ISSN 1870-4069

Modelo de agrupamiento de precios en el autoservicio

Leopoldo Javier Gómez Álvarez, Marco Antonio Aristeo García, Luis Alberto Álvarez Ayala

Universidad Panamericana,
Facultad de Ingeniería,
México

{0234446,0152238,0231942}@up.edu.mx

Resumen. En la actualidad el mercado de autoservicios en México presenta un gran dinamismo ya que la mayoría de las cadenas de autoservicios buscan ofrecer el mejor precio a sus clientes con el fin de aumentar su volumen de ventas, lo que las lleva a sacrificar grandes cantidades en margen de utilidad para lograr este objetivo. Por esto decidimos aplicar los siguientes modelos de aprendizaje de máquina que nos permitirán perfilar grupos de artículos que de acuerdo a su rango de precios promedio tengan un comportamiento similar a la diferencia que existe entre su precio y el precio de sus competidores, lo que permitirá generar estrategias puntuales para cada agrupación de productos haciendo más eficientes nuestras inversiones al aplicar rebajas de precio específicas a cada agrupación. Los modelos que empleamos son Agrupamiento Jerárquico,Mean Shift y K-Medias que nos permitirán generar las agrupaciones de acuerdo al comportamiento de los precios y Máquinas de Soporte Vectorial (MSV)que nos ayudará a medir el nivel de precisión de nuestros clusters seleccionados.

Palabras clave: Máquinas de soporte vectorial, clustering, retail, meanshift.

Retail Price Clustering Model

Abstract. Nowadays, the retail in Mexico presents a great dynamism since most of the retailers try to offer the best prices to their clients in order to increase their sales volume, which leads to lost large amounts in profit margin in order to achieve this target. For this reason, we decided to apply the following machine learning models that allow us to create groups of items according to their mean range price and similar patterns with the difference between their price and the competitor's price, this will let us develop unique strategies for each group of products, increasing the profit of their investments by applying specific price reductions to each group. The models used in this article are Hierarchical Grouping,

Mean Shift and K-Means, allowing us to generate clusters according to the patterns of the prices. Vector Support Machines (SVM) will help us to measure the level of precision in our selected groups.

Keywords: Suport vector machine, clustering, retail, meanshift.

1. Introducción

Debido a la estrategia de precios que tienen las cadenas de autoservicio, uno de los principales indicadores al que dan seguimiento es a la diferencia en precios contra sus competidores, lo que es conocido en la industria del autoservicio como diferencial de precios, el cual permite garantizar los mejores precios para el cliente. El cálculo del diferencial de precios se realiza sobre los artículos comparables, es decir aquellos que tienen el mismo código universal de producto o por sus siglas en inglés UPC en ambas cadenas, siendo el UPC un identificador único el cual podemos encontrar en el código de barras de cualquier artículo. El precio promedio de venta de la competencia se obtiene a través de un consolidado de precios del mercado que comparte una agencia de análisis de mercados en donde se detalla el precio de venta promedio a nivel artículo (identificándolo con su código de barras), cadena comercial que lo vende y zona del país en la que se encuentra. La fórmula del diferencial de precios (1)consiste en dividir el precio de venta promedio de la competencia entre el precio de venta promedio de la cadena menos 1, con lo que se obtiene un porcentaje de diferencial:

$$\frac{PrecioPromedioCompetencia}{PrecioPromedioCadena} - 1. \tag{1}$$

A raíz de esto surge la necesidad de identificar la relación que tiene el precio promedio de venta de los artículos con su diferencial de precios, los precios del mercado que usaremos se encuentran a nivel artículo-cadena comercial, les aplicaremos modelos de Agrupamiento Jerárquico,Mean Shift y K-Medias que nos permitirán generar las agrupaciones de acuerdo al comportamiento similar en diferencial de precios en relación a su punto de precio y Máquinas de Soporte Vectorial (MSV)que nos ayudará a medir el nivel de precisión de nuestros clusters seleccionados, todo esto con el objetivo de que una vez identificado el comportamiento del diferencial de los productos analizados se puedan generar estrategias puntuales que permitan hacer más eficientes las inversiones en competitividad de precio. Del set de datos original ejecutaremos los modelos únicamente en la categoría de dulces, ya que es la que representa una muestra más balanceada, como se detallará a continuación.

