**1. Entendimiento de negocio y procesamiento de información**

La empresa enfrenta actualmente un desafío en sus procesos, caracterizado por tiempos de ejecución prolongados debido a la naturaleza secuencial de las operaciones. Por ejemplo, se requiere enviar los datos pertinentes a cada área para que realicen sus respectivas tareas y extraigan sus indicadores individuales. Esta metodología se vuelve más problemática a medida que aumenta el número de clientes, ya que se incrementan los tiempos de extracción de datos, la carga en los sistemas para procesar la información de cada área y el tiempo dedicado al análisis correspondiente. Esta acumulación de factores contribuye a una complejidad creciente en términos de recursos computacionales y tiempo requerido por cada tarea, lo que eventualmente resulta en procesos más lentos y menos eficientes.

Para abordar este desafío, se pueden implementar varias acciones complementarias:

1. Establecer una conexión directa entre las bases de datos almacenadas en los sistemas ERP y aprovechar los entornos de procesamiento en la nube, como las soluciones disponibles en Google Cloud Platform (GCP), AWS o Azure.
2. Desarrollar procesos automatizados de extracción, transformación y carga (ETL y ELT) que estén integrados con los datos de los sistemas ERP, con el objetivo de crear un entorno más eficiente y que permita consultas en tiempo real, adaptándose a los cambios de manera dinámica.
3. Emplear herramientas de almacenamiento en un datalake o un almacén de datos (como Google Cloud Storage en GCP, Amazon S3 en AWS o Azure Data Lake), que simplifiquen tanto el acceso a los datos como la implementación de procesos automatizados para análisis recurrentes, adaptándose dinámicamente a las fluctuaciones del mercado.

Estudio de modelos:

1. Para el estudio de modelos, dado que la información general ya se encuentra almacenada en el almacén de datos junto con sus transformaciones para su uso, es posible desarrollar pipelines de modelado que abarquen desde la limpieza inicial de los datos hasta el proceso de ingeniería de características. Esto incluye la generación de nuevas variables, la eliminación de variables correlacionadas y la detección y eliminación de outliers, todos estos pasos adaptados al planteamiento específico del problema. Estos pipelines permiten una gestión eficiente y sistemática de todo el proceso de construcción y entrenamiento de modelos predictivos o analíticos.
2. Clasificar los problemas según su naturaleza, ya sea clasificación o regresión, y crear pipelines que se adecuen a la solución deseada. Esto implica diseñar flujos de trabajo específicos que aborden las particularidades de cada tipo de problema, desde la preparación de los datos hasta la evaluación del modelo. Además, es esencial incorporar técnicas apropiadas de preprocesamiento, selección de características y ajuste de hiperparámetros, con el fin de obtener resultados óptimos y precisos para la tarea en cuestión.
3. Para la automatización de procesos, es viable emplear herramientas en la nube que permitan el entrenamiento, la implementación y el monitoreo de los modelos propuestos. Esta estrategia posibilita la creación de sistemas automatizados y escalables, lo que conlleva a una mayor eficiencia operativa y la capacidad de adaptarse dinámicamente a las demandas cambiantes del entorno empresarial.

**Técnicas de modelación:**

Las técnicas de modelación a ejecutar van en pasos:

1. Limpieza de datos: Para asegurar la escalabilidad de un modelo, resulta fundamental llevar a cabo una exhaustiva limpieza de los datos utilizando herramientas como SQL, Pandas o Spark en Python. Este proceso implica estandarizar la información, eliminar valores irrelevantes y transformar las variables en formatos comprensibles para su análisis. En casos de desbalanceo de datos, se pueden implementar técnicas específicas como undersampling u oversampling, así como aplicar SMOTE o resampling para mitigar cualquier sesgo en las conclusiones de nuestros modelos. Además, es esencial realizar una validación cruzada adecuada y una evaluación continua del desempeño del modelo para garantizar su eficacia y fiabilidad en diferentes escenarios.
2. Planteamiento del modelo:

Considerando las necesidades particulares de cada área, podemos implementar diversos modelos que se ajusten a los objetivos específicos. Algunas alternativas a tener en cuenta incluyen:

* Técnicas de clusterización, como K-Means o DBSCAN, que posibilitan la segmentación de clientes, permitiendo así la personalización de estrategias de marketing y ventas en función de características similares entre grupos de clientes.
* Modelos de regresión, idóneos para analizar series temporales de productos. Estos modelos permiten identificar tendencias, detectar patrones y realizar predicciones futuras sobre el comportamiento de políticas, estrategias de marketing y ventas. Además, proporcionan insights valiosos para la toma de decisiones estratégicas basadas en datos.

