30-3-2019

Aplicación de algoritmos de Machine Learning para la estimación del precio y predicción de los productos más vendidos de ropa para perros en Amazon en base a sus atributos

Memoria

Stephanie Hevia Sahmkow

- ***Título***

Aplicación de algoritmos de Machine Learning para la estimación del precio y predicción de los productos más vendidos de ropa para perros en Amazon en base a sus atributos.

- ***Descripción del proyecto***

Se utilizará Machine Learning para predecir los productos más vendidos y precios de Ropa de perros en Amazon en base al tipo de prenda, clasificación de los más vendidos, valoraciones, estaciones del año de uso, si el producto es prime o no, su antigüedad de publicación, material, colores y talla.

- ***Motivo del desarrollo del proyecto:***

Se ha elegido este proyecto debido a que tengo un gran interés por la industria de las mascotas, específicamente para perros, ya que es uno de los sectores con mayor crecimiento y en un futuro próximo desearía desarrollar un Marketplace para comercializar ropa y accesorios para perros.

Con el desarrollo de este proyecto, se puede determinar las características que determinan el precio y posicionamiento de un producto en Amazon, lo cual lo podría aplicar en un futuro con mi propio data set de productos que se vayan a comercializar en mi Marketplace para determinar los precios apropiados de los productos y para saber cuáles son los más vendidos para ofertarlos en la página web.

Inicialmente se ha elegido Amazon debido a que, según el estudio realizado por la consultora Tandem acerca de los Anual Marketplace del 2019, Amazon es uno de los Marketplace más utilizado en España. Adicionalmente, Amazon es una de las pocas plataformas que proveen mayor información sobre los productos, incluyendo su posicionamiento de ventas, valoraciones y reseñas de los clientes, lo que la hace más confiable para los compradores.

- ***Antecedentes***

En un estudio realizado en el 2014, aplica Machine Learning para predecir compras por clientes en línea basado en el Precio Dinámico [1].

Actualmente, los precios dinámicos están mejorando y forman una importante parte en la industria del e-commerce. Los precios dinámicos consisten en ofrecer bienes a precios distintos que varían de acuerdo a la demanda de los clientes, a los precios de los competidores, a la oferta de los mismos y a las metas en ventas [2].

Hay varios modelos que son utilizados actualmente para determinar el precio dinámico como modelos basados en agentes, en inventarios, data driven, y en teorías del juego. Los modelos de Machine learning incorporan el uso del e-commerce para el entendimiento de las preferencias y patrones de los compradores y uso de los algoritmos para maximizar las ganancias.

El machine learning también es utilizado actualmente para predecir las ventas en el negocio del e-commerce. Hay un estudio que fue recientemente publicado en el que se utiliza el modelo

Long Short-Term Memory network (LSTM) para determinar las relaciones no lineales de la demanda en la jerarquía del surtido de los productos de e-commerce. Utilizan modelos unificados, incorporando información sobre productos relacionados los cuales pueden tener patrones de demanda que estén correlacionados. [3]

Otro estudio publicado en el 2015, se enfocó en predecir la venta de los productos en Amazon, en base a las valoraciones, reseñas y estrategias promocionales utilizando redes neuronales y finalmente concluye que las variables mencionadas anteriormente predicen la venta de los productos. [4]

Adicionalmente, hay otro estudio muy parecido al anterior, en el que incluye más variables para predecir las ventas utilizando redes neuronales. Estas variables fueron: ranking de ventas, volumen de reseñas, promedio valoración de estrellas, rating de las reseñas positivas más útiles, número de personas que encuentran las reseñas muy útiles (encuesta), análisis sentimental positivo y negativo, polaridad de los sentimientos y ranking de los que realizan las reseñas. El estudio concluyó que todas las variables son importantes para la predicción de las ventas en Amazon. [5]

- ***Objetivo general***

Predecir los productos más vendidos y precio de Ropa de Perros en Amazon en base a sus diferentes atributos.

- ***Objetivos específicos***

En el primer objetivo, se desea predecir los productos más vendidos en Amazon (por su posicionamiento) en base a su precio, talla, color, tipo de prenda, estaciones del año de uso, si el producto es prime o no, su antigüedad de publicación, material, colores y tallas. Para lograr esto, se aplicarán distintas regresiones lineales múltiples con diferentes métodos de validación. Si se obtiene un R2 alto, se procederá a utilizar clasificadores para predecir el posicionamiento. Finalmente se compararán los resultados de las regresiones y clasificadores y se elegirá el que tenga mejor resultado.

El segundo objetivo se basa en determinar el precio de los productos en base a los atributos nombrados anteriormente. Se realizará utilizando distintas regresiones lineales múltiples con distintos métodos de validación y si no se alcanza un R2 alto, se aplicarán distintos métodos de clasificación para predecir el mismo con distintos métodos de validación. Se elegirá el método con el mejor resultado.

Finalmente, se realizará un clustering para determinar como se relacionan los atributos entre ellos y ver como se podrían agrupar para extraer información relevante de ello.

- ***Metodología DS***

- ¿Cuál es el problema que se intenta resolver?

Saber si el producto entra dentro de los más vendidos.

¿Cuál es el precio ideal para asignar a un producto?

- ***Recolección de Datos y descripción del dataset***

Los patrones se obtienen de Amazon y se eligen en base a las siguientes clasificaciones:

- Los más vendidos en Camisas para perros
- Los más vendidos en Chubasqueros para perros
- Los más vendidos en Disfraces para perros
- Los más vendidos en Sudaderas con capucha para perros
- Los más vendidos en Abrigos y chaquetas para perros
- Los más vendidos en Jersey para perros

En un principio, varios de los productos fueron elegidos de la clasificación "Los más vendidos en Ropa y accesorios para perros", para tener una visión general de la ropa más vendida sin especializarse en ningún tipo de prenda. Luego dentro de esta clasificación, entran las otras nombradas anteriormente, lo cual refleja el posicionamiento del producto por su categoría.

Estas clasificaciones representan el ranking de ventas de Amazon y se utiliza para medir la popularidad de un producto dentro de su categoría en comparación con la competencia. La elegibilidad de un producto para la métrica se determina por el volumen de ventas recientes y los datos de ventas históricos relativos a cualquier otro producto que sea de la misma categoría.

Los atributos a incluir en el data set son los siguientes:

- Precio
- Posicionamiento
- Clasificación en los más vendidos de Amazon (posicionamiento)
- Talla
- Color
- Tipo de prenda
- Estaciones del año de uso
- Si el producto es prime o no
- Su antigüedad de publicación
- Material

Se obtendrán transcribiéndolos manualmente a un archivo de Excel para luego ser analizados en Jupyter.

Los datos recolectados se consideran estructurados y hay tanto numéricos como categóricos.