Existen trabajos que buscan estudiar el comportamiento de los precios en autoservicio [1,2]. En algunos casos desde la óptica de los descuentos basados en la temporalidad, en ese caso para los productos de salsa de tomate y tratando de encontrar la relación misma con la demanda [2] y en otros, bajo circunstancias

específicas para ciertos productos [1] tomando en cuenta una diversidad geográfica en la muestra de las observaciones con las que se desarrolla el artículo. Lo que podemos aprender de ambos enfoques es que siempre se buscará entender el funcionamiento de los precios en el autoservicio.

En [2] se examina los precios en el supermercado para explicar la reducción temporal de precios pasados Basado en el análisis de datos descriptivos, se concluye que el efecto de acumulación de demanda es asimétrico. Afecta la demanda durante precios bajos, pero no durante precios altos. Por lo tanto, las estimaciones de la elasticidad de la demanda de Ketchup basadas en un modelo estático pueden ser sustancialmente sobrevaloradas.

2. Descripción de la propuesta

Nuestra propuesta consiste en el uso de Agrupamiento Jerárquico, Mean Shift y K-medias para perfilar en distintos grupos el comportamiento del precio de los artículos frente a los precios del mercado.

2.1. Descripción General

Con la gráfica del codo calculamos el valor óptimo de centroides. A través de los modelos de Agrupamiento Jerárquico, MeanShift y K-medias buscamos definir los grupos de artículos de la categoría dulces de acuerdo a su precio promedio y diferencial de precios, con el objetivo de definir las etiquetas del set de datos y en el modelo de Máquinas de Soporte Vectorial logramos optimizar esta agrupación.En Agrupamiento Jerárquico con la función agglomerative clustering definimos la localización de los centroides y los productos que se localizarán en cada uno, también graficamos estos clusters en un dendograma, lo que nos permitió identificar el número de artículos que se localizaban en cada uno.Con Mean Shift convertimos nuestro set de datos a un arreglo Nunmpy y obtuvimos las etiquetas y los centroides. Al final lo graficamos para visualizar la localización de estos.En Kmedias se calcularon los centroides a través de la distancia Euclidiana e iteramos hasta que la suma de distancia se redujera hasta que los cambios en distancia fueran mínimos y por último lo graficamos.Con las etiquetas generadas en K-medias usamos el modelo de Máquinas de Soporte Vectorial, optimizando las agrupaciones en las que utilizamos un Kernel lineal por parte de los hiperparámetros, al final evaluamos el modelo. Las etapas a seguir se observan en Fig. 1.

2.2. Materiales

El dataset contiene 3804 productos de una cadena de servicio, con 7 columnas en las cuales podemos encontrar información relacionada a cada uno de los productos.

En primer grado, se valora la distribución de todas las categorías que contiene el set de datos con respecto al precio (ver Fig. 2): La gráfica nos indica que los

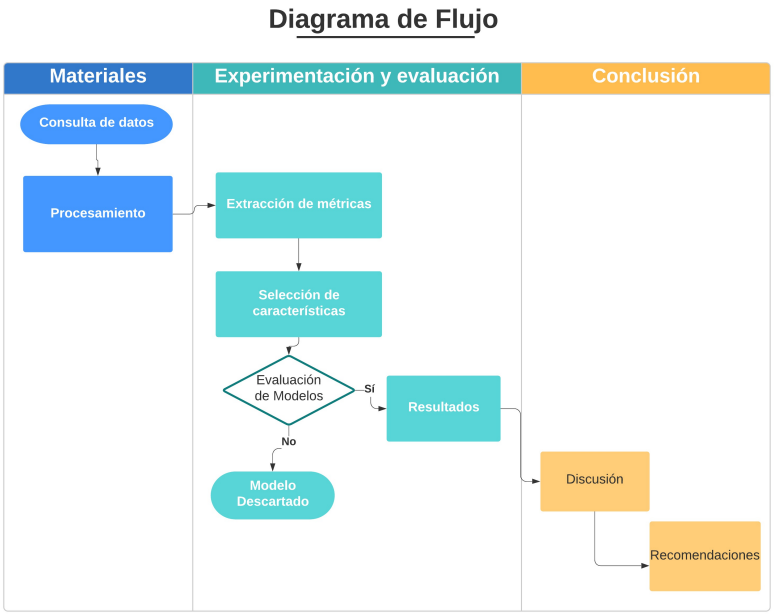


Fig. 1. Diagrama del flujo de trabajo.

