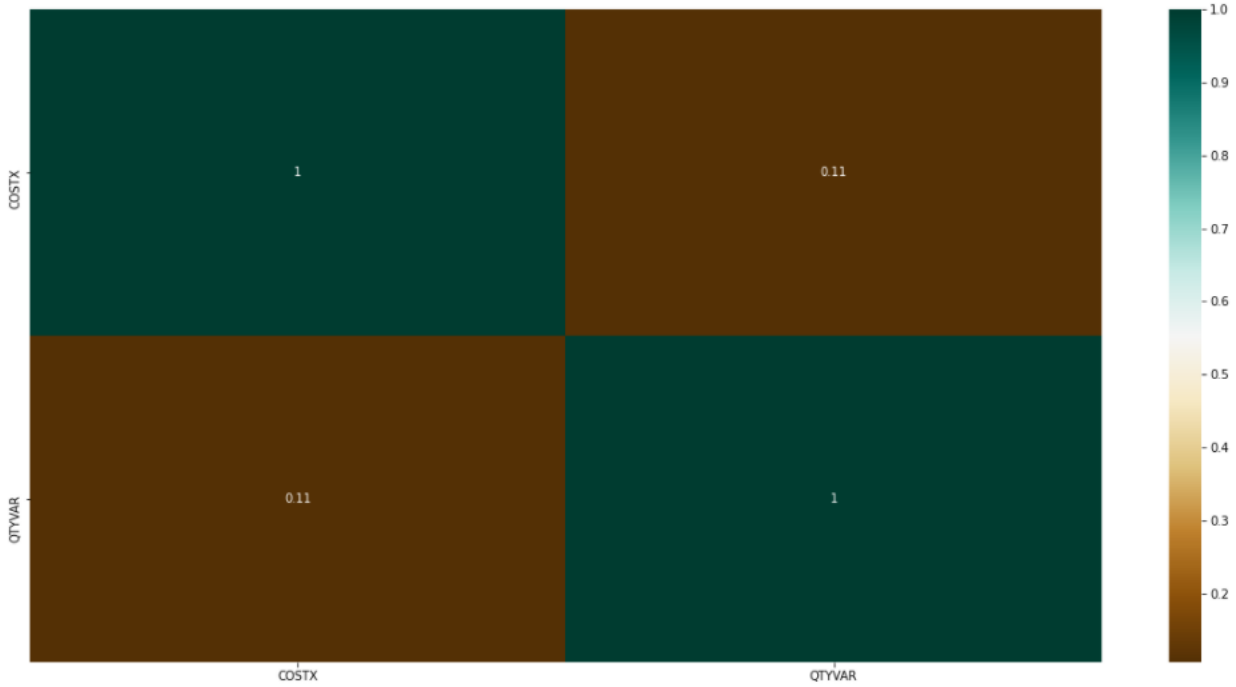


Fuente: Propia.

La Figura 15 muestra una muy baja correlación entre las variables continuas COSTX y QTYVAR, contrario a lo que se podría esperar intuitivamente, pero acorde con el negocio y el impacto conocido de las diferentes unidades de medida (UOM). Una caja (BX) de alto costo monetario (COSTX) puede variar en un pequeño porcentaje (QTYVAR) – por ejemplo 5% o 1 de 20, mientras que un producto inventariado en yardas (YD) o pies cuadrados (SQFT) puede variar de

manera notaria, en especial si hay un pequeño sobrante en la ubicación de bodega – por ejemplo, 20 SQFT “sobrantes” de un rollo de 300 yardas y se encuentran dos rollos de 300 SQFT cada uno, para un total de 600 SQFT en la ubicación de bodega y una variación del 3000% no proporcional con el costo de los dos rollos encontrados (ver Figura 15).

Figura 15. Correlación COSTX - QTYVAR.



Fuente: Propia.

6. CLUSTERING

Para seleccionar el número ideal de *clusters* (*k*) para el conjunto de datos de transacciones de ajuste de inventario de este proyecto, se procedió a ejecutar el algoritmo de *clustering k-prototypes* con valores de *k* entre 2 y 10, la inicialización de *Huang*, y un valor de peso *gamma* = 0.5. Se tomó la decisión de mantener constantes tanto la inicialización de *Huang* como el valor *gamma* = 0.5, en lugar de utilizarlos como parámetros a ajustar, y así centrar toda la atención y los esfuerzos en la elección del *k* ideal para el conjunto de datos y el análisis de los *clusters* resultantes del ejercicio. La inicialización de *Huang* fue seleccionada con el objeto de seguir la lógica y el proceso expuesto por Zhexue Huang en su artículo "*Extensiones al algoritmo k-Means para Clustering de grandes volúmenes de datos con variables categóricas*" (Huang, 1998), mientras que el valor de *gamma* fue seleccionado como un valor intermedio que atenúa un poco el peso de las siete variables categóricas sobre las dos variables numéricas, sin llegar a reducirlas al punto de extinguirlas y prácticamente convertir la ejecución del algoritmo en un *k-Means*.

Tal y como se mencionó en el marco teórico de este documento, la métrica seleccionada para esta evaluación fue el *coeficiente silueta*, la cual se distingue por valorar tanto la cohesión intra *cluster* como la separación entre *clusters*. Para el cálculo del coeficiente silueta se utilizó la función *silhouette_score* de la librería scikit-learn del lenguaje de programación *Python* (Pedregosa et al., 2021). Es importante mencionar que, aunque esta función permite seleccionar la métrica para calcular las distancias entre los diferentes puntos, ninguna de las métricas disponibles se ajustaba perfectamente a la ecuación propuesta por Huang en su artículo, por lo que fue necesario escribir un algoritmo que realizara el cálculo completo de la matriz de distancias del conjunto de datos.

La ecuación propuesta por Huang calcula la disimilitud o qué tan diferente es un punto en un conjunto de datos mixtos de otro punto del mismo conjunto. La ecuación tiene dos términos: el primero es la sumatoria cuadrada de las distancias euclidianas entre las variables numéricas; mientras que el segundo término corresponde a la sumatoria de las diferencias entre las variables categóricas, multiplicada por un factor de peso gamma que pretende equilibrar la ecuación, evitando favorecer a cualquiera de los dos componentes (ver Ecuación 3 en la página 24 de este documento).

El algoritmo desarrollado para este cálculo genera una matriz de distancias *N x N*, donde *N* es el número de filas o número total de transacciones de ajustes de inventario en el conjunto de datos de este proyecto. Cada celda [*i, j*] de la matriz representa la distancia desde el punto *i* (o fila *i* del conjunto de datos) al punto *j* (o fila *j* del mismo conjunto de datos) utilizando la ecuación propuesta por Huang. Es importante anotar que los cálculos matemáticos se limitan a poco menos de la mitad de la matriz, ya que la distancia del punto *i* al punto *j*, que correspondería a la celda [*i, j*] de la matriz, es igual a la distancia del punto *j* al punto *i*, que correspondería a la celda [*j, i*] de la matriz. En otras palabras, el orden de los puntos no altera el cálculo de las distancias, por lo que al calcular las distancias de las celdas por encima de la diagonal de la matriz se obtienen al mismo tiempo las distancias para las celdas por debajo de la diagonal. Luego se procede a rellenar la mitad de la matriz pendiente de cálculo con las distancias ya calculadas para las celdas “espejo” correspondientes, donde “espejo” simplemente implica que cada celda [*i, j*] = celda [*j, i*]. La diagonal de la matriz no requiere cálculo alguno, ya que las celdas [*i, j*] donde *i* = *j* tienen un valor de cero ya que la distancia de un punto (o fila del conjunto de datos) a si mismo es simplemente cero por tratarse del mismo punto (o fila del conjunto de datos), y ese es el valor de inicialización de la matriz de distancias previo a calculo alguno.

El cálculo del coeficiente silueta con la función *silhouette_score* de la librería *scikit-learn* se realiza con la opción de métrica = “precalculada” (*precomputed* por su nombre en inglés) para el cálculo de las distancias, por lo cual es necesario pasarle como parámetros la matriz de distancias precalculada con el algoritmo escrito siguiendo la fórmula de Huang, y las etiquetas de membresía de *cluster* (*cluster ID*) resultantes de la ejecución del algoritmo *k-protoypes* propuesto por Huang en su artículo (Huang, 1998). Tal y como se mencionó en el capítulo de metodología de este documento, fueron requeridas varias iteraciones mientras se afinaba el conjunto de datos, de manera que se pudiera lograr un buen ejercicio de *clustering*.

7. DISEÑO DE LOS EXPERIMENTOS DE VALIDACIÓN

Para validar los resultados obtenidos, se diseñaron dos experimentos de validación basados en técnicas de muestreo con reemplazo o *bootstraps*.