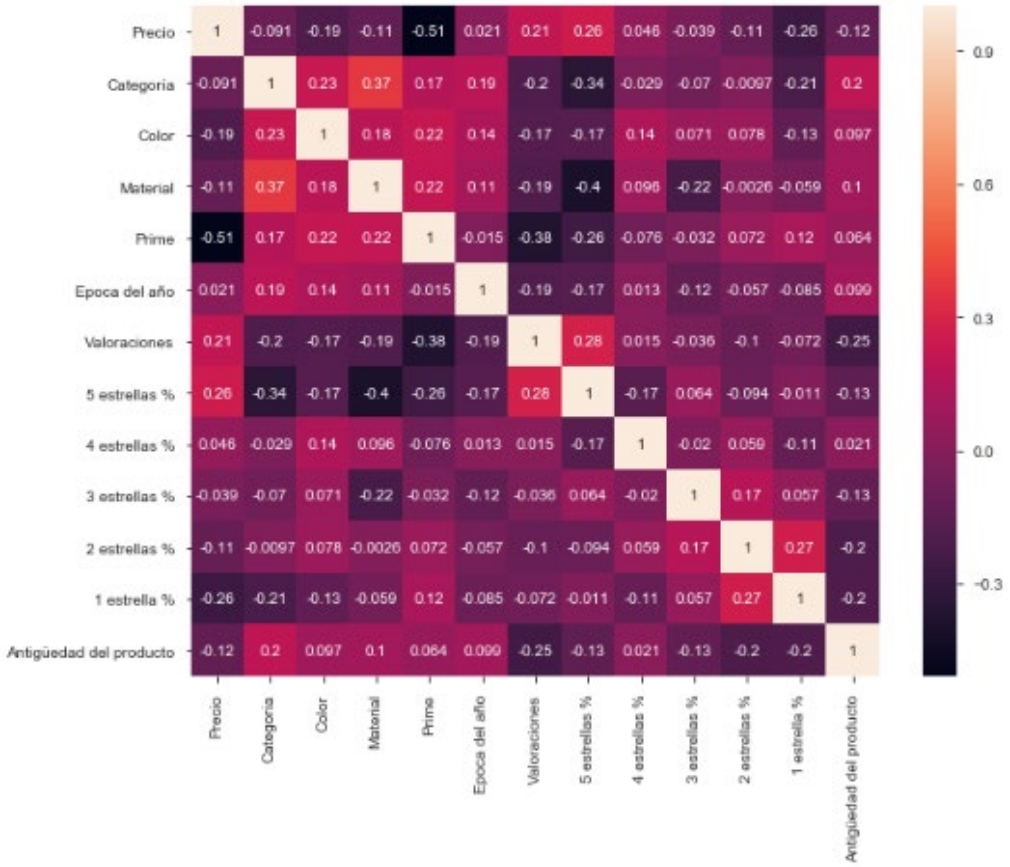
Dentro de los numéricos se encuentran el precio, valoraciones, promedio estrellas, los porcentajes de estrellas, antigüedad del producto y el resto son categóricos.

- ***Comprensión de datos***

Se representarán en gráficas las estadísticas de los productos en general.

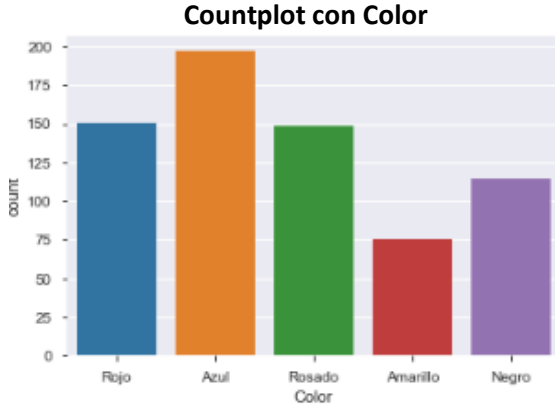
Se pueden visualizar los datos con las siguientes gráficas:

- Mapa de calor para determinar la correlación entre los atributos.



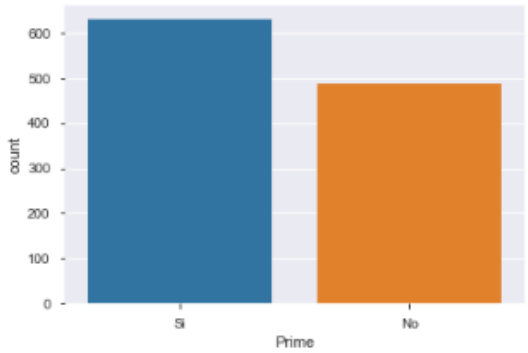
Se puede observar que los atributos que tienen mayor correlación son Categoría y Material, de resto no hay correlación entre ellos.

- Gráficas de frecuencia (countplot) con los atributos color, talla, época del año, material y categoría.



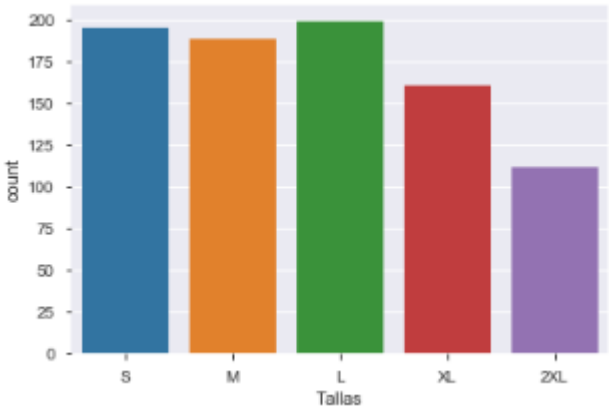
Los colores más vendidos son el azul, rojo y rosado.

Countplot con Prime



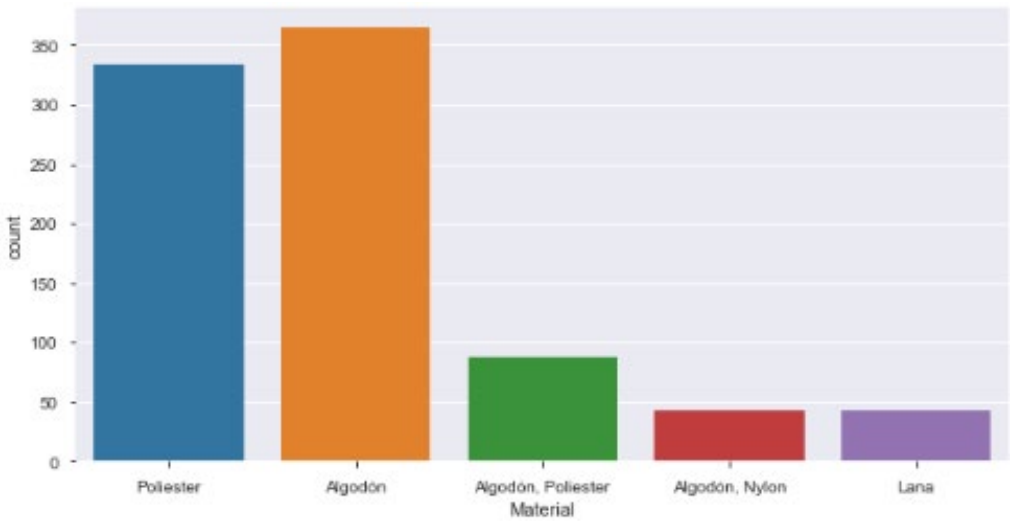
La mayoría de los productos tienen la opción de Prime.