**Tecnologías para utilizar:**

Los lenguajes de programación adecuados para plantear la solución del problema pueden incluir SQL, Python, Java o R. Entre ellos, Python destaca como una opción especialmente versátil y ampliamente utilizada en el ámbito del análisis de datos y modelado. Al emplear Python, es posible integrar una variedad de bibliotecas especializadas como Pandas, NumPy, Matplotlib, Seaborn, PySpark, Scikit-learn, TensorFlow o Keras. Estas herramientas facilitan el procesamiento, la visualización, el entrenamiento y la validación de modelos, permitiendo así un enfoque integral y eficiente para resolver los desafíos analíticos y de modelado de datos.

**Visualización:**

Con el fin de crear visualizaciones comprensibles durante el planteamiento de modelos, es recomendable utilizar cuadernos de Jupyter, que permiten incorporar descripciones gráficas con explicaciones directamente en el código del modelo. Además, para generar gráficos explicativos y tableros de control automatizados, se pueden emplear herramientas como Power BI, Tableau o BigQuery (en GCP). Estas herramientas son ideales para representar de manera clara y concisa los indicadores clave de rendimiento (KPI) de las diferentes áreas, proporcionando una visión global y detallada del desempeño del negocio.

**Metodologías:**

Una de las metodologías mas utilizadas y la cual podemos implementar para la solución del problema es CRISP-DM, la cual se divide en varios pasos:

1. Entendimiento del negocio

2.Entendimiento de los datos

3.Preparación de datos

5.Modelamiento

6.Evaluación

7.Despliegue

8.Monitoreo y mantenimiento

Si planteamos la anterior metodología nos permitirá una mejora continua y podremos adaptar el modelo según las necesidades del negocio.

**2. Entendimiento técnico Marketing Mix Modeling**

Dado el planteamiento en marketing como se desean modelos analíticos que mejoren la toma de decisiones de inversión, es ideal considerar todas nuestras ecuaciones de optimización, en las cuales observamos como se comporta la rentabilidad del producto. De este modo podemos iniciar con planeamiento de regresión para observar la tendencia de los presupuestos gastados.

De este modo una de las técnicas que aplica en la modelación Robin es Ridge Regression, la cual plantea el resultado de los indicadores mediante una expresión polinomial de grado 1, la cual permite considerar todas las variables, pero aquellas que presenten alta colinealidad o una contribución muy baja, las penaliza reduciendo en magnitud para su contribución en el resultado final. Para controlar esta penalización se considera el parámetro lambda, que permitirá limitar la penalización que se genere en cada variable, de este modo los indicadores o nuestras ecuaciones de optimización estarán planteados con el fin de evitar overfitting en nuestro modelo y reduce el sesgo humano.

Otra forma de optimizar nuestro modelo, esta basado en la selección del modelo planteando y modificando varios hiperparametros, de este modo Robin tiene la gran ventaja de realizar un modelo semi-automatico para selección de los mejores resultados. De este modo la forma de medir y ver si el modelo esta ajustado para la solución del problema se basa en el estudio de los errores que genera cada modelo y sumado a los tiempos que gasta el entrenamiento de los modelos.

Dado todo lo anterior es importante entender las variables que ingresaran en nuestro modelo, para su entrenamiento. Por lo que para calibrar el modelo de Marketing Mix Modeling debería plantearse:

1. Realizar estudios de incrementalidad y asegurar la alineación de los resultados se den en los diferentes términos, como granularidad y buenas métricas medidas en los mismos tiempos.
2. Recopilar los resultados de estudios de incrementalidad de acuerdo con la fecha de inicio, fecha de finalización y estudio incremental, para el entendimiento y estudios posteriores.
3. Se ingresan los resultados en el modelo Robyn como entradas de calibración del modelo.
4. Se utiliza la plataforma de optimización Nevergrand de META, para incluir el MAPE (minimizar el error absoluto porcentual medio) como otro medida de optimización.
5. Nevergrand Nevergrad generará un conjunto de soluciones de modelos óptimos de Pareto que minimizan y convergen a un conjunto de candidatos de modelos óptimos de Pareto, teniendo en cuenta los tres puntajes de optimización: NRMSE, decomp.RSSD, MAPE.
6. Por último se evalúan los modelos más óptimos y se selecciona el que mejor se ajuste al resultado de estudios incrementales.
7. El proceso se realiza de nuevo a medida que se sigan agregando nuevos estudios de incrementalidad.