El primer experimento de validación fue utilizado para confirmar el número de *clusters* o agrupaciones en los cuales segmentar el conjunto de datos de transacciones de ajuste de inventario del proyecto, y se le llamó "silueta *bootstrap*" por tratarse precisamente de una validación del coeficiente silueta con *bootstraps*; mientras que el segundo experimento tuvo como objetivo validar qué tan estables y estructurantes eran los clusters resultantes del ejercicio, y fue denominado "*clusterboot*" en referencia a la función del mismo nombre, implementada en el paquete *fpc* de *R* (RDocumentation, 2020), en la cual está inspirado el experimento. La función *clusterboot* de *R* está basada en el artículo "Evaluación de estabilidad de clusters" (*Cluster-wise assessment of cluster stability*) publicado por Christian Hennig en la revista "Estadística Computacional y Análisis de Datos" (Hennig, 2007). Ambos experimentos fueron ejecutados con cien (100) *bootstraps* del mismo tamaño del conjunto de datos original. El hecho de que se tratara de muestreo con reemplazo y que los *bootstraps* tuvieran el mismo tamaño del conjunto de datos original implica una alta probabilidad de contar con observaciones repetidas en cada uno de los *bootstraps* (Efron, 1997).

Los experimentos fueron desarrollados en el lenguaje de programación *Python*, y ejecutados con idénticos resultados, tanto en la nube utilizando *Google Colab Pro* con las opciones de "Alta capacidad de RAM" y Acelerador de hardware = GPU; como localmente en un computador con la siguiente configuración:

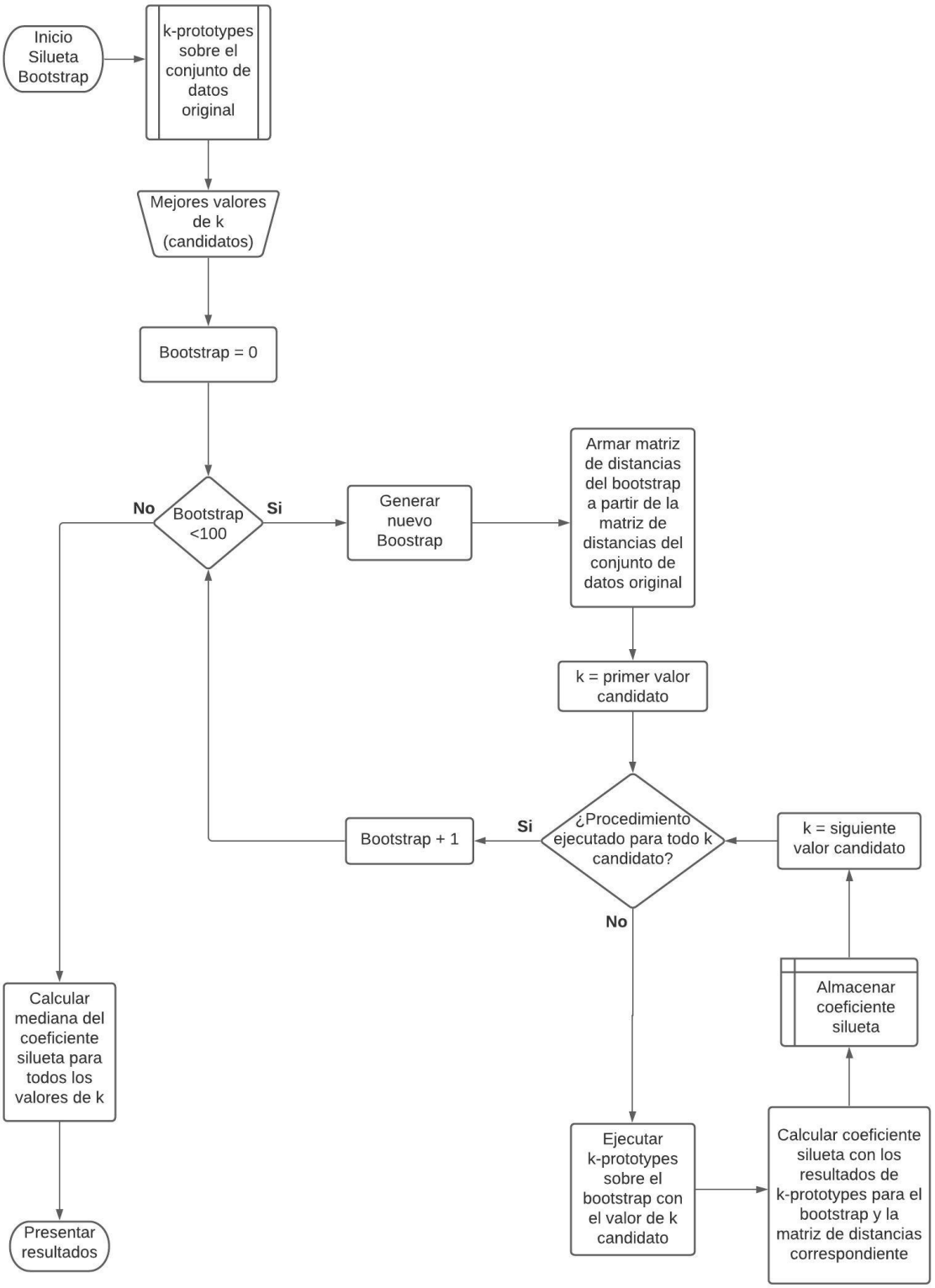
- Intel Core i7-8700 CPU @ 3.20 GHz 3.19 GHz.
- Memoria RAM: 32 GB.
- Sistema operativo de 64 bits, procesador x64.
- GPU: NVIDIA GeForce GTX 1060 6GB.

El detalle de los experimentos se presenta a continuación, y sus resultados en la sección 8 de este documento.

7.1 Silueta bootstrap

El primer paso consiste en elegir los valores de *k* con un mayor coeficiente silueta al ejecutar el algoritmo *k-prototypes* con el conjunto de datos original. Es importante aclarar que, ningún *k* es automáticamente considerado un buen candidato de hacer parte del experimento de validación, simplemente por haber ocupado un buen puesto al comparar su coeficiente silueta con los demás. No basta obtener el primer lugar, ni mucho menos el segundo o el tercero. El coeficiente silueta obtenido de ejecutar el *clustering* con ese valor de *k*, debe ser idealmente superior o al menos cercano a 0.5, para que ese valor de *k* sea considerado parte del experimento de validación. El siguiente paso implica ejecutar el algoritmo de *clustering k-protoypes* para 100 *bootstraps* del mismo tamaño del conjunto de datos original, con cada uno estos valores de *k* preseleccionados, y calcular el coeficiente silueta para cada una de las ejecuciones del algoritmo. Lo anterior implica que, si se seleccionan dos valores de *k* como los mejores según la ejecución del algoritmo *k-prototypes* con el conjunto de datos original, el algoritmo *k-prototypes* será ejecutado en doscientas ocasiones (dos veces para cada *bootstrap*, o sea, una vez por cada valor de *k* para cada *bootstrap*), y el coeficiente silueta será calculado el mismo número de veces. La mediana - valor medio que divide el conjunto de datos en dos partes iguales donde el 50% de los valores son mayores o igual a la mediana, y el 50% menores o igual a ella (Heumann, 2016) - fue la métrica elegida para calcular el coeficiente silueta final para cada valor de *k* preseleccionado. El resultado final, consiste simplemente en calcular la mediana entre los cien coeficientes silueta obtenidos para cada valor de *k* (uno por *bootstrap*). El valor de *k* con un mayor valor de la mediana al final del experimento es utilizado para validar el número de *clusters* (*k*) seleccionado (ver Figura 16).

Figura 16. Diagrama de flujo Silueta bootstrap



Fuente: Propia.

7.2 Clusterboot

Para evaluar la calidad del *clustering*, y confirmar qué tan estables y estructurantes son los *clusters* resultantes de ejecutar el algoritmo *k-prototype*s con el *k* seleccionado, se realizó otro experimento de validación con 100 *bootstraps* del mismo tamaño del conjunto de datos original. Este experimento compara los *clusters* resultantes de la ejecución del algoritmo *k-prototype*s con el conjunto de datos original y el valor *k* seleccionado como el número ideal de *clusters* del ejercicio, con los *clusters* resultantes de ejecutar el mismo algoritmo – con el mismo valor de *k* – sobre los *bootstraps*.

El procedimiento pretende replicar la función *clusterboot* del paquete *fpc* de *R* (RDocumentation, 2020). Al no encontrarse esta funcionalidad disponible en *Python*, fue necesario desarrollar un algoritmo escrito desde cero que la replicara. Para poder comprender el experimento, es fundamental entender primero como funciona la métrica utilizada para calcular la similitud de un *cluster* con otro, de acuerdo con el artículo “Evaluación de estabilidad de clusters” de Christian Hennig (Hennig, 2007): el coeficiente de *Jaccard*. El coeficiente de *Jaccard* mide la similitud entre dos conjuntos (*clusters* en este caso), y es por esta razón, que el algoritmo desarrollado utiliza operaciones de conjuntos (unión e intersección) para evaluar la similitud entre los *clusters* originales y los *clusters* resultantes del ejecutar el algoritmo *k-prototypes* para cada *bootstrap*. Si se asume por motivos de ilustración, que el *k* seleccionado como ideal para el conjunto de datos original es 2 (*k = 2*), se tendrían dos *clusters* originales *C0* y *C1*, y CB_i0 y CB_i1 serían los clusters resultantes de ejecutar el algoritmo *k-prototypes* con *k = 2* para el *bootstrap i*. El algoritmo compararía a *C0* tanto con CB_i0 como con CB_i1, calculando el coeficiente de *Jaccard* para cada pareja, y eligiendo el mayor valor como el coeficiente de *Jaccard* correspondiente al *cluster 0* original (*C0*) en comparación con el bootstrap *i*. Es importante anotar que, el *cluster 0* original (*C0*) no tiene por qué necesariamente coincidir con el *cluster 0* del *bootstrap i* (CB_i0), ya

que es posible que el *cluster 1* del *bootstrap i* (*CBi1*) sea el más similar al *cluster 0* original. Luego el ejercicio se repetiría para el *cluster 1* original (*C1*), con el fin de encontrar el *cluster* correspondiente en el *bootstrap i*, y de esta manera lograr calcular su coeficiente de *Jaccard* en comparación con el *bootstrap i*. Este procedimiento se puede generalizar sin ningún problema a cualquier número *k* de *clusters*.