Countplot con Tallas



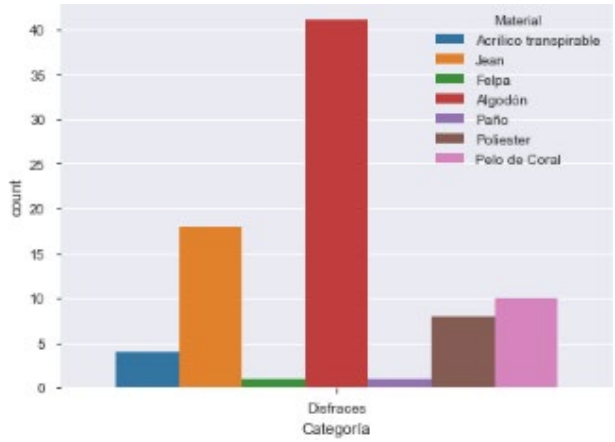
Las tallas más vendidas son la S, M y L.

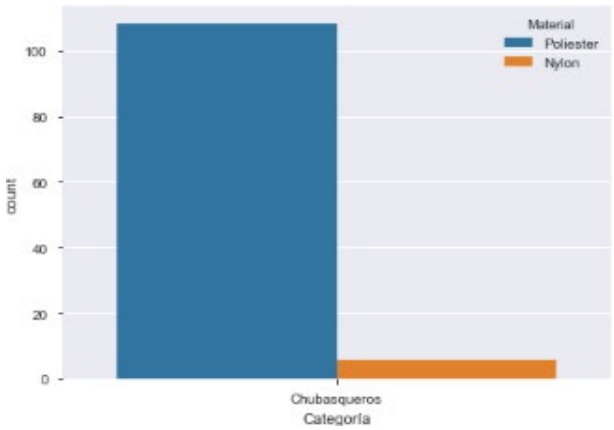
Countplot con Material

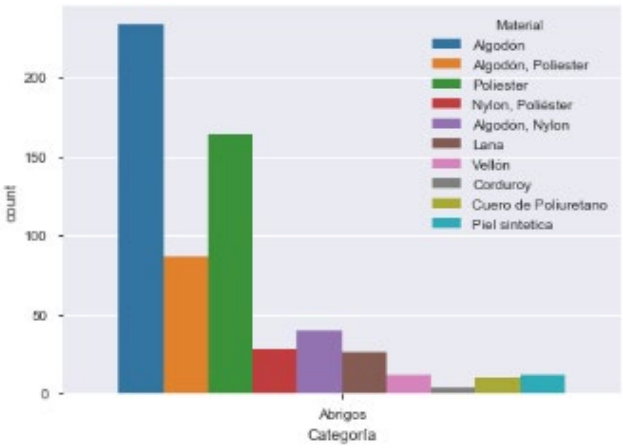


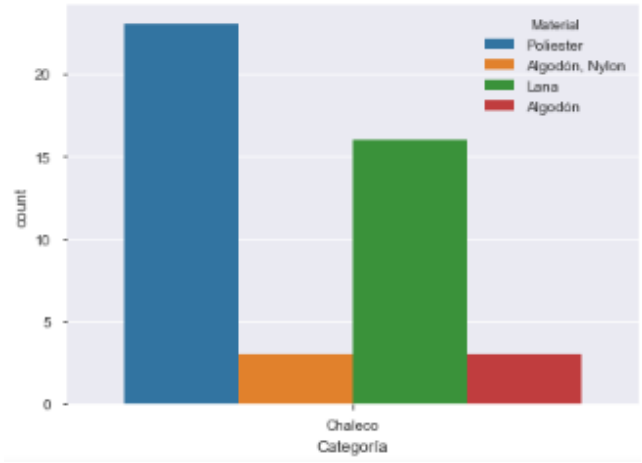
La mayoría de los productos más vendidos están hechos de Poliéster y el Algodón.

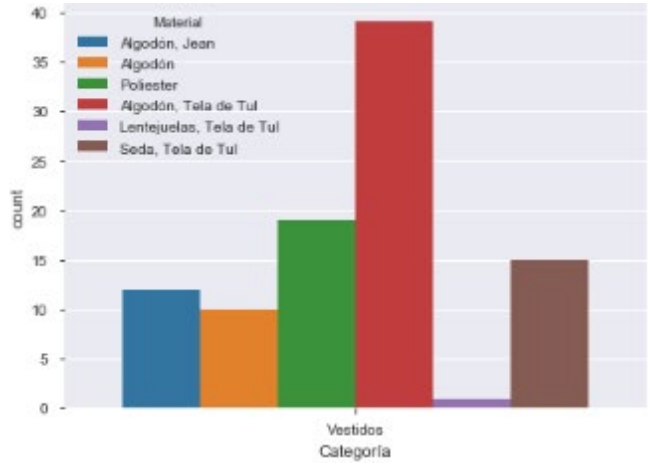
- Gráficas de frecuencia que representen los materiales que más se utilizan por cada categoría.

