### **1. Pregunta de investigación**

Los productos cosméticos y de belleza en Colombia son un mercado significativo en relación al volumen de ventas anuales que actualmente representan, según Juan Carlos Castro, director ejecutivo de la Cámara de la industria Farmacéutica y de Aseo de la Asociación Nacional de Empresarios de Colombia (Andi), en el 2019 esta industria registró ventas por US$3,572 millones, siendo este el tercer país con mayor número de ventas en América Latina [2].

Ante un mercado tan significativo y dinámico como lo es el caso de los productos cosméticos y de belleza, se vuelve esencial el uso de estrategias de mercadeo e-commerce para obtener una mayor promoción de productos relevantes para cada usuario, viéndose esto reflejado en un aumento en el volumen de ventas (Srivastava, Bala y Kumar, 2020). En la actualidad empresas líderes en este tipo de estrategias como lo son Netflix y Amazon, han logrado incrementar las utilidades hasta en un 29% (Walek y Fojtik, 2020).

Ahora bien, lo anterior, sumado a los grandes volúmenes de información que actualmente se captan en los procesos de compra en internet, se plantea como pregunta de investigación: **¿Cómo desarrollar un sistema de recomendación escalable en big data que permita la promoción de productos relevantes dependiendo del usuario?**

**2. Objetivo General**

Desarrollar un sistema de recomendación escalable y eficiente en Big Data que permita la promoción de productos relevantes dependiendo del usuario.

**3. Objetivo Específicos**

* Identificar metodologías clásicas y actuales utilizadas para desarrollar sistemas de recomendación de e-commerce.
* Construir un modelo baseline basado en la metodología: Collaborative Filtering-Memory Based (based on nearest neighbours).
* Construir un modelo con una metodología robusta, que sea escalable en Big Data
* Realizar un análisis comparativo entre los modelos desarrollados.

### **4.Revisión de la literatura y el estado del arte**

Los sistemas de recomendación fueron diseñados para mostrar contenido relevante y guiar al usuario hacia la información o el producto que pudiese ser de su interés (Subías, Laverón y Molina, 2018). En la actualidad existen 3 tipos de sistemas de recomendación: sistemas basados en contenidos, filtros colaborativos y sistemas híbridos.

El primer sistema (basados en contenido) utiliza información recolectada de las preferencias del usuario y la compara con productos que puedan ser recomendados con el fin de predecir el comportamiento futuro del consumidor (Subías, Laverón y Molina, 2018). El segundo sistema (filtrado colaborativo) por su parte procesa computacionalmente recomendaciones basadas en asociación de información de usuarios o productos que sean similares (Maheswari, Geetha, Selva, Kumar, 2019). Finalmente, los sistemas híbridos consisten en una combinación entre los sistemas de filtrado colaborativo y basados en contenido para llevar a cabo una recomendación.

En relación a los sistemas de filtrado colaborativo el cuál es el sistema de recomendación de interés para el presente proyecto, se pueden definir 2 enfoques, uno basado en los usuarios y otro basado en el producto (Kevin Liao, 2018)

Los sistemas de filtrado colaborativo utilizan las acciones de los usuarios para recomendar otras películas. En general, pueden basarse en el usuario o en el elemento. El enfoque basado en elementos generalmente se prefiere al enfoque basado en el usuario. El enfoque basado en el usuario es a menudo más difícil de escalar debido a la naturaleza dinámica de los usuarios, mientras que los elementos generalmente no cambian mucho, y el enfoque basado en elementos a menudo se puede calcular sin conexión y servirse sin tener que volver a capacitarse constantemente.

Si se quiere implementar un filtrado colaborativo basado en elementos, KNN es un modelo de aprendizaje perezoso no paramétrico. Utiliza una base de datos en la que los puntos de datos se separan en varios grupos para hacer inferencias para nuevas muestras.