El coeficiente de *Jaccard* se calcula como la cardinalidad de la intersección de los dos clusters (o subconjuntos) dividida por la cardinalidad de su unión, donde la cardinalidad es el número de elementos del conjunto (ver Ecuación 3). Lo cual se podría expresar en términos coloquiales como "qué porcentaje del total de los elementos son comunes entre los dos clusters".

Ecuación 4. Cálculo del coeficiente *Jaccard* para dos clusters o subconjuntos C y D de los n puntos del conjunto X.

$$\gamma(C, D) = \frac{|C \cap D|}{|C \cup D|}, \quad C, D \in X_n$$

Fuente: Cluster-wise assessment of cluster stability.

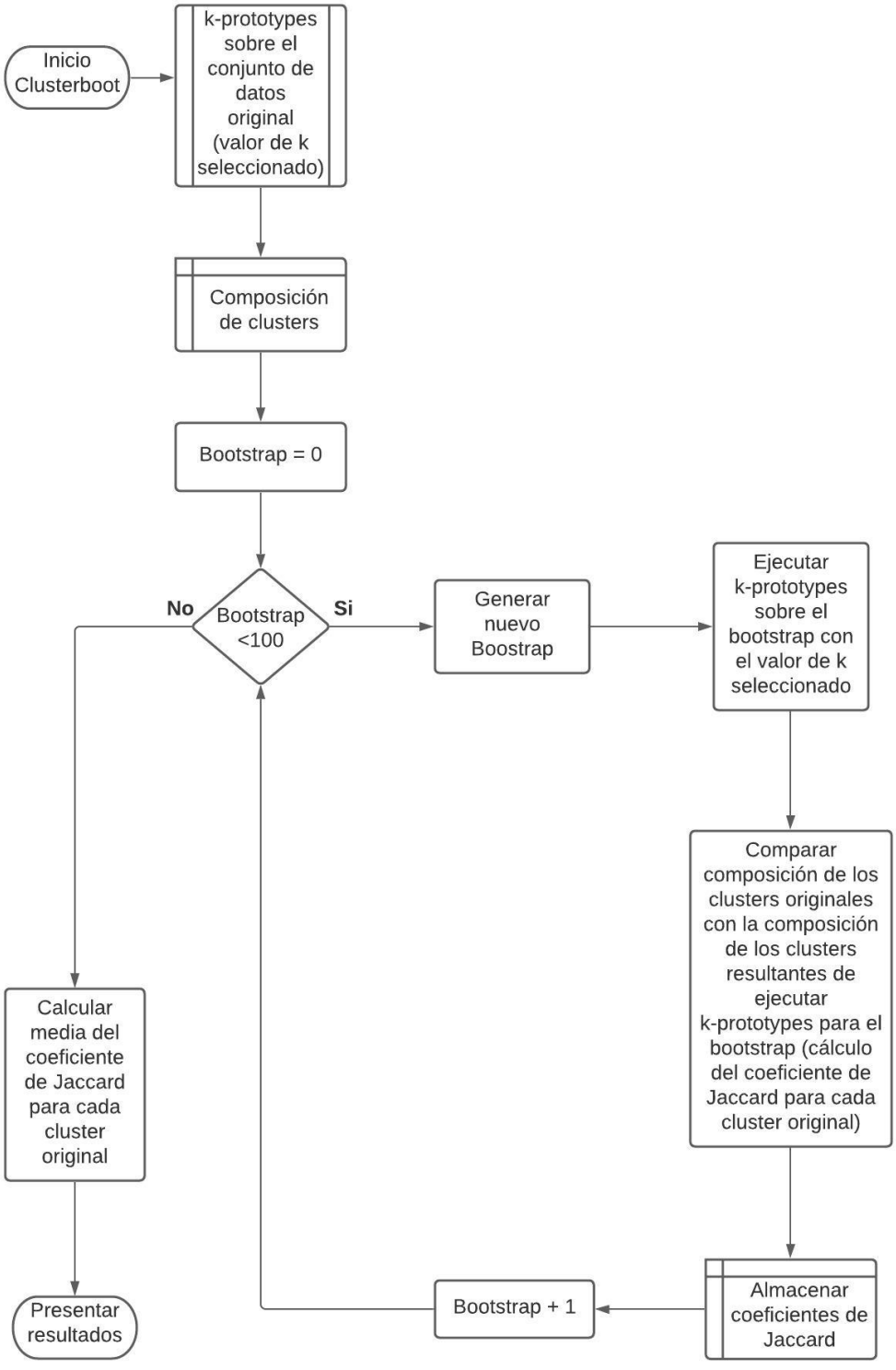
El algoritmo desarrollado se encarga de: 1) la generación de los 100 *bootstraps;* 2) de la ejecución del algoritmo de *clustering k-prototypes* para cada uno de los *bootstraps* con el valor de *k* previamente seleccionado para el conjunto de datos original; y 3) el cálculo del coeficiente *Jaccard* para cada *cluster* original para cada *bootstrap*. Después de realizar todas las comparaciones entre *clusters* y, por lo tanto, de calcular el coeficiente de *Jaccard* para cada *cluster original* en cada *bootstrap*, se procede a calcular el promedio del coeficiente de *Jaccard* de cada *cluster* original en el experimento, lo cual es simplemente el resultado de promediar los 100 valores de *Jaccard* obtenidos para cada *cluster* original.

Adicional al cálculo del promedio para cada *cluster* original, el experimento también incluye el conteo del número de instancias (número de *bootstraps*) en lo

que cada *cluster* original obtiene un coeficiente de *Jaccard* por debajo de 0.5 – en ese caso se dice que el *cluster* se disuelve, al no encontrar en ese *bootstrap* un *cluster* lo suficientemente similar; y el número de instancias en los que cada *cluster* original obtiene un coeficiente de *Jaccard* por igual o mayor a 0.75 – en ese caso se dice que el *cluster* se recupera, al encontrar en ese *bootstrap* un *cluster* muy parecido a sí mismo (ver Figura 17).

Es importante aclarar que, dada la naturaleza del conjunto de datos de este proyecto, y la experiencia obtenida a través de las diversas iteraciones de las diferentes etapas de la metodología CRISP-DM con el conjunto de datos, se tiene como objetivo un coeficiente de *Jaccard* alrededor de 0.6 (60%) con una baja desviación estándar, como parámetro para considerar que el experimento haya sido un éxito.

Figura 17. Diagrama de flujo Clusterboot



Fuente: Propia.

8. RESULTADOS OBTENIDOS

8.1. Consideraciones desde el punto de vista del negocio

Antes de discutir los resultados del proyecto, es importante recordar la naturaleza de los datos analizados, ya que ésta tiene importantes implicaciones en el análisis final desde el punto de vista del negocio.

Los datos corresponden a transacciones históricas de ajustes de inventario, es decir, correcciones tanto positivas como negativas, de los balances de los productos en las diferentes ubicaciones de bodega. El algoritmo de clustering busca patrones en estas transacciones de ajuste, para así lograr agruparlas en *clusters* compuestos por transacciones lo más parecidas posible entre si – intra *cluster* – y lo más diferente posible de las transacciones de los otros *clusters* – inter *cluster*. Los *clusters* resultantes están compuestos por transacciones de ajuste de inventario, y no de productos, que es lo que el proyecto buscar priorizar a partir de la identificación de aquellos más propensos a requerir ajustes. Lo anterior no es un problema ni mucho menos una contradicción. Al final de cuentas, cada transacción de ajuste está asociada con un producto, que a su vez pertenece a grupos o categorías de productos en el sistema que pueden ser sujetas a priorización. Estos grupos o categorías de productos son los que permiten la priorización, según la membresía del *cluster* seleccionado como aquel que incluye las transacciones en las cuales se debe enfocar la atención.