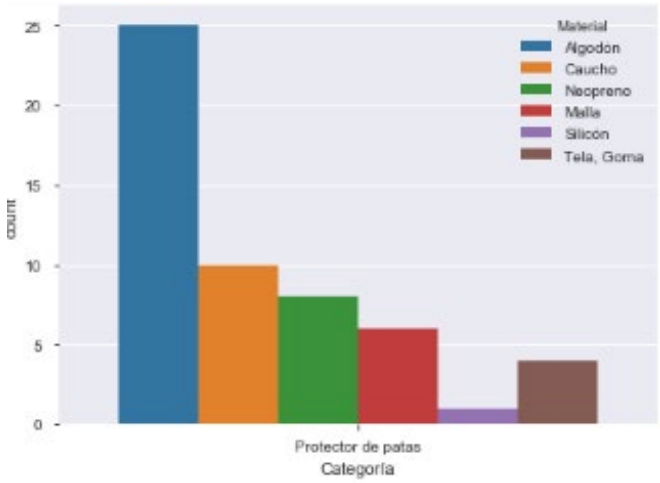


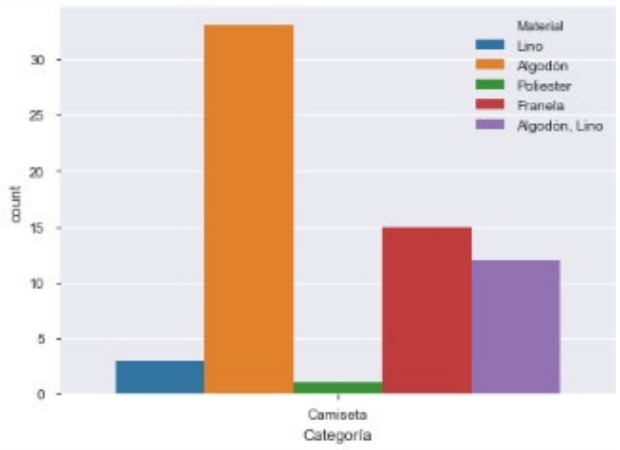


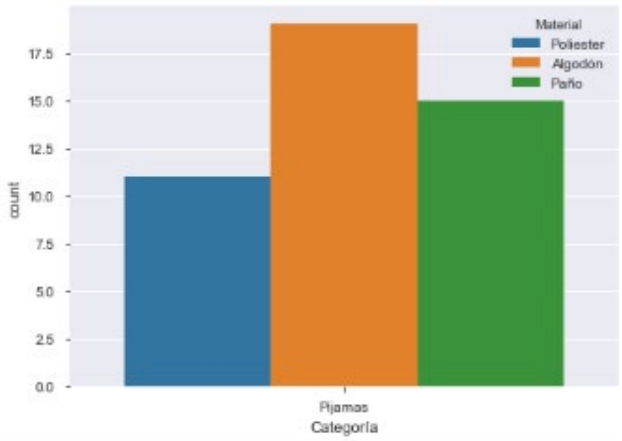



Material
Poliester
Algodón, Nylon
Lana
Algodón
count
0
5
10
15
20
Chaleco
Categoría


Material
Algodón, Jean
Algodón
Poliester
Algodón, Tela de Tul
Lentejuelas, Tela de Tul
Seda, Tela de Tul
count
0
5
10
15
20
25
30
35
40
Vestidos
Categoría


Material
Algodón
Caucho
Neopreno
Malla
Silicón
Tela, Goma
count
0
5
10
15
20
25
Protector de patas
Categoría

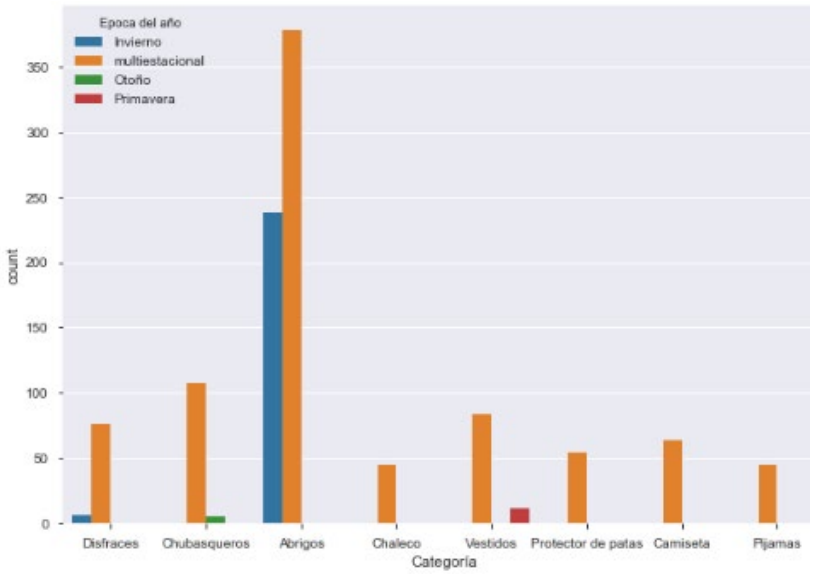




Se puede observar en estas gráficas lo siguiente:

- Los disfraces, protectores de patas, abrigos, camisetas y pijamas en su mayoría están hechos de algodón.
- Los chubasqueros y chalecos en su mayoría están hechos de poliéster.
- Los vestidos en su mayoría están hechos de Algodón y tela de tul.

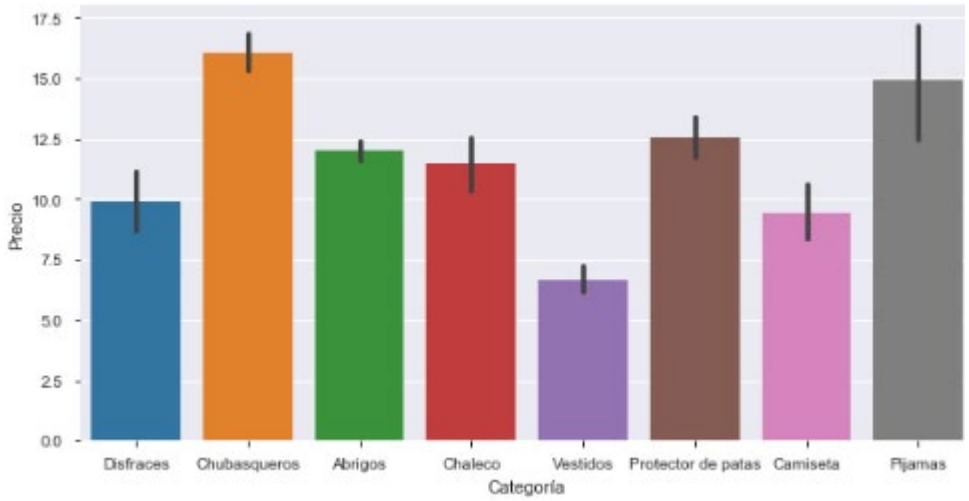
- Gráfica de frecuencia que represente la época del año en donde se utilizan las categorías.



Se puede observar que las categorías más vendidas son los abrigos, seguido por los chubasqueros y vestidos. También se puede observar que estos productos son utilizados más de una estación al año (otoño e invierno) y que los abrigos son utilizados en su mayoría en invierno.

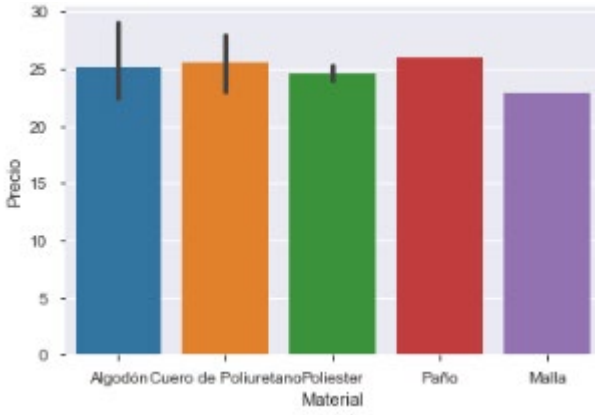
- Gráficas de barra del precio con respecto al resto de los atributos.