Tabla 1. Resumen de datos.

Artefacto	Descripción
Total de observaciones analizadas	3804
Formato del archivo	csv
Columnas Totales	203
Nuevas columnas	2

Tabla 2. Diccionario de datos en columnas seleccionadas.

Columna	Tipo de Dato	No. Valores
Categoría	String	22
UPC	String	3804
Descripción	String	3804
Precio Cadena	String	2391
Último Porcentaje Dif Reportado	String	482
imp_Precio	Float	2391
pct_ult_dif	Float	482

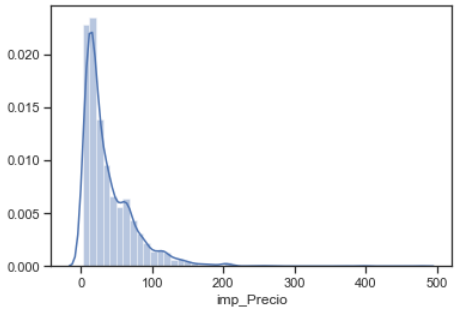


Fig. 2. Distribución de precios de todos los productos.

precios (ver Fig. 3) que mas se repiten en nuestro set de datos, están aproximadamente entre un rango del 0 al 60, por lo cual, decidimos utilizar la categoría de Dulces, debido a que esta categoría alimenta parte de esta gráfica por el rango de precios que contiene la categoría como se puede ver a continuación:

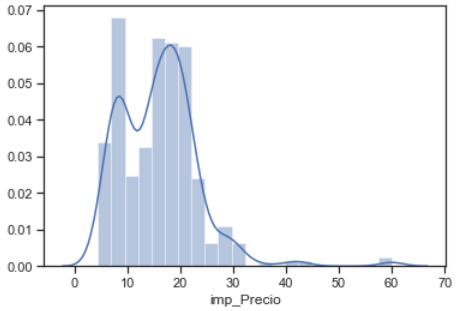


Fig. 3. Distribución de precios de la categoría de ´Dulces.

Realizamos una gráfica de caja (ver Fig. 4) para la categoría de precios, en la cual podemos notar claramente valores atípicos, los cuales pueden afectar algunos de los modelos de clusterización que utilizaremos a continuación. Es importante conservar los valores atípicos encontrados ya que cada punto representa un producto y se busca perfilar la totalidad de los productos.

2.3. Métodos de aprendizaje

Agrupamiento jerárquico aglomerativo Es un algoritmo de agrupación no supervisado, también se le conoce como aglomeramiento anidado ya que cada observación comienza en su propia agrupación y con cada movimiento la une

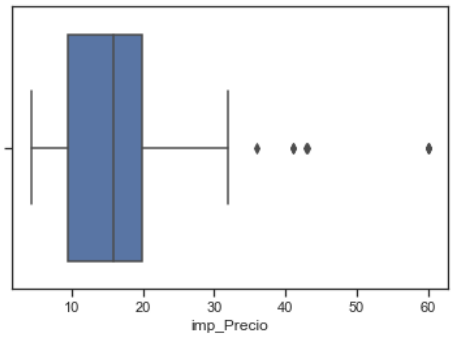


Fig. 4. Boxplot de precios de la categoría de ´Dulces.

con otra agrupación por lo que empieza con K= N clusters y procede uniendo el cluster con el otro que se encuentre más cercano fusinandolos obteniendo así K-1 agrupaciones, esto se repite hasta alcanzar el número deseado de agrupaciones, el criterio para saber que agrupaciones se fusionan es la distancia Euclidiana [4].

Mean shift Es un algoritmo de agrupación no paramétrico que no requiere que se establezca previamente el número de las agrupaciones. Se establece la función $N(X)$ para determinar los puntos cercanos x ∈ X, la medida de distancia usada es la distancia euclidiana y el Kernel $K(d)$ el cuál es un Kernel Gaussiano y la distancia entre dos puntos expresado en la siguiente fórmula:

$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)}. \tag{2}$$

K-medias Es un algoritmo de agrupación no supervisada que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Primero se define el número de grupos, estableciendose K centroides en el espacio de los datos, después cada grupo de datos se asigna al centoride más cercano, el siguiente paso es que se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo, después se iteran estos últimos dos pasos hasta que los centroides se mueven mínimamente. El algoritmo de k medias se centra en la optimización al minimizar la suma de las distancias cuadráticas de cada objeto al centroide de su agrupación [5]