Es importante resaltar que resulta absolutamente normal que productos con transacciones en el *cluster* seleccionado para priorización, también tengan transacciones en otros *clusters*. La razón es muy simple: es normal que los productos requieran ajustes, si no lo fuera, en lugar de buscar encontrar patrones en las transacciones de ajuste, simplemente se seleccionarían todos los productos que requirieron algún ajuste en la ventana de tiempo seleccionada. Pero lo

anterior resultaría en un muy alto número de productos, e incluiría productos que no requirieron mayores ajustes ni en cantidad, ni en costo, ni en frecuencia. Por tanto, al ser normal que un producto requiera ajustes de inventario, no se hacen relevantes sus ajustes menores. Esos ajustes menores quedarán muy probablemente agrupados con otros ajustes menores, ya que es muy posible que el peso de una sola variable categórica – o de un grupo pequeño de ellas –no sea una razón suficientemente fuerte para que dicha transacción sea considerada miembro del *cluster* que, por su composición, fue seleccionado para priorización. Es por esto, que la priorización se enfoca completamente en la composición del *cluster* seleccionado, e ignora de manera total y absoluta la composición o membresía de cualquier otro *cluster*.

8.2. k-prototypes

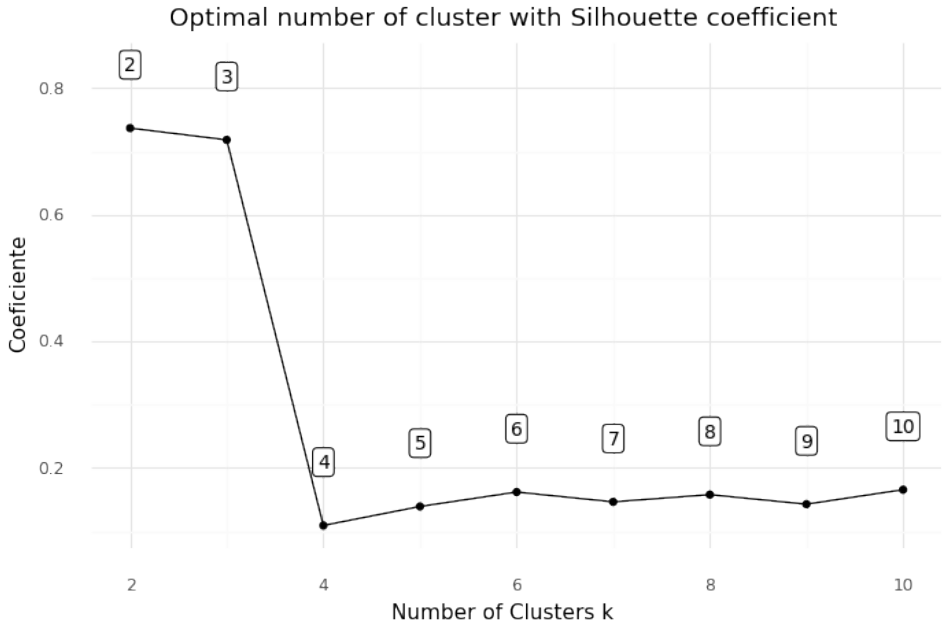
El mayor valor del coeficiente silueta para la ejecución del algoritmo *k-prototypes* con el conjunto de datos del proyecto fue para ***k = 2*** con **0.736858** seguido de cerca por *k = 3* con un coeficiente silueta de 0.718282, y con *k = 4* en un lejano tercer lugar con un muy bajo coeficiente silueta de 0.109149. La figura 18 confirma visualmente que la mejor partición está en *k* = 2 o *k* = 3.(ver Figura 18)

8.3. Silueta bootstrap

La elección del número de *clusters k = 2* fue validada ejecutando el algoritmo *k-protoypes* con *k = 2 y k = 3* para cien *bootstraps*, tal y como se describió en la sección de validación. Los resultados de la validación confirmaron ampliamente la superioridad del *clustering* del conjunto de datos con *k = 2* sobre el mismo proceso con *k = 3* (ver Tabla 9). La mediana del coeficiente silueta para *k* = 2 fue **0.75527**, incluso mayor al valor del coeficiente para el conjunto de datos original (0.736858). En cambio, la mediana fue muy baja, incluso inferior a 0.1 (exactamente 0.097002) y por lo tanto muy cercana a cero para *k* = 3. La Tabla 9 presenta tanto la

mediana, como la media y la desviación estándar del cálculo del coeficiente silueta para los 100 *bootstraps* (ver Tabla 9). No se planeaba utilizar la media, ya que es claro que algunos *bootstraps* pueden no ajustarse muy bien a los *k* seleccionados y por lo tanto castigar fuertemente el promedio, pero vale la pena destacar que, a pesar de esto, *k* = 2 obtuvo un promedio de 0.734424 muy cercano al valor del coeficiente obtenido para el conjunto de datos original. Los resultados avalan claramente la elección de *k* = 2 como el número de *clusters* ideal para el conjunto de transacciones de ajuste de inventario de este proyecto.

Figura 18. Número óptimo de *clusters* para el conjunto de datos.



Fuente: Propia.

Tabla 9. Resultados del experimento de validación *Silueta bootstrap.*

	k = 2	k = 3
count	100	100
median	0.755270	0.097002
mean	0.734424	0.290316
std	0.179109	0.314525

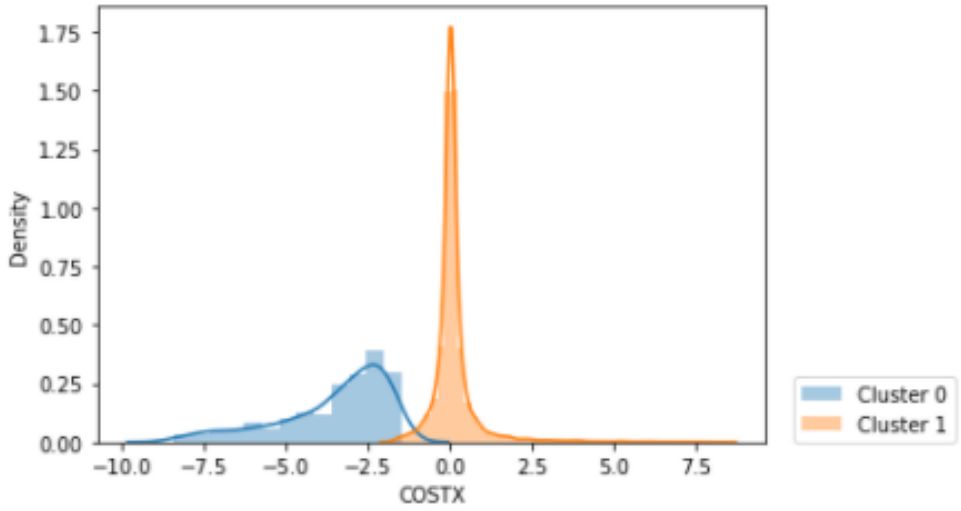
Fuente: Propia.

En el caso de ***k* = 2**, el primer *cluster* o *cluster 0* incluye 517 observaciones, y fue seleccionado como el *cluster* a priorizar. Mientras que el segundo *cluster* o *cluster 1* incluye 15.722 observaciones, y es considerado el *cluster* a ignorar por encontrarse compuesto por el grueso de las transacciones de ajustes de inventario.

8.4. Variables continuas

COSTX: la figura 19 presenta la distribución de probabilidad de los dos *clusters* respecto a la columna COSTX (ver Figura 19). La gráfica permite apreciar cómo el *cluster* seleccionado para priorización (*cluster 0*) incluye el grueso de los ajustes negativos, incluyendo la totalidad de sus valores extremos. Es importante recalcar que los ajustes extremos son principalmente negativos, con muy pocas transacciones con ajustes positivos importantes. Las transacciones de ajuste de inventario con valores negativos extremos son las más importantes para el negocio por que representan reducciones drásticas del nivel de inventario, o sea ubicaciones de bodega donde el alto costo de la diferencia entre el nivel de inventario que se creía tener en el sistema y lo que se encontró físicamente en ella implica una alta perdida para la compañía.

Figura 19. Distribución de probabilidad de COSTX en los *clusters*.

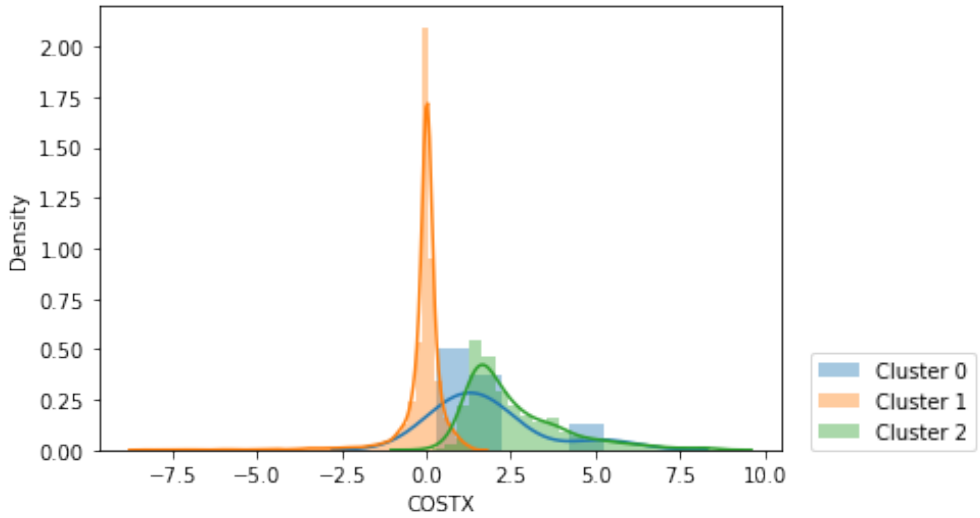


Fuente: Propia.