Barplot de Categoría contra Precio



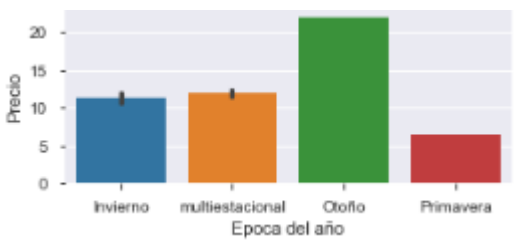
Se puede ver que, en promedio, los chubasqueros son los productos con mayor precio, siguiéndolo los pijamas y los protectores de patas.

Barplot del Material contra Precio



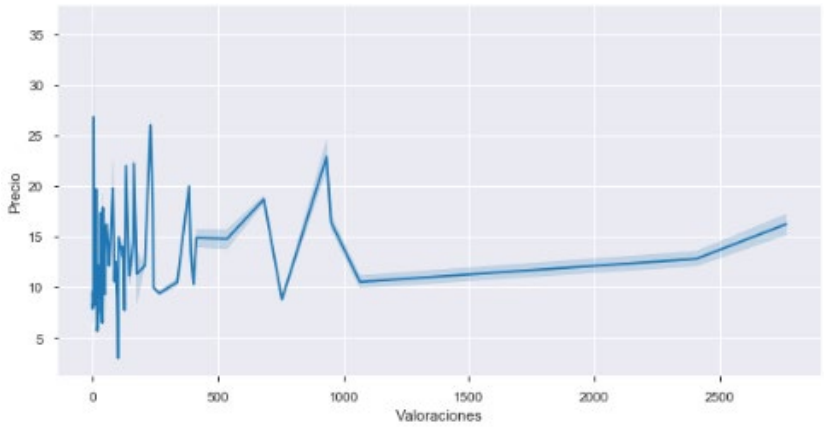
Se puede observar que los materiales con mayor precio en la ropa de perros son el poliéster, el paño y el algodón.

Barplot del Época del año contra Precio



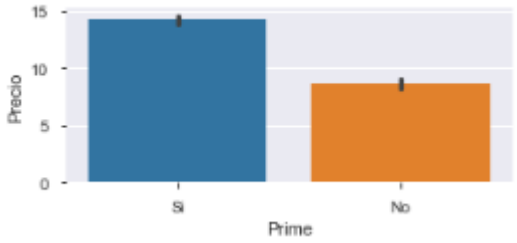
Se puede observar que la ropa que se utiliza en la estación de otoño, tiene mayor precio con respecto al resto de las estaciones. Relacionándolo con los atributos anteriores, en otoño se vende mayormente los Chubasqueros, el cual es el producto más costoso, y en su mayoría están hechos con el material poliéster que es uno de los que tienen mayor precio, como se nombró anteriormente.

Lineplot de Valoraciones contra Precio



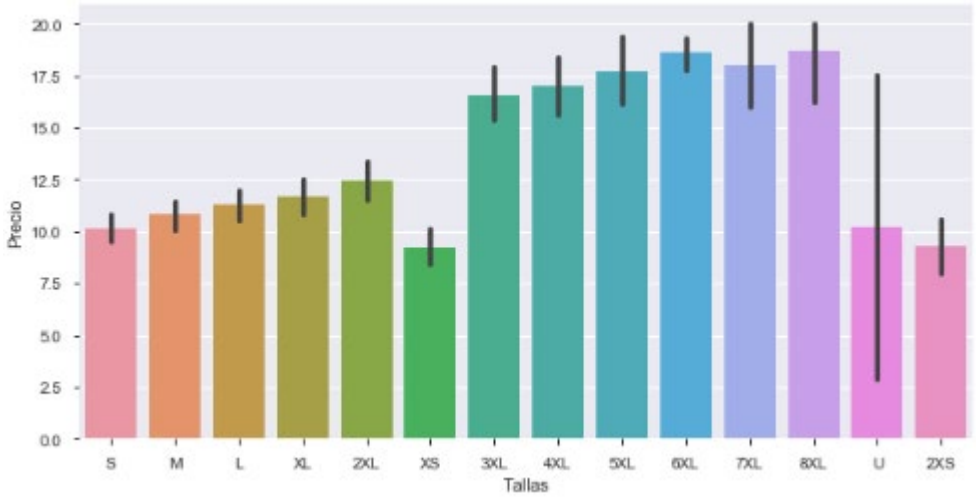
Se puede observar que el número de las valoraciones no afectan al precio del producto.

Barplot de Prime contra Precio



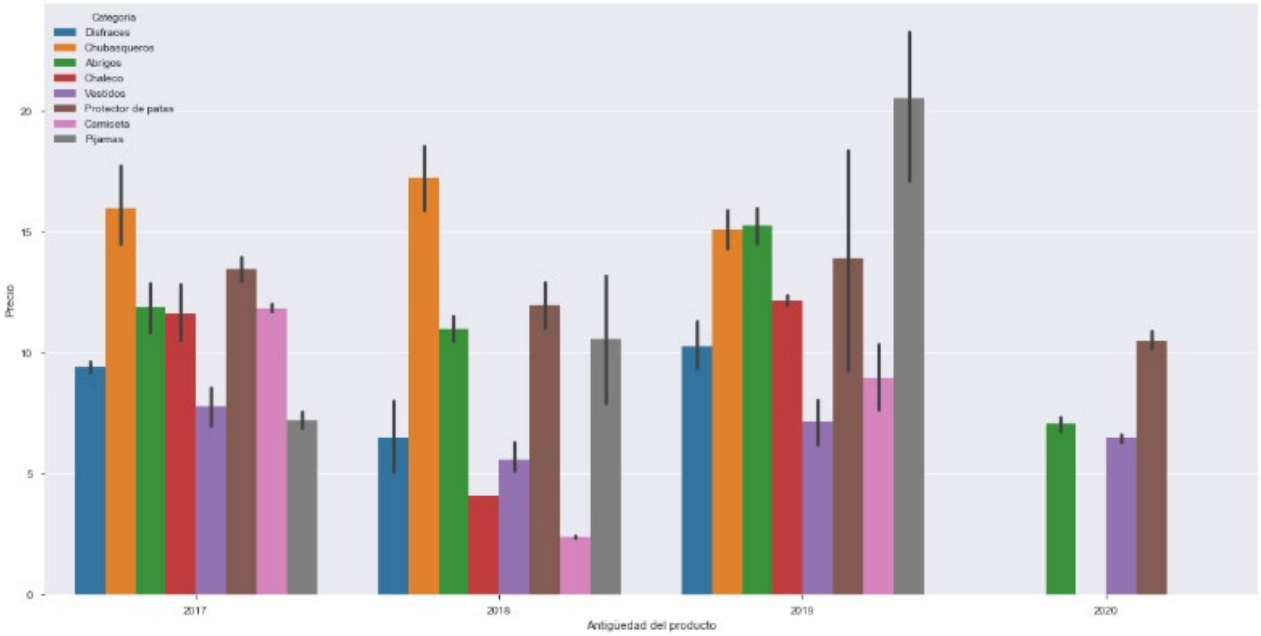
Se puede observar que los productos que tienen Prime tienen mayor precio que los que no lo tienen.

