Aplicación de Inteligencia Artificial para la Identificación de Clientes Potenciales en el Lanzamiento de Nuevos Productos en Arca Continental

Eliezer Cavazos Rochin A00835194 Saúl Francisco Vázquez del Rio A01198261

Facundo Colasurdo Caldironi A01198015 José Carlos Sánchez Gómez A01174050

Instituto Tecnológico de Estudios Superiores de Monterrey Inteligencia Artificial Avanzada para la Ciencia de Datos 28 de noviembre del 2024

Abstract

This project aims to identify potential customers for new products from Arca Continental through the use of artificial intelligence and data analysis. Tools such as PowerBI, Pandas, SKLearn, and TensorFlow were implemented to segment customers, clean and transform data, and train predictive models based on LSTM neural networks. The methodology included creating clusters based on purchasing patterns and socioeconomic characteristics, as well as using advanced techniques like embeddings and Bayesian optimization to improve prediction accuracy. Although the results showed significant progress, areas for improvement were identified to optimize recommendation precision, such as adjusting customer subgroups and refining product cluster filters. The project highlights the importance of artificial intelligence in making strategic decisions for product launches and concludes with proposals to enhance commercial effectiveness and maximize customer satisfaction.

Introducción

En un contexto donde la competencia en el mercado de productos de consumo es cada vez más intensa, Arca Continental, una de las principales embotelladoras de América Latina, enfrenta el reto de identificar de manera eficiente a los clientes potenciales para sus nuevos productos. Este reporte describe el proceso y los resultados de un proyecto desarrollado con el objetivo de aplicar herramientas de análisis de datos e inteligencia artificial para segmentar a los clientes y predecir su comportamiento de compra. Utilizando tecnologías avanzadas como redes neuronales y algoritmos de agrupamiento, se busca optimizar las estrategias de marketing y ventas, mejorando la precisión en la identificación de los consumidores más propensos a adquirir productos recién lanzados.

A través de un enfoque metodológico que abarca desde la exploración y limpieza de datos hasta la implementación de modelos predictivos, el proyecto busca proporcionar una solución que permita a Arca Continental tomar decisiones más informadas y basadas en datos. A lo largo de este reporte, se detalla el proceso seguido, las herramientas empleadas, y los resultados obtenidos, destacando tanto los avances logrados como las áreas de oportunidad

para futuras mejoras en la predicción y segmentación de clientes.

Planteamiento del problema

Arca continental planteó el reto de poder identificar grupos de clientes cuyas preferencias se alinean con las características de un producto de lanzamiento, optimizando estrategias de marketing e indicando si el producto de lanzamiento será exitoso o no.

Objetivo

Identificar potenciales clientes cuyas preferencias se ajusten a las características de los productos de lanzamiento.

Sub-objetivos

- Segmentar a los clientes en función de sus patrones de compra, permitiendo una mayor personalización de estrategias de marketing y ventas.
- Agrupar los productos según su rendimiento en el mercado, con el fin de optimizar la oferta y la

disponibilidad de los mismos en función de su demanda.

Objetivo

Identificar potenciales clientes cuyas preferencias se ajusten a las características de los productos de lanzamiento.

Sub-objetivos

- Generar modelos de predicción para predecir la venta de un producto nuevo de manera general
- Usando los clusters de clientes por nivel socioeconómico predecir la venta de un nuevo producto por cada cliente, para ver los comportamientos de este nuevo producto en los diferentes clientes que se tienen.

Antecedentes y proyectos relacionados

La inteligencia artificial (IA) ha transformado múltiples sectores, desde el comercio minorista y el entretenimiento hasta la salud y el transporte, facilitando el análisis y la predicción de patrones de comportamiento de los consumidores. No obstante, el uso de IA también plantea desafíos éticos y normativos, particularmente en cuanto a la privacidad, la transparencia de los modelos predictivos y el manejo responsable de los datos.

La aplicación de IA en el análisis de consumo y ventas requiere una consideración cuidadosa de estos aspectos, ya que el uso de datos personales y de consumo debe ajustarse a regulaciones de protección de datos, como el Reglamento General de Protección de Datos (GDPR) en Europa y sus equivalentes en América Latina.