A pesar de no haber sido el número de *clusters* elegido, vale la pena revisar los resultados para ***k* = 3**, ya que esta alternativa obtuvo el segundo mejor coeficiente silueta entre los valores de *k* evaluados. En este caso, el *cluster* que fue seleccionado para priorización bajo *k* = 2 (el *cluster 0*) no logra mantenerse, ya que sus 517 elementos terminan siendo parte del *cluster 1*, que con 15.528 incluye el grueso de los datos para *k* = 3. El *cluster 0* incluye solo 8 transacciones de costo positivo y las mayores variaciones de cantidad, y el *cluster 1* está conformado por 703 transacciones de alto costo positivo. (ver Figura 20). Vale la pena aclarar que tanto la figura 19 correspondiente al clustering para *k* = 2 como la figura 20 resultante del mismo ejercicio para *k* = 3, representan densidades de probabilidad, y por lo tanto el tamaño del área bajo la curva depende de la composición de cada cluster, lo cual puede generar alguna confusión a primera vista en el caso particular del *cluster* 0 para *k* = 3, que con solo 8 observaciones y con unos valores particulares que generan la agrupación, presenta prominencias en la gráfica que no son visibles en el caso de *k* = 2 con una conformación diferente de los *clusters*.

Podría decirse que, con la excepción del pequeño *cluster 0* de solo 8 elementos, la solución con *k* = 3 es una especie de “espejo” de la solución ideal con *k* =2, ya que la solución con *k* = 3 hace énfasis en las transacciones positivas extremas, mientras que la solución con *k* = 2 lo hace en las transacciones negativas extremas. Aunque todo tipo de discrepancia genera problemas al negocio, es importante recordar el mayor impacto en la operación de las transacciones negativas, ya que éstas implican una pérdida que impacta de manera directa los estados financieros de la compañía. Esto confirma aún más la elección de *k* = 2 como la mejor solución al problema, algo que fue refrendado por los resultados del experimento de validación *silueta bootstrap* descrito anteriormente.

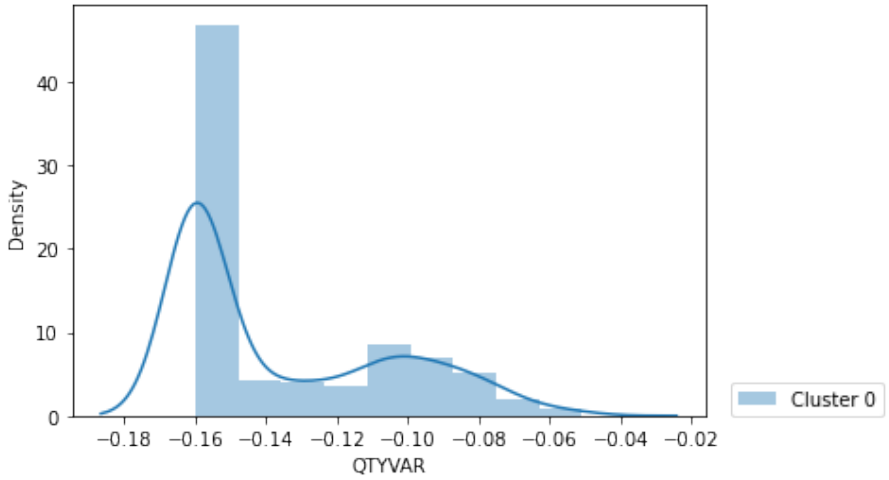
Figura 20. Distribución de probabilidad de COSTX en los *clusters (k = 3)*.



Fuente: Propia.

QTYVAR: las figuras 19, 20A, 20B y 20C presentan la distribución de probabilidad de los dos *clusters* respecto a la columna QTYVAR. La grafica permite apreciar cómo el *cluster* seleccionado para priorización (*cluster 0*) solo incluye ajustes negativos (ver Figura 21) . Es importante aclarar que la máxima variación negativa es de -100%, ya que, al no ser permitidos balances negativos, no se puede reducir la cantidad de un producto en una ubicación de bodega más allá de la totalidad del balance previo al ajuste. Una alta variación de la cantidad en una ubicación de bodega implica que se tenía físicamente una cantidad notoriamente inferior a la que se creía tener según el balance en el sistema. Esto puede llevar a hacer compromisos de despacho y entrega imposibles de cumplir, y resultar en alta insatisfacción de clientes, disminución de los ingresos e incluso perdida de negocios futuros, y en algunos casos, especialmente al volverse un fenómeno repetitivo, perdida de los clientes.

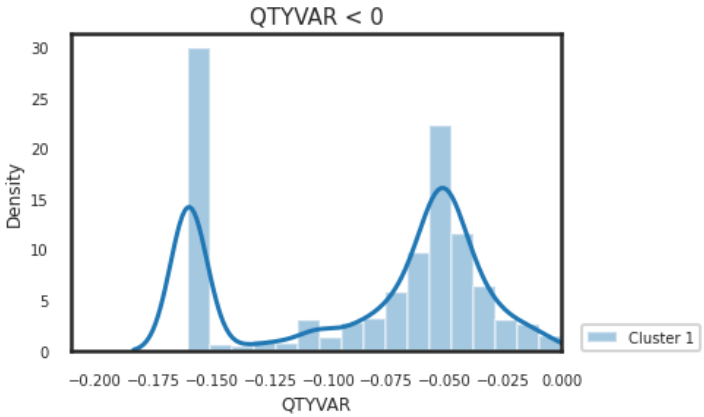
Figura 21. Distribución de probabilidad de QTYVAR (*cluster 0)*.



Fuente: Propia.

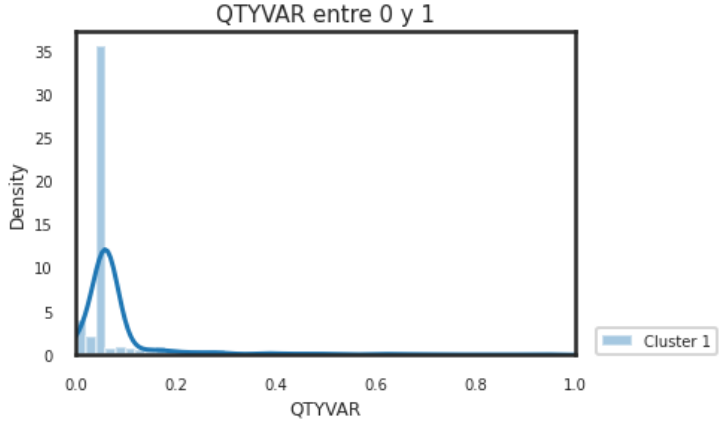
La grafica de las variaciones de QTYVAR para el *cluster* 1 es presentada en tres partes, dado su amplio rango de variación, desde un mínimo negativo similar al QTYVAR del *cluster 0* (ver Figura 22A), pasando por variaciones positivas pequeñas (ver Figura 22B), hasta unas pocas variaciones positivas extremas (ver Figura 22C), no hace posible presentarla claramente en una sola gráfica. Las variaciones positivas extremas representan aquellos casos en los que se encontró en la ubicación de bodega una cantidad muy superior porcentualmente al balance registrado en el sistema de información.

Figura 22A. Distribución de probabilidad de QTYVAR (*cluster 1)* - Parte 1.



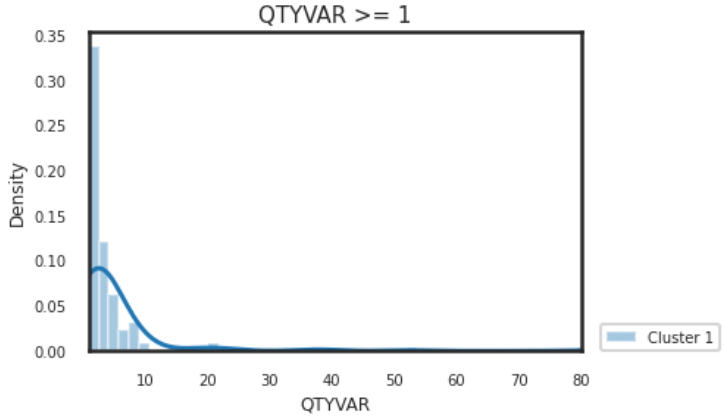
Fuente: Propia.

Figura 22B. Distribución de probabilidad de QTYVAR (*cluster 1)* - Parte 2.



Fuente: Propia.

Figura 22C. Distribución de probabilidad de QTYVAR (*cluster1)* - Parte 3.



Fuente: Propia.

8.5. Clusterboot

La ejecución del algoritmo de *clusterboot* con 100 bootstraps arrojó los resultados presentados en la Tabla 10.