Máquinas de soporte vectorial Es un modelo de aprendizaje supervisado que tiene como inicio el uso de un espacio de las hipótesis de funciones lineales en un espacio de mayor dimensión inducido por un Kernel, en el cual las hipótesis son entrenadas por un algoritmo tomado de la teoría de optimización, el cuál

utiliza elementos de la teoría de generalización. La forma en la que la MSV aprende la función objetivo es cambiando la representación de la función, es decir se mapean los espacios de entradas X a un nuevo espacio de características $F = \phi(x)|x \in X$ En donde las cantidades que se intoducen para describir los atributos son conocidos como características [3].

3. Experimentación

Gráfica de codo Por medio de la gráfica de codo, podemos obtener el valor óptimo de K, con base en la curva obtenida en la gráfica (ver Fig. 5), usando los parámetros de la tabla 3 se escoge entonces ese punto en donde ya no se dejan de producir variaciones importantes . En este caso, vemos que esto se produce a partir de k >= 3, por lo que evaluaremos los resultados del agrupamiento, con el valore de 3 a fin de observar el comportamiento del modelo.

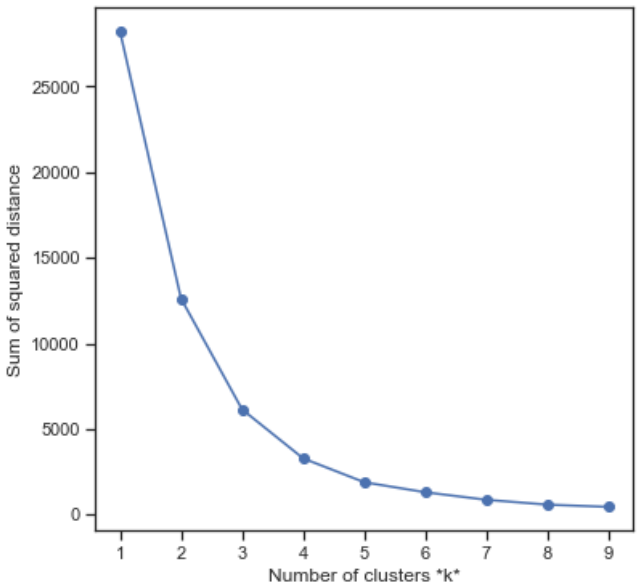


Fig. 5. Gráfica de codo.

Tabla 3. Parámetros para generar gráfica de codo.

Parámetro	Valor
Datos	'imp_Precio','pct_ult_dif'
Longitud de gráfica	range(1, 10)

Agrupamiento jerárquico De la librería Sklearn en Python, importamos la función agglomerative clustering (ver Fig. 6) en la cuál definimos que los datos se debían agrupar en tres clusters (ver Cuadro 4) y lo gráficamos como se muestra a continuación utilizando del set de datos, el precio y último diferencial. Las columnas utilizadas del set de datos son el precio promedio y el último diferencial.

Tabla 4. Parámetros para generar agrupamiento jerárquico.

Parámetro	Valor
Número de clusters	3
Linkage	average

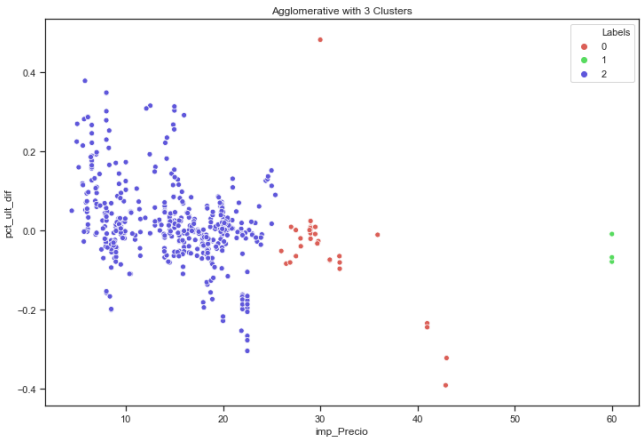


Fig. 6. Cluterización con agrupamiento jerárquico.