Desde una perspectiva ética, el uso de la IA tiene grandes responsabilidades esto debido a que implica que se manejen los datos de manera segura y que se cumplan con las normativas internacionales de seguridad, ademas, demás, es fundamental promover la transparencia algorítmica, explicando a los usuarios cómo se utilizan sus datos y cómo funcionan los modelos de predicción, ya que esto no solo genera confianza en los usuarios, sino que también evita posibles malentendidos en el uso de la información.

Durante el desarrollo del proyecto se decidió que se va a segmentar los clientes por patrones de compra para facilitar la recomendación de productos de lanzamiento parecidos a su patrón de compra y también segmentar los productos por sus características principales. Entre las herramientas relevantes para realizar esta segmentación se encuentran los algoritmos de clustering, como el k-means, en donde el uso de k-means ha demostrado su eficacia en la identificación de patrones clave en el comportamiento del consumidor dentro del sector minorista, permitiendo personalizar estrategias de marketing como se describe en un análisis práctico de segmentación disponible en la investigación realizadas por la analista Jennifer salazar: "Por otro lado, se considera de suma importancia efectuar una segmentación de clientes buscando detectar aquel segmento que genera mayor rentabilidad para el Banco. "

Asimismo, los modelos de clasificación, como los árboles de decisión y las redes neuronales, ayudan a predecir las probabilidades de que un cliente opte por un producto específico, proporcionando insights valiosos sobre sus preferencias. El análisis de patrones secuenciales es otra técnica que permite estudiar el orden y frecuencia de las compras, apoyando así en la creación de recomendaciones más efectivas.

En el contexto específico de este proyecto, la IA será utilizada para analizar patrones de consumo de los clientes de Arca Continental y de esta forma, apoyar la toma de decisiones estratégicas que maximicen las ventas de productos de lanzamiento.

Para entender de mejor maner a Arca Continenta, es necesario comprender que es una de las embotelladoras más grandes del mundo, la cual se caracteriza por usar estrategias avanzadas de mercado para sus productos, según el Reporte de Estrategias de Segmentación 2023, la empresa aprovecha los datos de consumo para diseñar campañas altamente efectivas. A través de la segmentación de clientes y el agrupamiento de productos según su rendimiento en el mercado, el proyecto busca maximizar las ventas de nuevos productos.

Metodologia

Herramientas y recursos

Para el desarrollo del proyecto se utilizaron una gran variedad de herramientos que no solo facilitaron el análisis de los datos, sino que también, ayudaron a la construcción de los modelos predictivos, estas siendo: PowerBi, Pandas, Matplotlib, Tensor Flow, SKLearn, google colab, drive, los conjunto de datos dados por Arca Continental

Power bi fue utilizado durante la exploración inicial para poder mostrar los resultados de nuestra investigación de una manera más visual y sencilla. Por otra parte, Pandas fue una herramienta esencial para lograr analizar los datos durante el proyecto, ya que esta nos permite trabajar con los datos dentro

de python que nos permiten una mejor interacción con los datos de Arca Continental, logrando así eliminar datos innecesarios y logrando manejar grandes cantidades de información de manera eficiente.

SkLearn es una librería de python la cual usamos para la clusterización de los datos tanto de los clientes como de los productos para de tal forma que pudiéramos agrupar y segmentar por los parecidos que habían entre los registros. De igual manera, se utilizó esta herramienta para hacer la codificación One-Hot de ciertas variables que utilizamos para el modelo de predicción.

Tensor Flow fue utilizado con el objetivo de crear modelos predictivos, en especifico redes neuronales, dentro del proyecto. Esto con la finalidad de poder contar con una herramienta que nos pudiera decir cuanto iba a comprar un cliente de cierto producto en un determinado mes, y así determinar si un producto es exitoso o no. Google Colab y Google Drive, fueron de vital importancia, ya que nos permitieron almacenar los datos y los resultados del proyecto, a su vez, que nos facilitaron la carga de trabajo al permitir a todos los miembros del equipo poder trabajar de una manera más organizada; además de esto Google Colab, nos permitió hacer el entrenamiento del modelo usando GPUs, lo cual aceleró este proceso.