El *cluster 1* – no elegido para priorización – obtuvo una media de 0.6125, con una desviación estándar de 0.0695, lo que lo muestra como un cluster estable y estructurante. No se recupera en ninguna ocasión al no lograr superar la barrera de *0.75*, pero solo se disuelve en 6 ocasiones, y su coeficiente *Jaccard* está por encima de 0.6 para un 94% de las instancias. A primera vista, los resultados no son tan claros para el *cluster 0* - elegido para priorización. El *cluster 0* obtuvo una

media de **0.5554**, con una alta desviación estándar de 0.1721, lo que invita a analizar más de cerca los resultados del experimento para entenderlos claramente. La media en un punto intermedio entre 0.5 – límite mínimo para no ser disuelto – y un prometedor 0.6, acompañada de una desviación estándar que da a entender que hay razones para grandes diferencias entre una instancia y otra, ofrecen esperanzas de encontrar resultados positivos cuando a primera vista el *cluster 0* no parecía haber pasado la prueba. Es verdad que no se recupera en ninguna ocasión al no lograr superar la barrera de *0.75* – igual que sucedió con el *cluster 1*, pero también es cierto que solo se disuelve en 10 ocasiones – lo que equivale a solo el 10% de las instancias. Además, su coeficiente *Jaccard* está por encima de 0.6 para un 61% de las instancias, y su mediana alcanza un valor de **0.6137** (Ver Tabla 10). Todo lo anterior invita a pensar que el *cluster 0* puede resultar tan estable y estructurante como el *cluster 1*.

Tabla 10. Estadísticas del *clusterboot* para *k* = 2.

	Cluster	
	0	1
count	100	100
mean	0.5554	0.6125
median	0.6137	0.6302
std	0.1721	0.0695
disuelve	10	6
recupera	0	0
Entre 0.5-0.6	29	0
> 0.6	61	94
min	0.0308	0.3136
max	0.6809	0.6389

Fuente: Propia.

Al mirar el detalle de las 100 instancias o bootstraps, se puede apreciar que ninguna de las 10 instancias en que el *cluster 0* se disuelve obtuvo un coeficiente *Jaccard* de siquiera 0.07. Esto implica que la conformación de los *bootstraps* (qué transacciones de ajuste de inventario conforman el nuevo conjunto generado con muestreo automático con reemplazo) no ofrece puntos intermedios al coeficiente

Jaccard obtenido de comparar el *cluster 0* con los *bootstraps* generados para cada iteración. Básicamente es una situación de "todo o nada": o se obtiene un sólido coeficiente *Jaccard* por encima de 0.6 (61% de las instancias), o al menos un coeficiente aceptable por encima de 0.5 que evita su disolución (29% de las instancias), o el coeficiente es tan bajo que el *cluster 0* queda lejos de alcanzar siquiera un valor de 0.1, tal y como sucede en el 10% de las muestras (ver Tabla 11). La explicación está en los datos: a pesar de que aproximadamente el 60% de los miembros del *cluster 0* hacen parte de esas 10 instancias en que el *cluster* se disuelve con un muy bajo coeficiente *Jaccard*, los puntos incluidos en esos *bootstraps* hacen que el resultado de *k-prototypes* resulte en dos grandes *clusters* que se reparten los puntos en una proporción cercana al 50% cada uno, por lo que un *cluster* relativamente pequeño, como lo es el *cluster* 0 del conjunto de datos original, no tiene mayor oportunidad de obtener un buen *Jaccard* al compararse con esos dos grandes clusters (Ver Tabla 11).

Tabla 11. *Bootstraps* en los que el *cluster 0* se disuelve.

Bootstrap #	Jaccard
10	0.031244
17	0.058047
23	0.067814
28	0.058088
38	0.030806
60	0.031066
70	0.030846
87	0.067005
90	0.063164
99	0.062538

Fuente: Propia.

Pero al revisar las instancias o *bootstraps* en las que el *cluster 0* no se disuelve con un muy bajo coeficiente de *Jaccard*, se puede apreciar un escenario completamente diferente: una media de 0.6115, una mediana levemente superior

de 0.6153, una pequeña desviación estándar de 0.0318, 29 instancias con un coeficiente *Jaccard* aceptable entre 0.5 y 0.6, y 61 instancias por encima de 0.6 con un valor máximo de 0.6809. A pesar de que en ningún caso se logra superar el umbral de recuperación de 0.75, los resultados obtenidos muestran un *cluster* sólido y estable. Por lo tanto, teniendo claro como la membresía particular de una minoría de *bootstraps* (10%) impacta la media total del experimento, y entendiendo que la gran mayoría de las instancias (90%) muestran un cluster sólido y estable, el experimento se considera un éxito y se procede a continuar con el análisis de las variables categóricas del conjunto de datos (Ver Tabla 12).

Tabla 12. Estadísticas *clusterboot* para *k* = 2 (detalle *cluster* 0)

	cluster 0
count	10
mean	0.0501
median	0.0581
std	0.0159
disuelve	10
recupera	0
Entre 0.5-0.6	0
> 0.6	0
min	0.0308
max	0.0678

	cluster 0
count	90
mean	0.6115
median	0.6153
std	0.0318
disuelve	0
recupera	0
Entre 0.5-0.6	29
> 0.6	61
min	0.5124
max	0.6809

Fuente: Propia.

8.6. Variables categóricas

CAT: el siguiente es el análisis por categoría del producto (CAT), la cual es la principal agrupación de productos en el sistema de información de donde provienen los datos (ver Tabla 13). Tal y como se explicó al comienzo de esta sección, no es relevante que una misma categoría pertenezca a varios *clusters*. Esto simplemente significa que la categoría tuvo ajustes de inventario de ambos tipos – ver por ejemplo CAT1 (categoría 1), la cual tuvo 5 ajustes de inventario que resultaron con membresía en el *cluster 0* y 140 con membresía en el *cluster 1*. La proporción 5:140 tampoco se considera particularmente relevante, ya que la

categoría podría tener un muy alto número de transacciones con membresía en el *cluster 1* (no priorizado) y aun así requerir priorización por su alta representatividad en el *cluster 0* (a priorizar). Ese no es el caso de CAT1, pero si el de CAT12 y CAT14, las cuales tienen respectivamente una participación de 22.1% y 38.7% en el *cluster 0*, para un total del 60.8% con el que se convierten en las principales categorías de productos a considerar para una priorización. Vale la pena mencionar que estas dos categorías tenían solo una participación del 15.7% y 13.2% para un total de 28.9% en el total del conjunto de datos previo al ejercicio de *clustering*.

Tabla 13. Composición de los *clusters* por CAT.

CAT	Cluster	Total
CAT1	0	5
	1	140
CAT10	0	6
	1	245
CAT11	1	1
CAT12	0	112
	1	2,445
CAT14	0	196
	1	1,942
CAT15	0	1
	1	82
CAT16	0	41
	1	2,414
CAT17	0	33
	1	3,953
CAT2	0	2
	1	188
CAT3	0	17
	1	502
CAT4	0	7
	1	97
CAT5	0	12
	1	188
CAT6	1	7
CAT7	0	28
	1	1,194
CAT8	0	57
	1	2,164
CAT9	1	160

Fuente: Propia.

Se realizó una prueba Chi Cuadrado sobre la tabla de contingencias, para verificar la relación de dependencia entre CAT y el cluster ID (cluster 0 o 1) y se obtuvo un p-valor prácticamente igual a cero (9.7×10^{-71}) que permite rechazar con un 95% de confianza la hipótesis nula (H_o) de independencia entre las variables (Heumann, 2016). Esto quiere decir que hay evidencia de la existencia de una relación de dependencia entre las variables CAT y *cluster* ID, y por lo tanto es posible sacar conclusiones a partir de la composición del *cluster 0* seleccionado para la priorización.

PLANNING Y ABC: las tablas 14 y 15 simplemente confirman reglas del negocio por las que se deben priorizar los productos planeados (1/1), los cuales representan un 85.9% del *cluster 0*; y los productos catalogados como A y B de acuerdo con la clasificación anual de inventarios ABC, que representan el 87.2% del *cluster 0* (ver Tablas 14 y 15).

Tabla 14. Composición de los *clusters* por código de PLANNING.

PLANNING	Cluster	Total
1/1	0	444
	1	12,984
2/0	0	25
	1	1,197
1/0	0	48
	1	1,541

Fuente: Propia.

Tabla 15. Composición de los *clusters* por código ABC.

ABC	Cluster	Total
A	0	392
	1	6,681
B	0	59
	1	3,397
C	0	66
	1	5,644

Fuente: Propia.

IG, HTS, UOM, Y AISLE: por simplicidad, ya que una priorización no será exitosa si no se hace verdadero foco en lo principal, se tomó la decisión de analizar las columnas IG (*Item Group*) o grupo de productos, HTS (*Harmonized Tax Schedule code*) o código de importación, UOM (*Unit of Measure*) o unidad, y AISLE o pasillo de la bodega donde se encuentra la ubicación en la que se realizó el ajuste, solo para las categorías de productos CAT12 y CAT14 alrededor de las cuales girarán los esfuerzos de priorización de productos en los conteos cíclicos diarios.

La tabla 16 presenta la composición del *cluster 0* según IG, y en ella se puede apreciar que sería recomendable hacer un énfasis especial en los grupos IG61 e IG64 dentro de la categoría CAT14 (ver Tabla 16).

Tabla 16. Composición del *cluster 0* por código IG.