Barplot de Tallas contra Precio



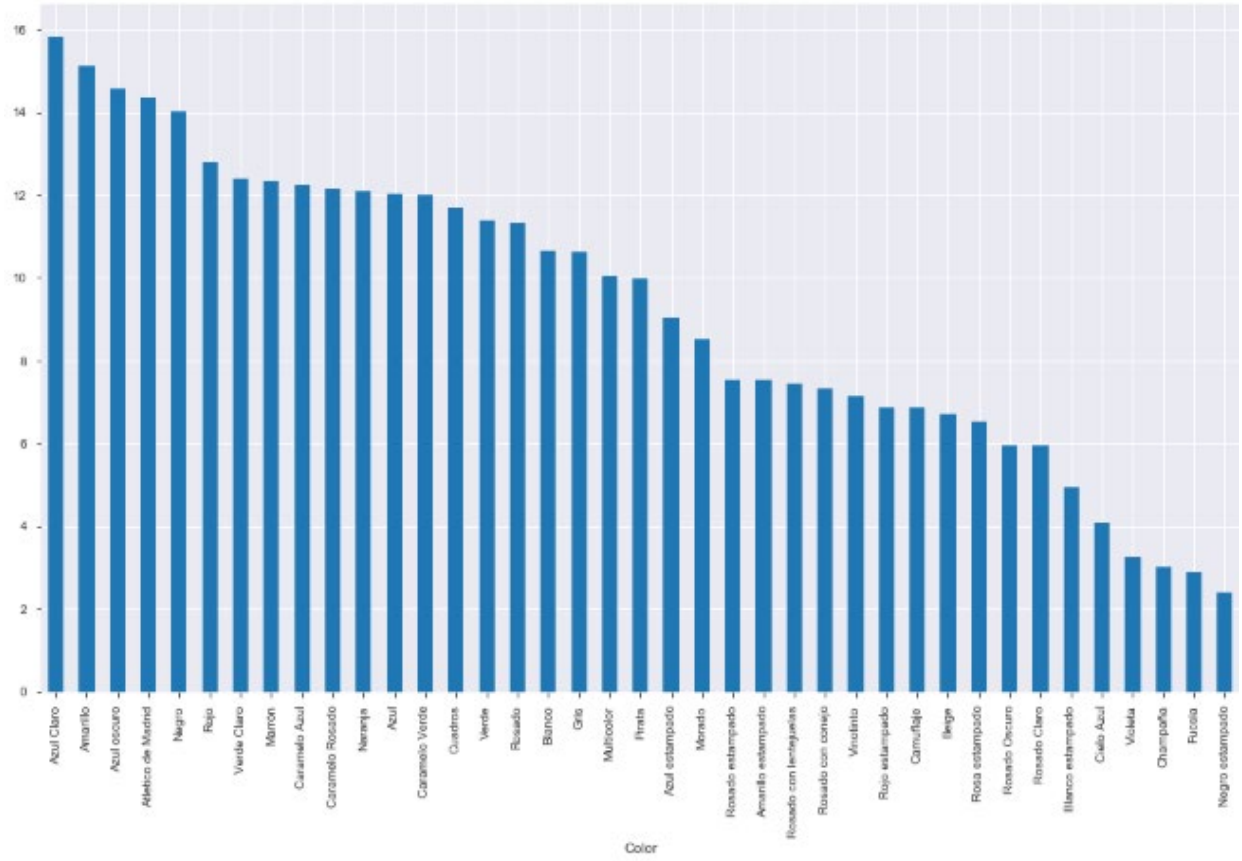
Se puede observar que las tallas de los productos con mayor precio con el 8XL, 6XL y 7XL.

Barplot de Antigüedad del producto y Categoría contra Precio



Los productos más caros han sido los pijamas durante el 2019, siguiéndolo los chubasqueros del 2017 al 2019 y los abrigos en el 2019.

Gráfica de Color contra Precio



Se puede observar que, en promedio los colores con mayor precio son el azul claro, el amarillo y el azul oscuro y los colores con menor precio son el negro estampado, fucsia y champaña.

- ***Preparación de datos***

Se prepararán convirtiendo las variables categóricas en numéricas y también agrupando alguna de ellas para poder llevar a cabo la clasificación.

Se normalizarán los datos si es necesario.

En cuanto a la preparación de los datos, hay muy pocos valores perdidos solamente en las valoraciones de algunos productos. De igual manera, para la predicción del precio se excluyen las valoraciones y para la predicción del posicionamiento se eliminarán los patrones que no tienen valoración.

Para predecir el posicionamiento utilizando regresiones, se realizarán 2 ensayos, incluyendo en el primero únicamente los patrones que tuvieran posicionamiento y en el segundo, se conservarán únicamente los que tienen posicionamiento y se eliminarán los patrones que tienen posicionamientos duplicados para el mismo producto con diferentes características.

Para predecir el posicionamiento utilizando clasificación, se fusionarán todas las columnas de los distintos posicionamientos por categoría en una sola. Posteriormente se aplicará un binning para dividirlo en las siguientes tres clases: Más vendidos, Mediamente Vendidos y Menos Vendidos.

Para la predicción del precio con la regresión, se agruparán ciertos atributos para mejorar las predicciones como, por ejemplo, los distintos tonos de colores en uno, los colores con estampados como estampados y para los productos que tengan más de dos materiales, agruparlos en el material que predomina en el producto. Se utilizará feature selección.

Para la clasificación del precio, se aplicará un binning para dividir al precio en diferentes clases.

- ***Materiales y métodos***

Para predecir los productos más vendidos en Amazon (por su posicionamiento) se utilizará lo siguiente:

Regresión lineal múltiple para predecir el posicionamiento de los productos del 1 al 3000 por categoría, utilizando los modelos Linear Regression, DecisionTreeRegressor, SVM y Adaboost con los métodos de validación Holdout utilizando un 80% de entrenamiento y el 20% de test y Leave One Out. Se escogerá el modelo con el R2 más alto, el cual representa el porcentaje de variación que explica su relación con una o más variables predictoras. Mientras mayor sea el R2, mejor será el ajuste del modelo a sus datos y siempre se encuentra entre 0 y 100%

Si no se obtienen buenos resultados, se llevará a cabo la predicción de la siguiente manera:

Se buscará predecir el posicionamiento en base a la categoría del producto, ya que en Amazon están rankeados por categoría. El primer paso sería predecir la categoría de los productos utilizando el modelo RandomForestClassifier con Leave One Out y como segundo paso, se predecirá el posicionamiento en base a su categoría utilizando el mismo clasificador y método de validación nombrado en el paso anterior. Se evaluarán los modelos con la métrica de exactitud (accuracy score), la cual determina cual es el mejor modelo en identificar relaciones y patrones entre las variables del dataset, basada en la clase predicha o en los datos en entrenamiento.