En el dendrograma (ver Fig. 7) podemos notar como se van formando los grupos conforme a la siguiente jerarquización con el enfoque de abajo hacia arriba.Se determina la distancia vertical más grande que no se cruza con ninguno de los otros grupos a partir del eje y. Comienza con muchos grupos pequeños y se combinan para crear grupos más grandes:

Mean Shift Primero convertimos el set de datos de un arreglo Pandas a un arreglo Numpy,después utilizamos Mean shift y lo ajustamos al set de datos,en el siguiente paso dejamos que el modelo estableciera las etiquetas y centroides usando los parámetros que se muestran en el cuadro 6. Al final graficamos los centroides y sus respectivos clusters de datos (ver Fig. 8). Las columnas utilizadas del set de datos son el precio promedio y el último diferencial.

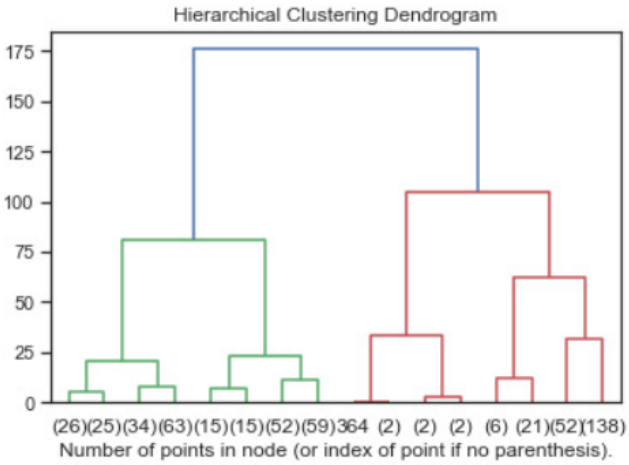


Fig. 7. Dendograma agrupación jerárquica.

Tabla 5. Parámetros para generar MeanShift.

Parámetro	**Valor**
Quantile	0.9
n_samples	500
bin_seeding	True

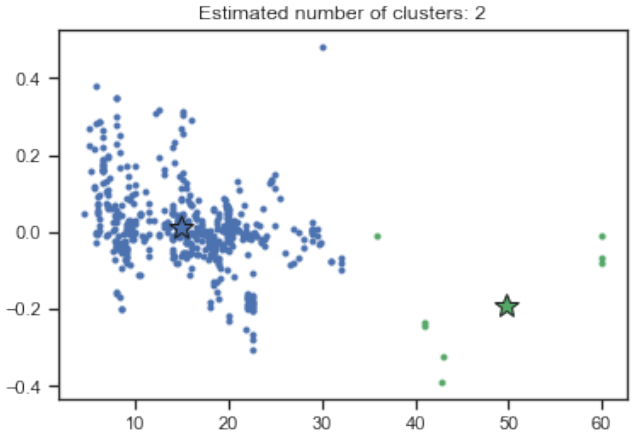


Fig. 8. Cluterización con agrupamiento mean shift.

K-medias En el primer paso cada fila de nuesto conjunto de datos se asigna al centroide más cercano basado en la distancia Euclideana como se aprecia en (3):

$$\arg\min_{c_i \in C} dist(c_i, x)^2. \tag{3}$$

Después recalculamos los centroides tomando una media de los puntos asignados en el paso anterior. Iteramos hasta que la suma de las distancias se minimiza y no hay cambios en los puntos asignados a los grupos. Graficamos los datos agrupados con un scatterplot diferenciandolos con colores (ver Fig. 9). Las columnas utilizadas del set de datos son el precio promedio y el último diferencial.

Tabla 6. Parámetros para generar K-medias.

Parámetro	Valor
Número de clusters	3

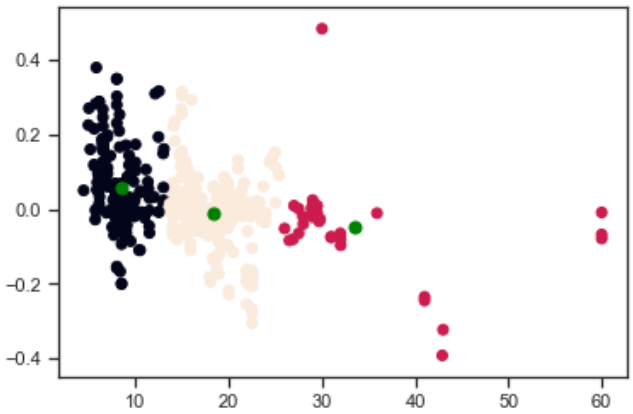


Fig. 9. Cluterización con k medias.