KNN no hace ninguna suposición sobre la distribución de datos subyacente, pero se basa en la similitud de las características del elemento. Cuando KNN hace inferencias sobre una película, KNN calculará la "distancia" entre la película de destino y todas las demás películas en su base de datos, luego clasifica sus distancias y devuelve las K películas vecinas más cercanas como las recomendaciones de películas más similares.[10]

Ahora se sabe que KNN funciona muy bien para aplicarlo a productos de belleza. En el documento “Recommender Systems Based on Collaborative Filtering Using Review Texts, A Survey” (Mehdi Srif, 2020), se expone que mediante las calificaciones basadas en texto según los usuarios que han usado algún producto, las palabras de revisión pueden mejorar la similitud tradicional en el modelo de KNN.

Dado que la base de datos que se emplea para el desarrollo del proyecto se utiliza el filtrado colaborativo basados en KNN, se adapta a la metodología adecuadamente esto según el documento “KNN Based Collaborative Filtering In Python using Surprise” (Pankaj kumar, 2019).

Otra metodología que se ha utilizado para sistemas de recomendación con alto flujo de datos es la factorización de matriz de mínimos cuadrados (ALS) en el filtrado colaborativo (kevin liao, 2018), Kevin Liao ayuda a entender que el sistema de recomendación se puede mejorar mediante factorización de matrices implementada en Spark.[14]

KNN tiene varios limitantes, el sesgo de popularidad, el problema de inicio en frio y escalabilidad.

Cuando se habla del Sesgo de Popularidad, se refiere al sistema que recomienda los artículos con más interacciones sin ninguna personalización, en la otra mano se tiene el problema de inicio en frío del elemento que es cuando los artículos agregados al catálogo tienen muy poca o ninguna interacción, mientras que el recomendador se basa en las interacciones de los artículos para hacer recomendaciones por lo tanto el problema de escalabilidad se evidencia en la capacidad de escalar a conjuntos de datos mucho más grandes cuando se agregan más y más usuarios y artículos a la base de datos.

Para resolver este problema se plantea la factorización matricial la cual es una familia de operaciones matemáticas para matrices en álgebra lineal. Es decir, una factorización matricial es una factorización de una matriz en un producto de matrices. En el caso del filtrado colaborativo, los algoritmos de factorización matricial funcionan descomponiendo la matriz de interacción usuario-elemento en el producto de dos matrices rectangulares de menor dimensionalidad. Una matriz puede verse como la matriz de usuarios donde las filas representan a los usuarios y las columnas son factores latentes. La otra matriz es la matriz de elementos donde las filas son factores latentes y las columnas representan elementos.[14]

Por otra parte, autores como Vaibhav Kumar (2020) plantean que los sistemas de recomendación aplicados a la descomposición de valores singulares utilizando modelos de filtro colaborativo pueden mejorar las estimaciones del sistema de recomendación. [15]

Descomposición de valor singular (SVD), es un método clásico del algebra lineal utilizado como una técnica de reducción de dimensionalidad en el aprendizaje automático. En otras palabras, es una técnica de factorización matricial, que reduce el número de características de un conjunto de datos.

Esta metodología se utiliza en los sistemas de recomendación cuando se aplica la técnica de filtrado colaborativo la cual utiliza una estructura matricial en la que cada fila representa un usuario y cada columna representa un elemento.

En este caso los elementos de la matriz son las calificaciones que los usuarios otorgan a los productos.

La factorización de esta matriz se realiza mediante la descomposición de valores singulares. El desarrollo consiste en que encuentra factores de matrices a partir de la factorización de una matriz de alto nivel (usuario-elemento-calificación).

Para acotar el concepto se tiene que la Descomposición de Valor Singular es un método para descomponer una matriz en otras tres matrices. Donde es una matriz de utilidad , es una matriz singular ortogonal izquierda , que representa la relación entre usuarios y factores latentes, es una matriz diagonal , que describe la fuerza de cada factor latente y es una matriz singular diagonal derecha, que indica la similitud entre ítems y factores latentes.