Exploracion de los datos

Cada miembro de equipo fue explorando los datos proporcionados por Arca Continental, lo que más nos llamó la atención fue el conjunto de datos de clientes ya que este contenía muchas tablas con diferentes tipos de información, como los niveles socioeconómicos de las personas que viven a 300 metros alrededor del cliente, las zonas que estaba cerca del cliente en unos 300 metros, también se mencionaba establecimientos cerca del cliente como parques, supermercados, escuelas, hospitales, por otra parte, habían indicadores de la movilidad de autos alrededor de la zona, al mismo tiempo, se encontraba indicadores que mencionaba el tipo de cliente y más indicadores con información demográfica.

Una vez que terminamos de explorar los datos de clientes nos surgieron varias preguntas como ¿Necesitamos tanta información para el proyecto?, ¿Podemos quitar columnas de los datos?, ¿Qué información afecta más la compra de productos?, etc. La mayoría de estas preguntas fueron contestadas tanto por el socio formador y por nuestra investigación que nos indicaron hacia qué camino ir. Una vez establecidas las preguntas exploramos los datos de ventas y de productos, pero en estos no tuvimos tantas dudas por la simplicidad de las tablas.

Limpieza de los datos

Para nuestra limpieza de datos, decidimos seguir pasos para los tres conjunto de datos que se nos proporcionaron está siendo.

- Resaltar variables que creíamos importantes
- Buscar valores con NaN
- Buscar relaciones entre los conjunto de datos

Primeramente se empezó en los datos de los clientes resaltando las variables de sub_canal_economico, población alrededor de 300 m, estatus económico, zonas alrededor de 300m y los nivel de ocupación de las personas alrededor. Posteriormente se buscaron los valores que tuvieran valores nulos donde se obtuvieron 73 clientes con algunos datos vacíos que decidimos rellenar con valor numérico de 0.

Empezamos a relacionar las tablas de ventas y clientes, donde nos dimos cuenta que había clientes que no estaban en los datos de ventas. Esto indicaba que no habían realizado una compra y decidimos borrar estos valores para que no afectará al análisis, también decidimos borrar clientes con compras a partir de septiembre de 2022 ya que al igual que los clientes sin compra no iban a servir para el análisis final.

Continuamos relacionando la tabla de ventas y productos donde identificamos que habían productos en nuestra tabla de ventas que no existían en nuestros datos de producto, con esto decidimos también borrar los datos de ventas donde no hubiera una relación con los productos.

Para la selección de las variables importantes en el dataset de productos se decidieron usar el tipo de producto, envase, tamaño, sabor, cantidad de productos por envase y cantidad en mililitros por envase. La razón del por que seleccionamos esas variables fue gracias a que en nuestro análisis identificamos cuáles eran las variables que más afectan a la venta de un producto, tanto en unidades como en continuidad.

Continuando con la selección de variables pero ahora en el dataset de clientes decidimos usar las variables del conteo de personas por nivel socioeconómico alrededor del clientes, ya que de los datos demográficos era el que más afectaba la compra de productos de ciertas características.



Figure 1: Correlacion Nivel Sociecono de clientes con venta

Transformación de datos

Se optó por un modelado estrella de datos para dimensionar las variable categóricas y características tanto de clientes como de productos en diferentes tablas, ya que al momento de dimensionar las tablas da una mejor transparencia de los datos, además de optimizar el uso de los datos al no tener que manejar todas las variables, sino que se va decidiendo qué variables usar en base a lo que se esté buscando en nuestras consultas de datos.

Siguiendo la idea de dimensionamiento se generaron diferentes dimensiones para clientes usando las características de zonas alrededor, conteo de personas por nivel socioeconómico alrededor y tipo de cliente esto para tener datos categóricos de estas variables y poder analizar si se puede usar para un futuro modelo de predicción con estas variables. Para la tabla de productos se dimensionan las columnas de marcas, contenedor del producto, tamaño del envase, rentabilidad, la categoría del producto y el tipo del producto para simplificar su uso en consultas de los datos de productos.



Figure 2: Modelado Estrella

Después del dimensionamiento el segundo paso en nuestro modelado de datos es la generación de tablas de hechos, que al final serán las tablas que usaremos para nuestro análisis y clusterización de clientes. Se generaron 5 tablas de hechos, empezando con las tablas de hechos de clientes se generaron 4 tablas, la primera tabla hace referencia a la cantidad que un cliente compra de productos de cada categoría, nuestra segunda tabla identificaba cuantos productos distintos compraba un cliente por categoría de producto, después nos dimos cuenta que el uso de la variable del "tipo de producto" se ampliaba más la cantidad de categorías que la variable de "categoría de producto", así que decidimos generar las mismas dos tablas pero ahora con el uso de "tipo de producto" que con la "categorías de producto". Nuestra última tabla de hecho la generamos para identificar los productos exitosos por cada cliente, usando la definición que nos dio Arca continental donde cada cliente compra el mismo producto durante 5 meses

seguidos, además que su uso también funciona para identificar la cantidad de mejores clientes por cada producto.