Mediante una tabla de correlación confirmamos la independencia entre el precio promedio y el porcentaje de diferencial. Por último enlistamos los artículos que pertenecen a cada clasificación 7.

Máquinas de Soporte Vectorial Utilizamos el modelo de máquinas de soporte vectorial junto con el modelo de K-medias, ya que SVM funciona por medio de etiquetas y este modelo no es capaz de realizarlo, por lo cual optamos por utilizar las etiquetas obtenidas previamente en el modelo de K-medias.

Tabla 7. Tabla de correlación.

Métrica	UPC	Precio Promedio	Último Diferencial
UPC	1.000000	-0.032489	-0.121398
Precio Promedio	-0.032489	1.000000	-0.386194
Último Diferencial	-0.121398	-0.386194	1.000000

Para utilizar este modelo, dividimos los datos en 70 % entrenamiento y 30 % para pruebas (ver Cuadro 8). En este método, es recomendable estandarizar el subconjunto de datos de entrenamiento y apartir de esto empezar con el modelado, utilizando un kernel lineal por parte de los hiperparámetros. Al final, evaluamos nuestro modelo por medio de una matriz de confusión (ver Fig. 10), la cual nos arrojo una exactitud del 98 %, utilizando un K=3.

Tabla 8. Parámetros para generar MSV.

Parámetro	Valor
Kernel	lineal
test_size	0.3
random_state	0

4. Resultados y discusión

Al realizar el análisis de los modelos planteados en este documento encontramos los siguientes pros y contras en cada método.

En el modelo de Kmedias tuvimos un error alto, ya que nuestro set de datos contenía puntos atípicos, lo que nos llevaría a tener que descartar productos del mismo para disminuir el error. Mientras que en Mean Shift esto no afectó las agrupaciones obtenidas debido a que el modelo define por si mismo el número de agrupaciones y la ubicación de los centroides, permitiendonos conservar todos los artículos de la categoría de dulces sin presentar un error alto en el modelo, lo que asegurará que las estrategias de competitividad de precios que generemos a partir de estas agrupaciones contemplen todo el catálogo de productos. Además de que el no incluir un producto en la estrategia de precios afectaría la percepción que tiene el cliente de la cadena de autoservicio y dejaría en desventaja al artículo frente al precio más bajo ofrecido por la competencia.

Como se muestra en las siguientes 4 graficas, podemos notar que hay consitencia en los grupos definidos por los clusters. Los primeros grupos definidos por el cluster jerarquico y por meanshift son similares respecto a los precios de los productos. Mientras que en K-medias, notamos que ese primer bloque, el algoritmo lo divide en 2 grupos:

	precision	recall	f1-score	support
0	1.00	0.93	0.97	46
1	1.00	1.00	1.00	7
2	0.97	1.00	0.99	101
accuracy			0.98	154
macro avg	0.99	0.98	0.98	154
weighted avg	0.98	0.98	0.98	154

```
[[ 43   0   3]
 [  0   7   0]
 [  0   0 101]]
```