Los factores latentes aquí son las características de los elementos, por ejemplo, el tipo de producto de cosmético. La SVD disminuye la dimensión de la matriz de utilidad A al extraer sus factores latentes. Mapea a cada usuario y cada producto en un espacio latente r- dimensional.

### **5. Implementación de Metodología**

A continuación, se define el proceso que se llevó a a cabo en cada una de las iteraciones

**Iteración 1**

Entendimiento y extracción de la información:

En esta iteración se realizaron búsquedas de palabras clave en la web sobre cosméticos y belleza con el objetivo de entender el comportamiento del mercado a nivel nacional y Latinoamérica. Se encuentra que el sector cosméticos en Colombia es uno de los pilares de la economía del país y el tercer en Latinoamérica, datos que son explicados por Carlos Castro, director ejecutivo de la Cámara de la industria Farmacéutica y de Aseo de la Asociación Nacional de Empresarios de Colombia (Andi) por otro lado se encontró que el gasto promedio en las mujeres en productos de belleza está entre 300 y 400 mil pesos colombianos cifra cerrada en el 2019 lo cual incrementaron la producción de cosméticos en tasas de 5,7%. Posteriormente se investiga sobre estudios previos en sistemas de recomendación de productos de belleza finalmente se evidencia que no existen artículos concluyentes de un sistema claro de recomendación utilizando técnicas de machine learning en el mercado de cosméticos.

**Iteración 2**

Estructuración de la información:

La base de datos fue extraída desde la WEB y posteriormente guardada en “parquet” para reducir el peso del archivo. Se construye una arquitectura para almacenar los datos en S3 Amazon, donde se crea un clúster en EMR para preparar los datos a la exploración.

**Iteración 3**

Estructuración y exploración de datos:

Dentro de EMR se genera un ambiente en Jupyter notebook para aplicar la metodología de un modelo Baseline, KNN, SVD y ALS en Python y PySpark. Inicialmente se limpia la base de datos y se estructura la información realizando filtros específicos que ayudan a construir la capa semántica donde se analizarán con técnicas de estadística descriptiva y prescriptiva el comportamiento de los datos y finalmente se deja la base lista para ser consumida por la capa de modelamiento la cual se normaliza de acuerdo al rating de calificación cada review del producto.

**Iteración 4**

Construcción de modelos

En la capa de modelamiento de acuerdo a la investigación previa sobre sistemas de recomendación se emplean metodologías clásicos y actuales de ALS, KNN y SVD. Adicionalmente se identifican las métricas óptimas para la evaluación de sistemas de recomendación y se aplican a los modelos desarrollados

**Iteración 5**

Comparación de modelos finales

En esta iteración se comparan los modelos por medio de RMSE para seleccionar el mejor sistema de recomendación que se ajusta a la realidad que se requiere del el usuario consumidor.

### **6. Análisis de los datos**

### **Hallazgos importantes**

De acuerdo a la información obtenida en la base de datos se evidencia lo siguiente:

* Se han registrados reviews desde el año 2000 y se observa un crecimiento significativo a partir del año 2011demostrando así un incremento en compras por medio del e-commerce en el sector de cosméticos y belleza