Selección y entrenamiento de modelos

Se realizó una clusterización de clientes y productos, esto nos ayudó a determinar en que cluster iban a ir los productos de lanzamiento y también a poder agrupar a los clientes por sus patrones de compra. El cluster de productos siendo el primer filtro que determina en qué categoría irá el producto de lanzamiento, en este cluster se usaron las columnas de contenedor, tipo del producto, el tamaño de su empaque, sabor, productos por empaque, mililitros por envase y retornabilidad. Usamos estas variables porque, tras realizar múltiples pruebas con diferentes combinaciones de variables, estas fueron las que nos ofrecieron mejores resultados. Esto nos llevó a definir un total de diez clusters para categorizar los productos, determinamos este número después de probar con diferentes cantidades de clusters y observar que con diez se logró la mejor distribución de los produc-

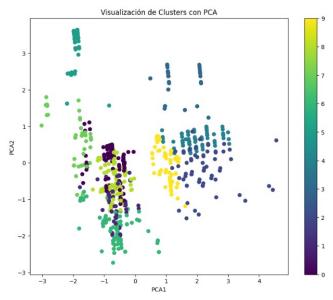


Figure 3: Clusters de productos

Para la clusterización de clientes se optó hacerlo con base a su nivel socioeconómico usando las columnas de conteo de personas por nivel socioeconómico a 300 metros teniendo un total de tres clusters (nivel alto, medio y bajo). Una vez generados los clusters iniciales y separar a los clientes por valores demográficos, decidimos hacer sub clusters en donde se muestran los patrones de compra de cada grupo de clientes viendo cuáles productos compran más y cuales productos compran variedad de esté usando nuestras tablas de hechos de compra por tipo de

producto. Para decidir cuántos sub clusters hacer decidimos a prueba y error en donde los clientes se separaran de mejor manera y terminamos con 12 subclusters por cluster de cliente terminando con 36 clusters y subclusters finales.

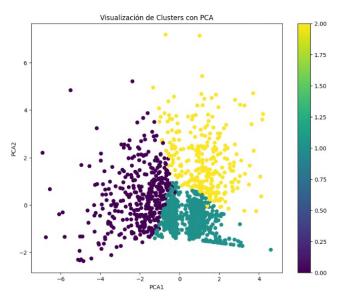


Figure 4: Clusters de clientes

Se segmentaron los clusters creados previamente con el objetivo de analizar en detalle tanto los productos como los clientes agrupados en cada uno. Esto permite identificar con mayor precisión los elementos que conforman cada cluster y comprender las características principales de los grupos generados.

Predicción de Clientes Potenciales para nuevo producto

Para la predicción, utilizamos la librería Tensor Flow para crear modelos de inteligencia artificial que sean capaces de predecir la venta de un nuevo producto con respecto a un cliente en los próximos cinco meses, además utilizando los clusters determinar los clientes que más comprarían este producto, dependiendo del cluster de producto donde entre el nuevo producto, y también analizar cómo se desempeñaría la venta del producto dependiendo del nivel socioeconómico.

Estos modelos son redes neuronales, las cuales cuentan con capas de Dropout y BatchNormalization para garantizar que el modelo pueda abstraer la información del conjunto de datos y no se sobreajuste. Para la función de pérdida se usa la función huber, la cual es menos sensible a los datos atípicos que la de error cuadrado promedio. De igual manera usamos las funciones EarlyStopping y ReduceLROnPlateau para asegurarnos que el modelo pudiera obtener un valor menor en la pérdida de los datos de validación.

La entrada del modelo tiene un formato diferente

al que tienen nuestros set de datos. En un inicio intentamos utilizar la técnica de OneHotEncoding para poder usar nuestras variables categóricas sin darles un peso mayor a una que la otra. Sin embargo, nos percatamos de que este acercamiento no iba a ser eficiente, ya que nuestro modelo terminaría recibiendo 147 datos por entrada. Para resolver esto, optamos por usar embeddings para cada columna, esto redujo la cantidad de columnas en un 58.5

Llegamos a la solución de la problemática creando una LSTM, para predecir la venta promedio en un mes que va a tener un producto con un cliente en específico. En un inicio se tenía pensado usar una red neuronal con capas Dense, y funciones de activación de ReLu.