El segundo objetivo se basa en determinar el precio de los productos en base a los atributos nombrados anteriormente. Se realizará utilizando una regresión lineal múltiple con lo siguiente:

LinearRegression y DecisionTreeRegressor con Holdout.

DecisionTreeRegressor, SVM y con el GradientBoostingRegressor con Leave One Out.

Si no obtiene R2 alto, se utilizarán métodos de clasificación para predecir el precio de la siguiente manera:

Dividiendo el precio en rangos de n números utilizando el método de binning y luego utilizando los clasificadores kNN, Adaboost, Random Forest y Decision Tree con Leave One Out y Holdout utilizando un 80% de los datos para entrenamiento y el 20% para test. Se seleccionaría el modelo con el mayor Accuracy score.

Finalmente, se realizará un clustering con K-means.

- ***Evaluación del modelo***

Predicción del precio:

Para la predicción del precio, se utilizó una regresión lineal múltiple y los resultados del R2 fueron los siguientes:

- Exclusión de las variables del porcentaje de las estrellas ya que están representadas por "Promedio Estrellas (5)" con el modelo de LinearRegression y Hold Out: 0.27
- Exclusión de las variables del porcentaje de las estrellas, con el modelo DecisionTreeRegressor y Hold Out: 0.75.
- Exclusión de las variables del porcentaje de las estrellas con el modelo DecisionTreeRegressor y Leave One Out: 0.72.
- Exclusión de las variables del porcentaje de las estrellas con el modelo SVM y Leave One Out: 0.53.
- Exclusión de las variables del porcentaje de las estrellas con el modelo Adaboost y Leave One Out: 0.57.

	R2 Holdout	R2 Leave One Out
Linear Regression	0.27	-
Decision Tree	0.75	0.72
SVM	-	0.53
Adaboost	-	0.57

Estos resultados confirman que el modelo más apropiado para predecir el precio es el DecisionTreeRegressor con Hold Out, obteniendo un R2 de 0.75.

También se probaron los pasos anteriores con la data set agrupando los productos de distintos tonos de colores en uno, agrupando los colores con estampados como solo estampados, y agrupando los productos de 2 materiales en el que lo predominaba y los resultados no fueron tan buenos como los mencionados anteriormente. Por último, se probó cada modelo eliminando cada una de las variables, pero de igual manera los resultados no mejoraron.

Se utilizó como feature selection el Recursive Feature Elimination y el Lasso Model, los cuales no proveyeron mejores resultados.

Finalmente, cabe destacar que no hiso falta la normalización porque no mejoró ninguno de los R2.

Adicionalmente, se intentó predecir el precio por clasificación aplicando un binning, dividiendo al precio en 22 clases. Se dividió en 22 clases porque nos ha dado mejores resultados quitándole uno a uno, comenzando desde 28, ya que el precio varía desde 2 hasta el 27.

Los resultados del Accuracy de los clasificadores utilizados fueron los siguientes:

- Knn con Leave One Out: 0.66
- Adaboost con Leave One Out: 0.19
- Decision Tree con Leave One Out: 0.80
- Decision Tree con Holdout: 0.71
- Adaboost con Holdout: 0.19
- Knn con Holdout: 0.58

	Accuracy Leave One Out	Accuracy Holdout
Knn	0.66	0.58
Adaboost	0.19	0.19
Decision Tree	0.8	-
Random Forest	0.75	0.71

Se puede observar que el clasificador con el mejor resultado es el Decision Tree con Leave One Out con un 0.80 de Accuracy.

No fue necesario normalizar los datos debido a que no dieron mejores resultados.

Predicción del Posicionamiento (productos más vendidos):

Como se mencionó anteriormente, se realizaron dos ensayos para la predicción del posicionamiento, utilizando regresión lineal múltiple, incluyendo en el primero únicamente los patrones que tuvieran posicionamiento y en el segundo, igual que el primero y se eliminaron los patrones que tenían posicionamientos duplicados para el mismo producto con diferentes características.

El R2 se las regresiones fueron los siguientes:

- Regresión aplicando Hold out con el modelo Linear Regression: 0.10 y 0.02.
- Regresión aplicando Hold out con el modelo DecisionTreeRegressor: 0.14 y 0.39.
- Regresión aplicando Leave One Out con Árboles de Decisión: -0.62 y 0.05.
- Regresión aplicando Leave One Out con Máquinas de Soporte Vectorial (SVM): 0.16 y 0.16.
- Regresión aplicando Leave One Out con Adaboost: 0.24 y 0.16.

Ensayo 1

	R2 Holdout	R2 Leave One Out
Linear Regression	0.1	-
Decision Tree	0.14	-0.62
SVM	-	0.16
Adaboost	-	0.24

Ensayo 2

	R2 Holdout	R2 Leave One Out
Linear Regression	0.02	-
Decision Tree	0.39	0.05
SVM	-	0.16
Adaboost	-	0.16

Luego de estos resultados, se decidió utilizar los métodos de clasificación para predecir el posicionamiento de los productos de la siguiente manera:

Primero, se buscó predecir la clase de las categorías utilizando Leave One Out con el clasificador RandomForestClassifier, excluyendo los atributos de los posicionamientos y los porcentajes de estrellas, dando un **Accuracy** de **0.9928.** Posteriormente, se creó la variable "Producto" en la cual se colocará la lista de los atributos del producto en específico del cual se desea predecir su categoría.

Segundo, se unieron todas las columnas de los distintos posicionamientos por categoría en una sola y se creó un Data Frame filtrando el data set solamente por la Categoría predicha en el paso anterior. Luego se aplica un binning para los posicionamientos de los productos dividiéndolos en 3 clases: "Más vendidos", "Medianamente Vendidos" y "Menos Vendidos". Consecutivamente se utilizó de nuevo el clasificador RandomForestClassifier con Leave One Out, y se excluyeron los atributos del porcentaje de estrellas, obteniendo un **Accuracy** de **0.59375.** Finalmente, el yhat nos dará el posicionamiento predicho del producto en base a su categoría.

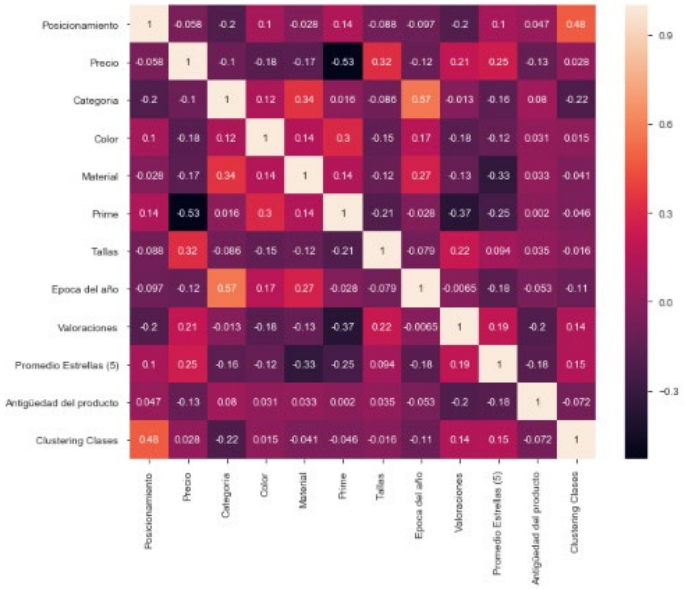
	RandomForestClassifier_LOO_Accuracy
Categoría	0.99
Posicionamiento	0.59

No fue necesario normalizar los datos debido a que no dieron mejores resultados.