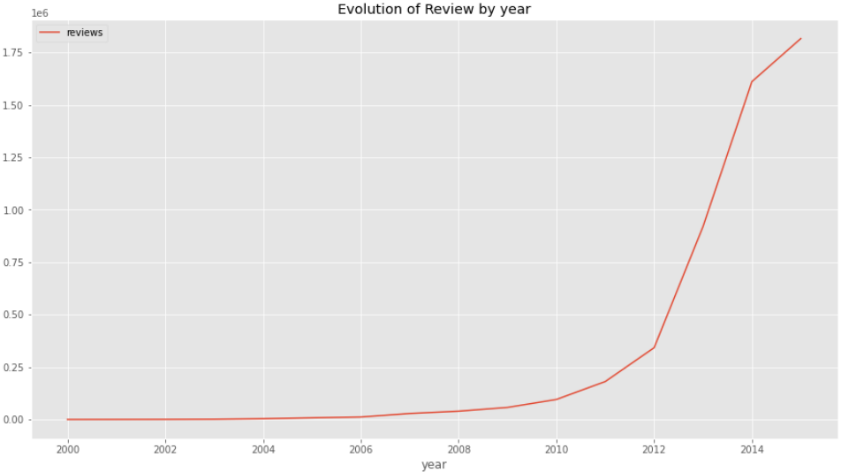


Imagen 3: Gráfica de tendencia de reviews (2000-2015)

* Se evidencia que del top 10 de los productos con más ventas registradas en la historia es la “plancha para el cabello” con ventas de 15.410 ítems vendidos.

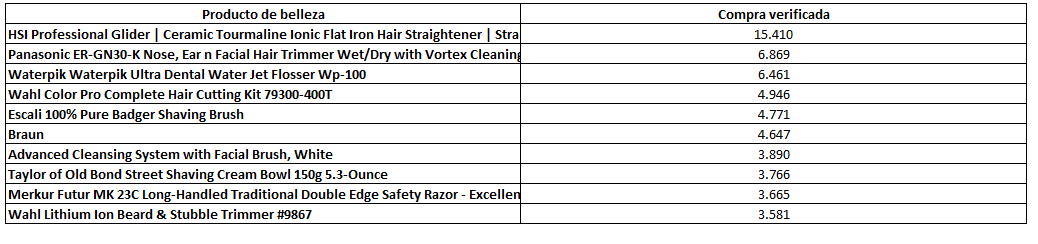


Imagen 4: Gráfica top 10 de productos más vendidos online

* De acuerdo a la reviews realizadas por los compradores, el 64% ha calificado en 5 en calidad de los productos de belleza y cosméticos vendidas en la red siento esta la mayor calificación de calidad en comparación de los otros niveles.

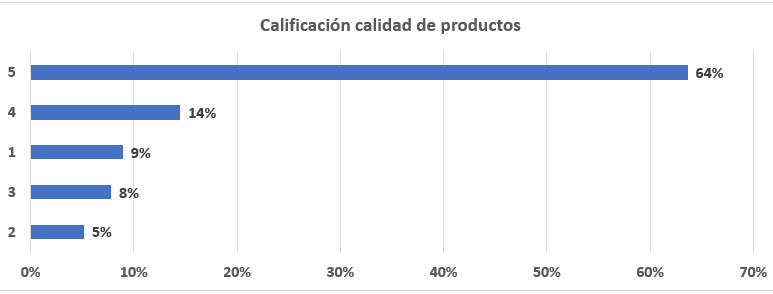


Imagen 5: Gráfica de calificación de productos online

### **7. Modelamiento para el sistema de recomendación**

### **Modelo KNN**

Para realizar el modelo con KNN se filtró la base de datos por usuarios e ítems más frecuentes, los criterios fueron los siguientes:

* Ítems con más de 60 reviews
* Clientes con más de 12 reviews

Se realizaron estos criterios aleatoriamente con el objetivo de identificar los clientes más frecuentes y aplicar el modelo de KNN con **small data.**

La base se dividió la base entre test y train (25%- 75%) respectivamente

* El RMSE fue de 1.36
* El tiempo de ejecución fue de 9 segundos

### **Baseline**

Se inició un modelo baseline con los siguientes procedimientos

* La base se dividió la base entre test y train (25%- 75%) respectivamente.
* El entrenamiento del modelo registró un tiempo de 60 segundos
* El RMSE final fue de 1.515 siendo un error muy alto para la validación.