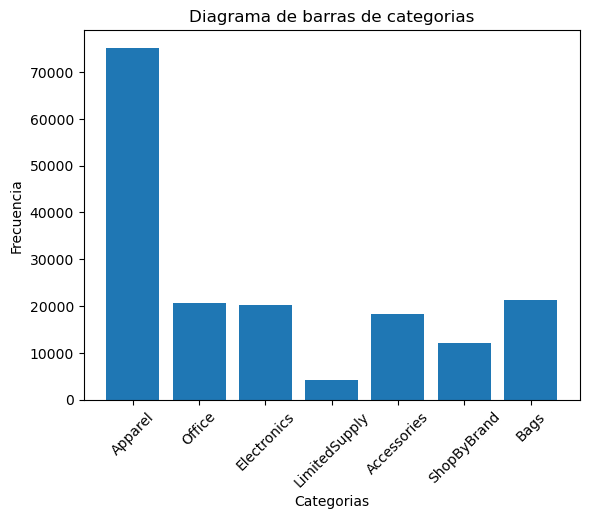
**3. Ejercicio de Clustering:**

*El cuaderno de Jupyter se encuentro en el archivo* [***3.Ejercicio de Clustering.ipynb***](http://localhost:8888/notebooks/3.Ejercicio%20de%20Clustering.ipynb)*, ubicado en la ruta* [*https://github.com/ajrianop/Revision1*](https://github.com/ajrianop/Revision1)*. De igual modo se adjunta el PDF del notebook con las pruebas y código.*

Dada la información del comportamiento de clientes en e-commerce, se desea dividir por lo menos en tres segmentos diferentes, los clientes registrados. De este modo, se revisan las variables, que son:

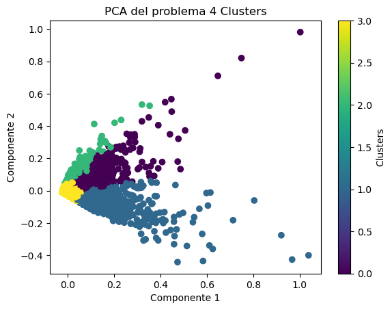
* fullVisitorID: Identificador del cliente
* channelgrouping: Medio por el cual llegó al sitio web
* date: dia
* OS: Tipo de dispositivo con el que entro al sitio web
* Apparel, Office, Electronics, LimitedSupply, Accessories, ShopByBrand, Bags: Categorias de los productos comprados
* totalSpent\_USD: Total gastado

De estas variables para observar los cluster, se consideraron las variables númericas. Que son Apparel, Office, Electronics, LimitedSupply, Accessories, ShopByBrand, Bags y totalSpent\_USD, de manera inmediata Podemos observer que la mayor cantidad de productos comprados se dio para la categoría Apparel.