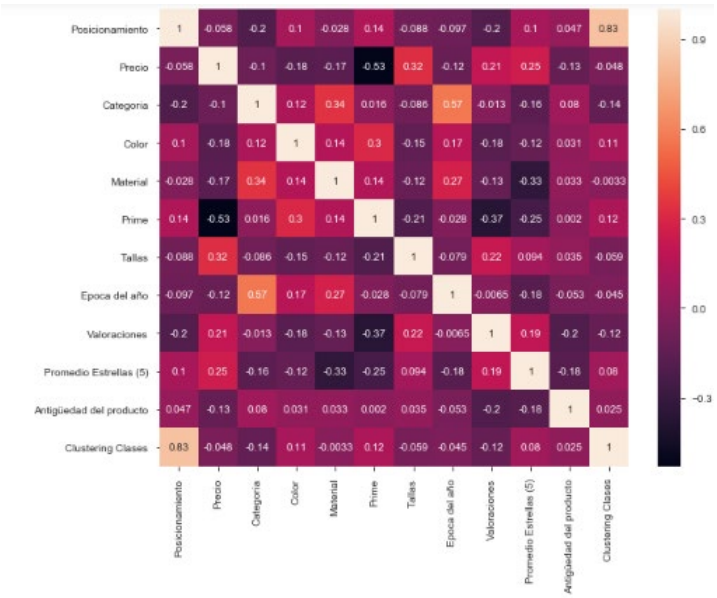
Clustering:

El clustering se realizó con K-means el cual da como resultado lo siguiente:

Agrupando los datos en 4 clúster, se encontró únicamente una correlación con Posicionamiento de 0.48.



Agrupando los datos en 2 clústeres, se encontró solamente correlación con Posicionamiento de un 0.83.



Por lo tanto, se podría decir que los atributos se agrupan por su Posicionamiento. Entonces, puede haber una relación de esto con las tres clases definidas anteriormente para predecir el posicionamiento.

- ***Problemas y soluciones***

El primer problema se presentó al momento de recolectar los datos en Amazon debido a que no se logró hacer web scrapping de los productos por tener características tan específicas y el URL cambiaba por cada patrón del mismo producto por cada característica. Por lo tanto, se han recolectado los datos manualmente.

El segundo problema se basa en que los precios de Amazon cambian constantemente en un mismo día, ya que el algoritmo de Amazon ajusta los precios automáticamente en base a la competencia y este factor no está incluido en los modelos aplicados anteriormente.

El tercer problema consiste en que el ranking o posicionamiento de los productos más vendidos en Amazon también cambia constantemente debido a que esta métrica se determina por el volumen de ventas recientes y los datos de ventas históricos relativos a cualquier otro producto que sea de la misma categoría.

Por lo tanto, las predicciones realizadas son estáticas porque se basan en datos tomados en un tiempo en específico.

- ***Conclusiones:***

Tomando en cuenta que las predicciones fueron realizadas con datos estáticos, se puede concluir lo siguiente:

Los modelos con mejores resultados que predicen el precio son la regresión múltiple aplicando DecisionTreeRegressor con Hold Out con un R2 de 0.75 y el clasificador de Árboles de decisión con Leave One Out dando como resultado un 0.80 de Accuracy.

Para la predicción del posicionamiento, el modelo con mejores resultados se basa en predecir la clase de las categorías utilizando Leave One Out con el clasificador RandomForestClassifier y con esta información predecir el posicionamiento en base a la categoría predicha en el paso anterior. Se utilizó de nuevo el clasificador RandomForestClassifier con Leave One Out, del cual se obtuvo un **Accuracy** de **0.59375.**

Con respecto al clústering, solamente se encontró correlación entre el clúster y Posicionamiento, por lo que los atributos se agrupan por su posicionamiento.

Finalmente, a pesar de obtener buenos resultados provenientes de los modelos nombrados anteriormente, sería importante considerar el precio de los competidores y los posicionamientos en tiempo real para poder realizar predicciones más exitosas. Se podría continuar este estudio aplicando estas consideraciones.

- ***Fases y tiempos:***

Día	Tareas a realizar
10/03/2020	Definir, delimitar y completar la estructura del proyecto.
11/03/2020	Completar estructura del proyecto y recolección de datos.
12/03/2020	Recolección de datos.
13/03/2020	Recolección de datos.
14/03/2020	Recolección de datos.
15/03/2020	Recolección de datos.
16/03/2020	Recolección de datos.
17/03/2020	Recolección de datos.
18/03/2020	Recolección de datos.
19/03/2020	Recolección de datos.
20/03/2020	Preprocesamiento y limpieza de datos.
21/03/2020	Visualización de los datos y estadísticas.
22/03/2020	Construir el modelo de las regresiones lineales y métodos de clasificación para el precio. Evaluación de los modelos.
23/03/2020	Continuación del trabajo del día anterior.
24/03/2020	Construir el modelo de las regresiones lineales y métodos de clasificación para determinar los productos más vendidos. Evaluación de los modelos.
25/03/2020	Continuación del trabajo del día anterior.
26/03/2020	Aplicar el clustering, Conclusiones y elaboración de la Memoria.
27/03/2020	Realizar la presentación en PTT

- ***Referencias***

1. Guptan, R. and Pathak, C. (2014) "Machine Learning Framework for Predicting Purchase by online customers based on Dynamic Pricing". *Procedia Computer Science* 36 (2014) 599 – 605. [Accedido el 28 de marzo, 2019].
2. Shpanya, A. (2013) "5 Trends To Anticipate In Dynamic Pricing". *Retail Touch Points.* [Accedido el 28 de marzo, 2019.]
3. Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., Seaman, B. (2019) *"Sales Demand Forecast in E-commerce using a Long Short-Term Memory Neural Network Methodology*". Lecture Notes in Computer Science, vol 11955. Springer, Cham. [Accedido el 28 de marzo, 2019].
4. Chong, Alain & Ngai, Eric & Ch'ng, Eugene & Li, Boying & Lee, Filbert. (2015). "*Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach*". International Journal of Operations & Production Management. [Accedido el 28 de marzo, 2019].
5. *Hou, Fangfang & Li, Boying & Chong, Alain & Yannopoulou, Natalia & Liu, Martin. (2017). "Understanding and predicting what influence online product sales? A neural network approach. Production Planning & Control"*. 28. 964-975. 10.1080/09537287.2017.1336791.