### **Modelo SVD**

Para mejorar el baseline se propone un modelo SVD y se obtuvo lo siguiente:

* La base se dividió la base entre test y train (25%- 75%) respectivamente.
* El entrenamiento del modelo registró un tiempo de 325 segundos (5.4 minutos)
* El RMSE fue de 1.516 teniendo un error similar al del baseline

### **Escalamiento de datos modelamiento en Spark desde AWS**

### **Modelo ALS**

De acuerdo a los resultados anteriores se procede a escalar los datos a AWS usando Spark para el procesamiento de grandes volúmenes de datos y se obtuvo lo siguiente:

* La base se dividió la base entre test y train (25%- 75%) respectivamente.
* El tiempo de ejecución para los dos tratamientos con ALS fue de 230 segundos (3.5 minutos)
* Inicialmente se normalizó el conjunto de datos y se obtuvo un RMSE de 1.11 mejorando el error anterior del SVD
* Se realizó un tratamiento adicional a los datos el cual consiste en normalizar restando la media de los ratings por customer\_id, este procedimiento mejoró significativamente el RMSE teniendo un registro de 0.95 con un solo parámetros
* Se cambian el parámetro por otro diferente y se realiza otra iteración obteniendo finalmente un RMSE de 0.98

### **Tabla comparativa entre los modelos aplicados**

En la siguiente tabla se muestra una comparación de los modelos desarrollados por RMSE y tiempo de ejecución.

* El modelo final que mejor resultados obtuvo por RMSE fue el ALS con el último tratamiento de datos que se aplicó
* El modelo que tuvo mayor tiempo de ejecución fue el SVD con casi 5 minutos de ejecución
* El modelo que tuvo menor tiempo de ejecución fue KNN que se realizó con una pequeña de datos registrando 9 segundos