CAT	IG	Total
CAT12	IG16	22
	IG17	18
	IG30	25
	IG31	26
	IG33	6
	IG34	3
	IG35	12
CAT14	IG61	77
	IG62	1
	IG63	3
	IG64	95
	IG67	10
	IG69	4
	IG70	6

Fuente: Propia.

Se realizó una prueba Chi Cuadrado sobre la tabla de contingencias, para verificar la relación de dependencia entre IG y el cluster ID (cluster 0 o 1) y se obtuvo un p-valor prácticamente igual a cero (5.34×10^{-87}) que permite rechazar con un 95% de confianza la hipótesis nula (H_o) de independencia entre las variables (Heumann, 2016). Esto quiere decir que hay evidencia de la existencia de una

relación de dependencia entre las variables IG y *cluster* ID, y por lo tanto es posible sacar conclusiones a partir de la composición del *cluster 0* seleccionado para la priorización.

La tabla 17 presenta la composición del *cluster 0* según HTS, la cual no ofrece mayores posibilidades de priorización dada la alta variedad de categorías y sus bajas frecuencias (ver Tabla 17)

Tabla 17. Composición del *cluster 0* por código HTS.

CAT	HTS	Total
CAT12	HTS119	11
	HTS15	8
	HTS150	1
	HTS155	1
	HTS16	1
	HTS166	1
	HTS19	1
	HTS36	6
	HTS37	2
	HTS40	6
	HTS41	4
	HTS43	2
	HTS45	1
	HTS46	7
	HTS51	10
	HTS52	1
	HTS53	2
	HTS55	24
	HTS58	2
	HTS59	2
	HTS94	2
	HTS95	13
	HTS98	4

CAT	HTS	Total
CAT14	HTS117	7
	HTS120	1
	HTS15	98
	HTS150	3
	HTS16	1
	HTS161	1
	HTS170	1
	HTS6	77
	HTS76	4
	HTS93	1
	HTS95	2

Fuente: Propia.

La tabla 18 presenta la composición del *cluster 0* según UOM, y sugiere concentrar esfuerzos en productos inventariados en cajas (BX) de las categorías CAT12 y CAT14, y los productos inventariados en yardas (YD) de la categoría CAT12 (ver Tabla 18).

Tabla 18. Composición del *cluster 0* por UOM.

CAT	UOM	Total
CAT12	BDL	2
	BG	5
	BX	35
	RL	1
	YD	69
CAT14	BDL	2
	BX	194

Fuente: Propia.

Se realizó una prueba Chi Cuadrado sobre la tabla de contingencias, para verificar la relación de dependencia entre UOM y el cluster ID (cluster 0 o 1) y se obtuvo un p-valor prácticamente igual a cero (3.88×10^{-52}) que permite rechazar con un 95% de confianza la hipótesis nula (H_o) de independencia entre las variables (Heumann, 2016). Esto quiere decir que hay evidencia de la existencia de una relación de dependencia entre las variables UOM y *cluster* ID, y por lo tanto es posible sacar conclusiones a partir de la composición del *cluster 0* seleccionado para la priorización.

La tabla 19 presenta la composición del *cluster 0* según AISLE o pasillo en el que se encuentra la ubicación física de bodega del ajuste, y sugiere prestar particular atención a los pasillos 32 y 37 (ver Tabla 19) para los productos de la categoría CAT14.

Se realizó una prueba Chi Cuadrado sobre la tabla de contingencias, para verificar la relación de dependencia entre AISLE y el cluster ID (cluster 0 o 1) y se obtuvo

un p-valor prácticamente igual a cero (1.095×10^{-73}) que permite rechazar con un 95% de confianza la hipótesis nula (H_o) de independencia entre las variables (Heumann, 2016). Esto quiere decir que hay evidencia de la existencia de una relación de dependencia entre las variables AISLE y *cluster* ID, y por lo tanto es posible sacar conclusiones a partir de la composición del *cluster 0* seleccionado para la priorización.

Tabla 19. Composición del *cluster 0* por AISLE.

CAT	AISLE	Total
CAT12	2	26
	3	48
	4	116
	5	5
	6	18
	7	28
	8	40
	9	90
	10	30
	11	22
	21	42
	23	46
	24	72
	25	100
	26	52
	27	81
	28	28
	34	68
	99	693

CAT	AISLE	Total
CAT14	4	4
	5	5
	8	16
	9	27
	10	50
	11	341
	28	252
	29	261
	30	210
	31	372
	32	1,088
	33	264
	37	2,664
	99	198

Fuente: Propia.

En resumen, el resultado del ejercicio sugiere priorizar los productos A (ABC = A según la última clasificación ABC anual) de las categorías CAT12 y CAT14 que se planean para inventario con base en su demanda histórica (PLANNING = 1/1). Bajo la categoría CAT14, se recomienda prestar particular atención a los grupos de productos IG61 e IG64 inventariados en cajas (UOM = BX) almacenados en los pasillos (AISLE) 32 y 37. En el caso de la categoría CAT12, se recomienda enfocar esfuerzos en los productos inventariados en cajas (BX) y yardas (YD).

9. CONCLUSIONES Y FUTURO TRABAJO

La combinación de los datos de las transacciones de ajuste de inventario, con información inherente al producto, y la información del negocio implícita en los códigos ABC, permitió estructurar un conjunto de datos que resultó ser un buen insumo para la búsqueda de respuestas al problema de priorización propuesto en este proyecto. Es importante mencionar que se inició trabajos con un amplio número de variables que se redujo drásticamente durante la fase de preprocesamiento, bajo un enfoque de “menos es más”, permitiendo armar un conjunto de datos a la medida del proyecto, sin información redundante, y sin omitir datos que pudieran resultar valiosos a la hora de la priorización. El trabajo de limpieza y preprocesamiento de los datos podrá ser fácilmente replicado en futuras iteraciones por fuera del ámbito académico, para iniciar nuevos procesos de priorización acordes con las que puedan ser las nuevas necesidades y políticas de la empresa.

El *clustering*, como técnica de aprendizaje no supervisado, y en particular el algoritmo *k-prototypes* que permite trabajar con datos mixtos – tan comunes en el mundo real, resultó ser un camino válido y exitoso para encontrar respuestas al problema de priorización de productos en los conteos cíclicos periódicos.

La priorización sugerida por los resultados del ejercicio de *clustering* incluye 818 productos o SKU de 10.827 activos en la bodega principal (donde cada variación de color implica un nuevo SKU). Esta priorización de 818 productos corresponde al 7.6% del total de los productos, mientras que la priorización corporativa recomienda enfocarse en los productos A y B, que corresponden al 21.84% del total. Esto implicaría una reducción del 65.2% de los productos a priorizar, lo cual permitiría diseñar una estrategia de conteos cíclicos más efectiva.

Desde el punto de vista del negocio, la priorización resulta particularmente interesante, ya que las dos categorías sugeridas ocupan el primer y tercer lugar en ventas de la compañía con un 31 y 14% respectivamente, alejadas de las categorías que ocupan el cuarto y quinto lugar con un 7% cada una. Lo interesante no radica simplemente en que las categorías hagan parte del *Top* 3 de ventas, si no en que la categoría que ocupa el segundo lugar – con un 21% del total - no haga parte de la priorización sugerida. Un ejercicio de priorización tradicional basado en el volumen de ventas, sin la utilización de técnicas de *clustering*, podría haber sugerido priorizar el *Top* 2, el *Top* 3, o incluso el *Top* 5, y en todos estos casos se habría hecho énfasis en categorías que no lo requieren según los resultados del ejercicio de *clustering*.

La priorización a un mayor nivel de detalle – teniendo en cuenta los resultados al nivel grupo de productos, código ABC, y unidad de medida – también se encuentra alineada con el negocio, ya que corresponde al 81 y 83% del total de las ventas para cada una de las categorías sugeridas. Esto implica que se hace énfasis en productos de alto volumen de ventas, y representativos de cada una de las categorías, mientras que al mismo tiempo se encuentra la manera de reducir un poco más el número de productos a priorizar al no tomar el camino fácil y seleccionar todos los productos de la categoría. Es claro que un menor número de productos a priorizar hace más manejable los conteos cíclicos periódicos, ya que éstos no son más que conteos manuales realizados por personas en la bodega, aunque se cuente con herramientas tecnológicas como lectores de código de barras que facilitan un poco la actividad.

La reducción en el número de productos a priorizar en comparación con la estrategia corporativa de priorización ABC, la alineación de los resultados del ejercicio de *clustering* con las ventas de la compañía, y la manera como la priorización sugerida logra llegar a un nivel de detalle que se puede incorporar fácilmente a la planeación de los conteos cíclicos, hacen que este proyecto pueda

catalogarse como un éxito. La sola reducción del número de productos a priorizar permitirá planear un menor número de productos por conteo cíclico, con lo que se logrará reducir las oportunidades de error, y al mismo tiempo facilitar la realización de los conteos, lo cual seguramente tendrá un impacto positivo tanto en el clima laboral como en la calidad de vida de los empleados a cargo de las labores de la bodega. Se espera que la nueva priorización basada en datos permita mejorar el nivel de servicio al cliente, al brindar mayores oportunidades de corrección oportuna de las discrepancias, y así evitar consecuencias negativas en el negocio por el mal manejo de órdenes, promesas a clientes, y niveles de inventario. Lo anterior podría tener un efecto cascada y llegar a impactar positivamente tanto las utilidades de la compañía como su posicionamiento en el mercado.