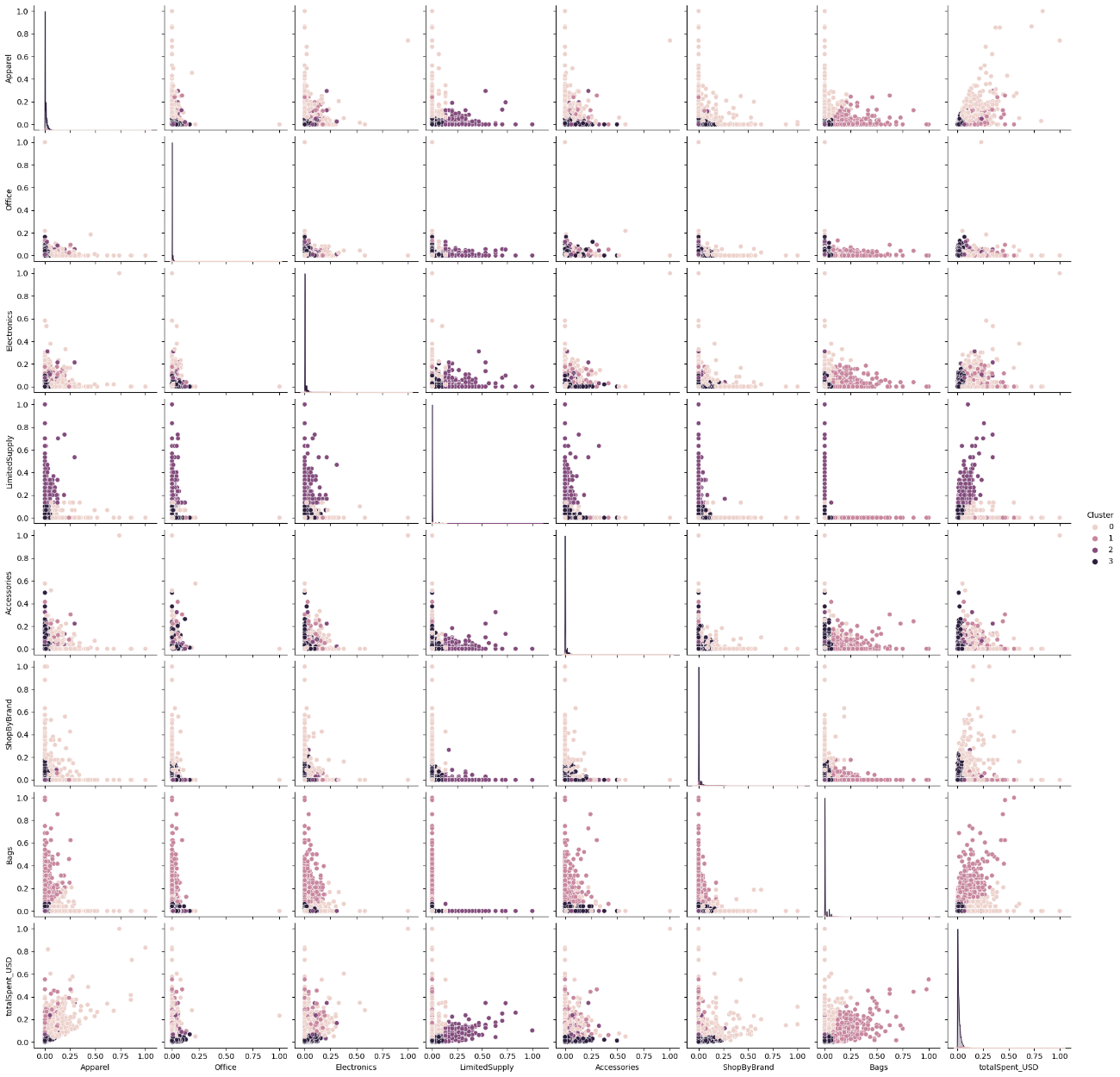


Al aplicar el método Elbow y encontrando el coeficiente Silhouette podemos observar que la mejor cantidad de clusters que podemos considerar el 4. Donde utilizando PCA podemos evidenciar el gráfico del lado derecho. En contraste con considerar 3 clusters que era lo inicialmente solicitado, podemos evidenciar una clara distribución un poco desbalanceada de acuerdo a la localización de nuestra data, veasé gráfico de la izquierda para planteamiento con 3 clusters.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente 

De este modo, la representación de los 4 clusters, permite evidenciar en un esquema mediante pairplot que:



1. Para el cluster 0: se evidencia una mayor compra de Apparel que los demás productos, salvo en ShopByBrand, en la cual se evidencia que gran parte de los clientes que compran Apparel, y también ShopByBrand a nivel global. De igual modo aquello clientes que compran una cantidad de ítems en la categoría Apparel, presentan también un mayor gasto en dólares (totalSpent\_USD), lo que representa una relación directamente proporcional entre las variables. Se puede agregar que este cluster presenta aquellos clientes que generan un mayor gasto totalSpent\_USD en general de acuerdo a cada ítem.
2. Para el cluster 1: se evidencia una mayor compra de Bags, que de los demás productos. En este caso, también se muestra que a pesar de tener una cantidad más grande de Bags el incremento en totalSpent\_USD, no es tan pronunciado como en las demás variables. Para este cluster se presentan aquellos clientes que generan un gasto totalSpent\_USD medio, compartiéndolo con el cluster 2.
3. Para el cluster 2: se evidencia una mayor compra en LimitedSupply, que los demás productos. De este modo para las demás variables no es tan significativo este cluster. Para este cluster se presentan aquellos clientes que generan un gasto totalSpent\_USD medio, compartiéndolo con el cluster 1.
4. Para el cluster 3: El cluster 3 representa todos aquellos clientes que generan compras de pocos ítems en cada una de las categorías y de igual modo, representan un menor gasto en dólares totalSpent\_USD.

**5. Test SQL**

**Pregunta 1:** Solución A)

1. SELECT abi\_email FROM usa\_web\_form WHERE abi\_email LIKE 'john%' OR abi\_email LIKE 'phillip%'

**Pregunta 2:** Solución A)

1. SELECT COUNT(DISTINCT abi\_email) FROM usa\_web\_form WHERE td\_host = 'www.budweiser.com' AND abi\_age BETWEEN 21 AND 34

**Pregunta 3:** Solución A)

1. Will not retrieve anything because some WHERE statements are conflicting

**Pregunta 4:** Solución A)

1. SELECT COUNT(DISTINCT v.visited\_page), f.abi\_email FROM usa\_web\_form f JOIN usa\_page\_views v ON f.cookie\_id = v.cookie\_id GROUP BY f.abi\_email

**Pregunta 5:** Solución A)

1. SELECT cookie\_id, timestamp, abi\_email CASE WHEN lower(abi\_gender) IN ('male', 'men') THEN 'M' WHEN lower(abi\_gender) IN ('f', 'female', 'girl') THEN 'F' WHEN lower(abi\_gender) IN ('other') THEN 'O' END AS abi\_gender FROM usa\_web\_form

**Pregunta 6:**

SELECT brand\_name, COUNT(DISTINCT client\_id) FROM abi\_consents

WHERE consent = 'TC-PP'

GROUP BY brand\_name;