Layer (type)	Output Shape	Param #
dense_12 (Dense)	(None, 32)	1,984
dropout_2 (Dropout)	(None, 32)	Θ
batch_normalization_2 (BatchNormalization)	(None, 32)	128
dense_13 (Dense)	(None, 8)	264
dropout_3 (Dropout)	(None, 8)	Θ
batch_normalization_3 (BatchNormalization)	(None, 8)	32
dense_14 (Dense)	(None, 4)	36
dense_15 (Dense)	(None, 1)	5

Figure 5: Arquitectura de la NN con capas Dense

Sin embargo, los resultados que obtuvimos no eran tan acertados como los deseábamos, por lo que decidimos transicionar a una LSTM que pudiera entender la temporalidad de las ventas. Este nuevo modelo generó mejores resultados, sin embargo no son tan precisos como nos hubiera gustado.

Layer (type)	Output Shape	Param #
1stm (LSTM)	(None, 12, 64)	32,256
dropout (Dropout)	(None, 12, 64)	Θ
batch_normalization (BatchNormalization)	(None, 12, 64)	256
lstm_1 (LSTM)	(None, 12, 32)	12,416
dropout_1 (Dropout)	(None, 12, 32)	0
batch_normalization_1 (BatchNormalization)	(None, 12, 32)	128
lstm_2 (LSTM)		8,320
dropout_2 (Dropout)	(None, 12, 32)	Θ
batch_normalization_2 (BatchNormalization)		128
lstm_3 (LSTM)	(None, 16)	3,136
dropout_3 (Dropout)	(None, 16)	Θ
batch_normalization_3 (BatchNormalization)	(None, 16)	64
dense (Dense)	(None, 1)	17

Figure 6: Arquitectura de la NN LSTM

La entrada para la LSTM consiste en un vector de esta forma: (n, 6, 61). Dónde no es la cantidad de

registros, cada registro cuenta con una secuencia de longitud seis (6) con 61 datos cada una. La razón por la que usamos seis como nuestra longitud de la secuencia fue por un tema de recursos de cómputo. Al estar trabajando en Google Colab, este terminaba la ejecución si se excede del límite de RAM, que con una secuencia mayor sucedía. Especulamos que de ser posible, si se utiliza una secuencia de longitud 12, se podrían obtener mejores resultados, ya que el modelo tendría un contexto mayor de la temporalidad de los datos.

Este modelo se entrenó con 810,908 registros que son las ventas de cada producto iniciando desde enero de 2020, hasta junio de 2022. Y se hicieron pruebas con 1,344 que son las ventas de los últimos seis meses, esto a petición del socio formador. Las variables que se usaron para entrenar al modelo fueron las siguientes:

- Sabor del producto
- Marca del producto
- Mililitros del producto
- Contenedor del producto
- Cantidad de productos en un envase
- Retornabilidad del producto
- Categoría del producto
- Tipo del producto
- Mes de la compra
- Espacios recreativos alrededor de 300 metros del cliente
- Conteo de personas cerca del cliente
- Viviendas alrededor de 300 metros del cliente
- Accesibilidad de la zona

Esto con el objetivo de predecir la compra en uni_box (galones) de un producto en un mes determinado por un cliente en específico.

De igual manera, se implementó durante la metodología para la elaboración del modelo una optimización bayesiana, la cuál busca los mejores hiper parámetros para el modelo con base a los resultados de iteraciones pasadas. Esto con el objetivo de asegurarnos de que el modelo final que usaramos fuera el que devolviera mejores resultados.

Comparamos ambos modelos para averiguar cuál de los dos era el más óptimo, y concluimos que el mejor es el de LSTM.

Modelo	MSE	Mejor val_loss
NN Dense	16.8679	5.2275
NN LSTM	15.2819	4.5455

Table 1: Resultados de los modelos entrenados.