Un observador externo podría pensar en la simplicidad de limitarse a ejecutar *k-Means* solo con variables numéricas – utilizando variables *dummies* o dicotómicas, o cualquier otra técnica para ese efecto. Este “atajo” no permitiría entablar la conversación que este proyecto pretendía, y logró establecer con las transacciones de ajuste de inventario de manera integral, incluyendo información de producto, del negocio, y el detalle de los ajustes en sí.

Los esfuerzos adicionales en que debió incurrir el equipo de trabajo para poder calcular la matriz de distancias con la métrica utilizada por el algoritmo *k-prototypes*, y para implementar los algoritmos utilizados en los experimentos de validación – dada la ausencia de herramientas públicas disponibles que se ajustaran perfectamente a las necesidades del proyecto -, ofrecieron un cierre perfecto a una maestría de profundización, obligando al equipo a lograr el tipo de compresión de los conceptos, las métricas, y los procesos, necesario para lograr desarrollar los diferentes algoritmos, y utilizarlos de manera efectiva para el buen desarrollo del proyecto.

Una importante lección aprendida fue que por más conocimiento previo que haya de los datos, y la aparente certeza de que algunas combinaciones no pueden darse, es importante validar cada una de ellas durante la etapa de preprocesamiento de los datos, para evitar sorpresas posteriores, y la necesidad de iteraciones adicionales completas o parciales.

Es importante resaltar el inmenso apoyo para el equipo de trabajo, y el incalculable valor para el proyecto, que implicó contar con un asesor con amplia experiencia y basto conocimiento, con la capacidad de poder apreciar el panorama global y rápidamente identificar problemas y/o caminos, sin necesidad de estar involucrado en la minucia o en el día a día de las actividades. Su disposición, y capacidad de liderazgo desde afuera, fueron sin duda elementos claves para poder sacar adelante el proyecto en el corto tiempo disponible.

Un interesante aporte de este proyecto es la utilización de técnicas *de machine learning*, y en particular de aprendizaje no supervisado, en un área del negocio donde – de acuerdo con los proyectos identificados durante la investigación - los esfuerzos se han centrado principalmente en un par de nichos: la planeación o predicción de niveles de inventarios, y la clasificación de productos según los códigos ABC. El incursionar en los terrenos de la priorización de productos para solucionar el tradicional problema de la atención oportuna de las discrepancias de inventario, puede convertirse en una interesante invitación para futuros investigadores a atreverse a adelantar este tipo de proyectos, logrando ampliar el rango de cobertura de los aportes del *machine learning* al área de inventarios de cualquier empresa productora o comercializadora de bienes.

Un trabajo futuro interesante sería repetir el ejercicio, pero como primera acción consolidar transacciones al nivel SKU – dejando por tanto completamente de lado las ubicaciones de bodega, para así trabajar los cambios netos de inventario del producto en cada evento – donde un evento correspondería a un conteo cíclico o a

un inventario físico anual. Lo anterior resultaría en menos observaciones por efecto de la consolidación, y no tendría en cuenta los productos mal ubicados, con todas las consecuencias que esto acarrea, pero permitiría centrarse totalmente en la priorización de aquellos productos con desbalances a nivel bodega, o seá, aquellos productos para los cuales hubo una reducción o un incremento neto al final del evento.

Un ejercicio interesante por fuera del campo del *machine learning* sería la realización de una validación en el campo de los resultados del ejercicio de *clustering*, que permita confirmar en la práctica que la priorización sugerida es realmente pertinente, y que su incorporación a la estrategia de inventarios cíclicos periódicos implica mejoras notables en la detección y corrección oportuna de las discrepancias de inventario.

Otro trabajo futuro interesante, aunque con objetivos diferentes, consistiría en empezar a registrar la razón de cada desbalance – en caso de conocerse por supuesto – al momento de realizar la corrección o transacción de ajuste de inventario. En caso de no poderse identificar la razón, podría simplemente usarse un código genérico de “desconocido”. El uso de códigos explicando la razón de la discrepancia, y por lo tanto de la corrección, permitiría – al acumular un buen volumen de observaciones con estos códigos – realizar un análisis buscando entender las principales razones de los desbalances, realizando un proceso de clustering con las mismas variables utilizadas en este proyecto más el nuevo código de razón de desbalance, para tratar de entender las diferentes causas de desbalance entre los diferentes grupos de producto en inventario.

BIBLIOGRAFÍA

Axsäter, S. (2015). Inventory Control (3rd ed.). Springer International Publishing.

Boylan, J. E., Syntetos, A. A., & Karakostas, G. C. (2008). Classification for forecasting and stock control: a case study. Journal of the Operational Research Society, 59(4), 473–481. https://nebulosa.icesi.edu.co:2144/10.1057/palgrave.jors.2602312

Cao, F., Liang, J., Bai, L. (2009). A new initialization method for categorical data clustering. Expert Systems with Applications, Volume 36, Issue 7, 10205-10800. https://doi.org/10.1016/j.eswa.2009.01.060

Chollet, F. (2018). Deep Learning with Python (1st ed). Manning Publications Co.

DeHoratius N., & Mersereau A. J., Schrage L (2008). Retail Inventory Management When Records Are Inaccurate. Manufacturing & Service Operations Management 10 (2), 257-277. https://doi.org/10.1287/msom.1070.0203

de Vos, N. (2021). Documentación paquete k-modes (nicodv/kmodes). https://github.com/nicodv/kmodes/blob/master/kmodes/util/init_methods.py

Efron B, Tibshirani R (1997) Improvement on cross-validation: the 0.632+ bootstrap method. J Am Stat Assoc 92:548–560.

Geher, G., Hall, S. (2014). Straightforward Statistics : Understanding the Tools of Research. Oxford University Press.

Hennig, C. (2007). Cluster-wise assessment of cluster stability, Computational Statistics & Data Analysis, Volume 52, Issue 1, 258-271.

Heumann, C., Schomaker, M., Shalabh (2016). Introduction to Statistics and Data Analysis. Springer International Publishing.

Huang, Z. (1997). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Dept. of Computer Science, The University of British Columbia, Canada, pp. 1–8.

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, 283–304.

IBM Knowledge Center (2019). Guía de CRISP-DM de IBM SPSS Modeler. https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html

James, G., Witten, D., Hastie T., & Tibshirani R. (2013). An Introduction to statistical learning with application in R (8th ed). Springer Science Business Media.

Kang, Y., & Gershwin, S. B. (2005) Information inaccuracy in inventory systems: stock loss and stockout, IIE Transactions, 37:9, 843-859. https://www.tandfonline.com/doi/abs/10.1080/07408170590969861

Magal, S.R., Word, J. (2011). Integrated Business Processes with ERP Systems. Wiley.

Martin, W., & Stanford, R. E. (2007). A methodology for estimating the maximum profitable turns for an ABC inventory classification system. IMA Journal of Management Mathematics, 18(3), 223–233. https://nebulosa.icesi.edu.co:2144/10.1093/imaman/dp1013

Mueller, M. (2003). Essentials of Inventory Management (3rd. ed.). Harper Collins Leadership.

Mukhiya, S.K., Ahmed, U. (2020). Hands-On Exploratory Data Analysis with Python. Packt Publishing.

Ochella, S., Shafiee, M., Sansom, C. (2021). Adopting machine learning and condition monitoring P-F curves in determining and prioritizing high-value assets for life extension, Expert Systems with Applications, Volume 176. https://doi-org.ezproxy.uniandes.edu.co:8443/10.1016/j.eswa.2021.114897

Pedregosa *et al.*, Scikit-learn: Machine Learning in Python version 0.24.2 (abril de 2021). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

RDocumentation, fpc versión 2.2-9. (2020) clusterboot: Clusterwise cluster stability assessment by resampling (diciembre de 2020). https://www.rdocumentation.org/packages/fpc/versions/2.2-9/topics/clusterboot

SciPy.org , SciPy versión 1.6.2 (abril de 2021). Recuperado de; https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.zscore.html

Real Academia Española (s.f.a). Recuperado de: https://dle.rae.es/modelo

Swamynathan, M. (2017). Mastering Machine Learning with Python in Six Steps. Apress.

Vidal, C. J. (2010). Fundamentos de control y gestión de inventarios. Programa Editorial Universidad del Valle.

Waller, M.A., Esper, T.L. (2014). The Definitive Guide to Inventory Management: Principles and Strategies for the Efficient Flow of Inventory across the Supply Chain. Pearson FT Press.

Xu, R., Wunsch, D.C. (2009). Clustering. Wiley.

ANEXO 1

Enlace al repositorio de GitHub del proyecto: incluye los algoritmos de cálculo de la matriz de distancias utilizando la medida de disimilitud de Huang, *clusterboot*, y la validación del coeficiente silueta con *bootstraps*.

https://github.com/FTPGitHub/TDG.git