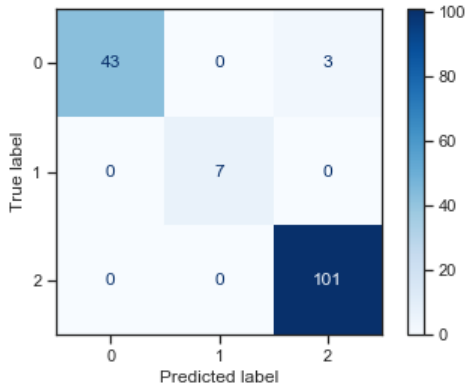
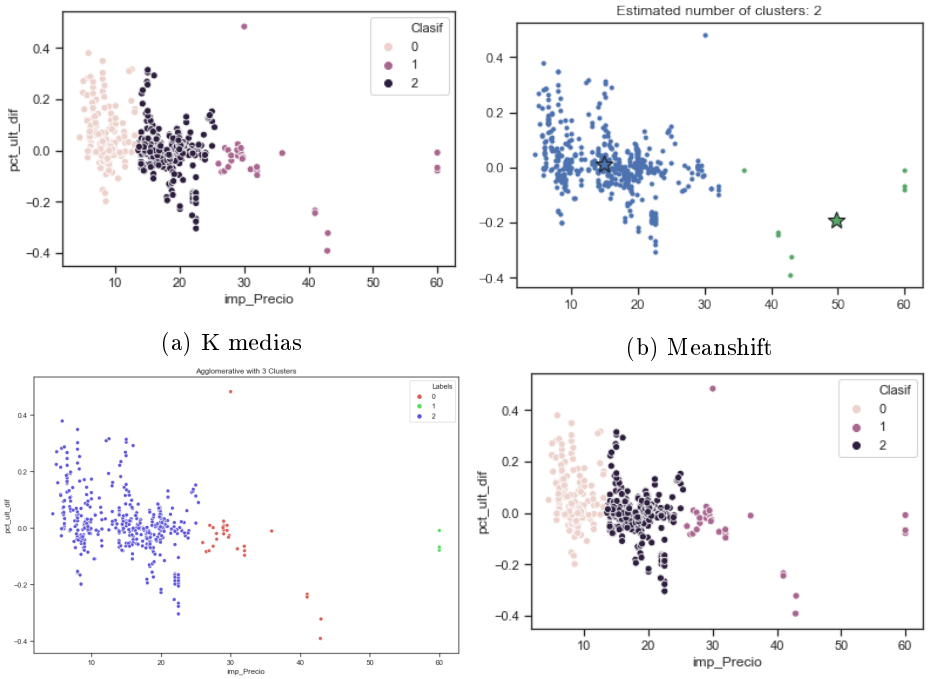


Fig. 10. Matriz de confusión para SVM.

Con los dos grupos identificados en los agloritmos Mean Shift y jerárquico podemos definir que el grupo que se encuentra entre cero y treinta pesos (cluster 0) debe tener una estrategía diferente al que tiene un precio mayor a treinta pesos (cluster 1), ya que los artículos contenidos en el primero presentan una mayor rotación debido a su punto de precio bajo y por lo tanto es importante profundizar la inversión en diferencial de precios para contar con un precio competitivo mientras que los productos del cluster 1 son principalmente importados u orgánicos, que se encuentran dirigidos a un mercado de clientes con un mayor nivel socioeconómico y por lo tanto la inversión en diferencial de precios no debe ser tan agresiva para no afectar los márgenes de utilidad alto que aporta este segmento.

Por lo que podemos afrimar que aquellos productos con un precio menor a treinta pesos (cluster cero) deben continuar con una estrategia inversión agresiva que se traduzca en un diferencial mayor y en los productos con un precio mayor a treinta pesos (cluster 1) la inversión debe de ser mínima, lo que nos ayudará a ser más eficiente con los recursos invertidos y al mismo tiempo a asegurar la competitividad de nuestros precios donde es necesario.



(a) K medias (b) Meanshift

(c) Jerarquico (d) SVM

Fig. 11. Agrupaciones por modelos.

5. Conclusiones

Podemos concluir que los modelos de Agrupamiento Jerárquico y Meanshift son convenientes para identificar el comportamiento de precios de los artículos en el autoservicio lo que se traduce en una herramienta útil para la creación de una estrategia de precios efectiva que nos permita maximizar el margen de los productos e invertir de forma eficiente para tener la competitividad necesaria frente al mercado.

Referencias

1. Hosken, D., Reiffen, D.: Patterns of retail price variation. RAND Journal of Economics pp. 128–146 (2004)
2. Pesendorfer, M.: Retail sales: A study of pricing behavior in supermarkets. The Journal of Business 75(1), 33–66 (2002)
3. Resendiz Trejo, J.A.: Las maquinas de vectores de soporte para identificación en línea. Centro de Investigación Y de estudios avanzado del Instituto Politécnico Nacional Departamento de Control Automático 1(1), 1–68 (2006)

4. Teichgraeber, H., .B.A.R.: Systematic comparison of aggregation methods for input data time series aggregation of energy systems optimization problems. Computer Aided Chemical Engineering 44(1), 955–960 (2018)
5. Wu, J.: Advances in K-means Clustering, Springer Science & Business Media, vol. 1, chap. 1, pp. 1–50. Springer (2012)