Resultados

Los resultados que nos han dado nuestras predicciones de clientes recomendados ha ido mejorando pero aún no está dando los resultados esperados. Usando los productos de pruebas que se nos proporcionaron donde se podrá ver el comportamiento que tienen nuestras predicciones en la siguiente tabla comparativa, además dejamos de considerar lo clientes exitosos para identificar las métricas ya que muchas veces estos valores no contaba muchos clientes que compraban mucha cantidad de estos productos:

Producto	Clientes	Clientes	Clientes	Clientes Exi-
	Reales	Exitosos	Prop-	tosos que co-
		Reales	uestos	inciden
MONSTER	176	Solo	Se propo-	Ninguno co-
ENERGY 473	Clientes	hay un	nen 471	incide
ML NO RE-	com-	cliente	clientes	
TORNABLE	praron	exitoso		
	este pro-	de este		
	ducto	producto		
	en los			
	últimos 5			
	meses			
TCH HARD S	126	Solo	Se propo-	Ninguno co-
LIMA I 355	clientes	hay un	nen 471	incide
ML NR LAT 6	com-	cliente	clientes	
REFO	praron	exitoso		
	este pro-	de este		
	ducto	producto		
	en los			
	últimos 5			
	meses			
COCA COLA	295	No tiene	Se propo-	113 coinciden
S/A MARS	clientes	ningún	nen 471	
355ML NR	com-	cliente	clientes	
LSL 6B	praron	exitoso		
	este pro-			
	ducto			
	en los			
	últimos 5			
	meses			

 Table 2: Resultados de las predicciones anteriores.

Producto	Clientes Reales	Clientes Exitosos Reales	Clientes Prop- uestos
MONSTER ENERGY 473 ML NO RE- TORNABLE	176 Clientes com- praron este pro- ducto en los últimos 5 meses	Solo hay un cliente exitoso de este producto	Se propo- nen 491 clientes
TCH HARD S LIMA I 355 ML NR LAT 6 REFO	clientes com- praron este pro- ducto en los últimos 5 meses	Solo hay un cliente exitoso de este producto	Se proponen 486 clientes
COCA COLA S/A MARS 355ML NR LSL 6B	295 clientes com- praron este pro- ducto en los últimos 5 meses	No tiene ningún cliente exitoso	Se proponen 486 clientes

Table 3: Resultados de las predicciones nuevas.

Con estas métricas podemos identificar que el modelo aún se tiene que mejorar para que los Clientes que se recomienda coincidan más con lo real. Las métricas de evaluación que estamos usando es identificar cuántos clientes recomendados por el modelo coinciden con la venta real del producto. Nuestro filtro final de los clientes recomendado donde quedamos con los mejores 50 clientes de un producto, normalmente se puede esperar que el 30% de estos terminen dentro del top 20 de los clientes reales mientras que el otro 70% de los clientes recomendados serán clientes que no se tuvieron en los datos reales pero que se pueden considerar por tener patrones de compra similares.

Conclusiones

En este proyecto se buscó implementar herramientas de analitica de datos y herramientas de inteligencia artificial para identificar clientes potenciales para nuevos productos de Arca Continental.

Aunque las predicciones muestran clientes potenciales para los nuevos productos, la predicción no es tan satisfactoria tenemos en mente mejoras que podrían ayudar a la mejora de este modelo de predicción de clientes potenciales. Cómo optimizar los sub clusters para generar mejores agrupaciones, mejorar el filtrado de los clientes por cluster de producto.

El uso de las herramientas como PowerBI, Pandas, SKLearn y Tensor Flow demostraron ser esenciales para la limpieza, análisis, segmentación y modelado de datos. Estas herramientas ayudaron para el procesamiento de grandes volúmenes de datos. En conclusión, en este proyecto se nos mostró la importancia de la inteligencia artificial para la toma de decisiones estratégicas para el lanzamiento de productos. Sin embargo quedan mejoras para realizar para optimizar las predicciones y garantizar que estas sean más precisas y efectivas. Implementar estas mejoras propuestas nos permitirá realizar una solución que maximice la rentabilidad y satisfacción de los clientes con los productos de lanzamiento.

Referencias

- Arca. C. (2024) Balance General de Arca Continental. Recuperado de https://www.arcacontal.com/media/373544/2020 _ac_consolidated_financial_statements.pdf
- SalesForm. C.(2020) Clústeres: ¿qué son y para qué sirven?. Recuperado de https://www.salesforce.com/mx/blog/clusteres/