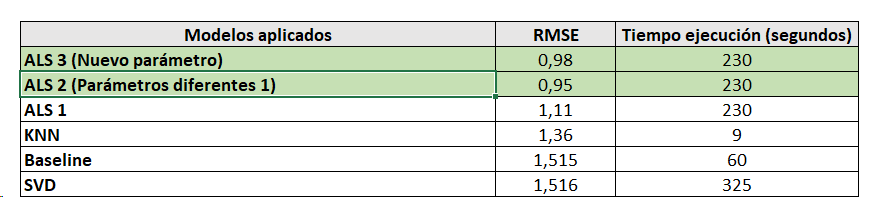


Imagen 6: Tabla comparativa de modelos

### **8. Uso de herramientas de Big Data**

### **Arquitectura de ETL en AWS**

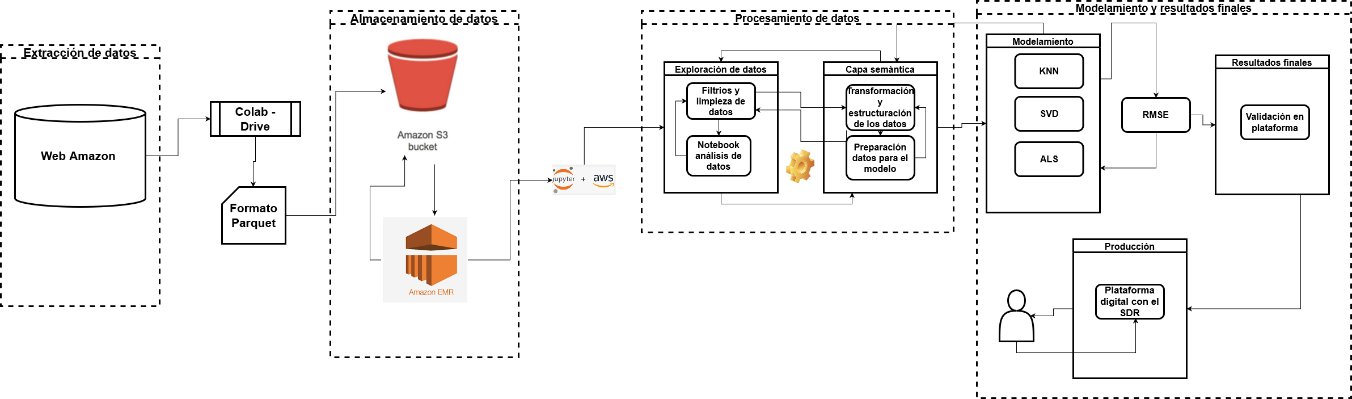


Imagen 7: Arquitectura ETL AWS

Nota: Para analizar mejor la arquitectura entrar al siguiente [link](https://eafit.sharepoint.com/sites/ProyectoMinera/Documentos%20compartidos/Forms/AllItems.aspx?RootFolder=%2fsites%2fProyectoMinera%2fDocumentos%20compartidos%2fGeneral%2f8%2e%20Arquitectura&FolderCTID=0x012000DAADB7E50628854C9AA8502DD330C1CD)

### **Extracción y almacenamiento**

La ubicación del conjunto de datos se encuentra en la web de Amazon tomando el siguiente link para descargarlos:

[Amazon\_reviews\_us\_beauty\_v1\_00.tsv.gz](https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Beauty_v1_00.tsv.gz)

Luego de obtener la fuente de información son llevados a Google colab ubicado en el drive de cada cuenta de Gmail, inmediatamente son guardados en formato “parquet” el cual ayuda a reducir el peso del archivo, tenga utilizada para manejar grandes volúmenes de información.

Para tener una fuente sólida donde almacenar los datos crudos se utiliza la instancia de S3 habilitada en Amazon y allí se crea un clúster en EMR con el fin de consumir la información. Los datos son actualizados constantemente desde S3 y EMR.

### **Procesamiento de los datos**

Gracias a la creación del clúster en EMR, generamos un ambiente en Jupyter creando notebooks con kernel en Python y librería para lenguaje Spark, lenguajes que soportan cualquier conexión con plataformas digitales. En este ambiente se inicia la fase de exploración y transformación de datos. Luego de tener los datos limpios se genera una capa semántica la cual tiene el objetivo de estar preparada para ser consumida por los algoritmos del modelamiento predictivo sistema de recomendación. Esta fase de preparación de la base de datos tiene varias iteraciones ya que se van generando nuevas actualizaciones o ideas en el transcurso del proceso con el fin de afinar la base final para el modelo.

### **Modelamiento y resultados finales**

En la etapa de modelamiento se desarrollaron varios modelos, KNN con small data en python, baseline en Python, y escalados en Spark SVD y ALS en estos se entrenan y se prueban los modelos evaluándolos con RMSE. Se realiza una comparación entre los modelos y de acuerdo a los resultados se seleccionó quien es el modelo que se mejor recomiende al usuario final en una plataforma digital.

Esta arquitectura cuenta con capacidad de escalamiento ya que dado a la tendencia de crecimiento de este mercado en e-commerce cuenta con el soporte del flujo de datos desde Amazon y con la capacidad de procesar los modelos en tiempo real desde el clúster creado en EMR.

### **8. Ejecución del plan diseñado en diagrama Gantt**

Se diseñó un plan con la metodología del diagrama de Gantt para llevar a cabo el desarrollo del proyecto

1. Investigación profunda sobre el mercado de cosméticos:

En este paso tuvo una ejecución del 100%, donde participaron 2 integrantes del equipo quienes extrajeron y documentaron toda la información sobre el mercado de los artículos de cosméticos y belleza, cifras más importantes, conocimiento del cliente consumidor de productos de belleza, identificación de sistemas de recomendación con metodologías KNN y ALS. Se identificó además que el plan de ejecución en cuanto a los días estimados para realizar esta operación fue exacta y óptima dejando por un lado retrasos o perdida de información, es decir no hubo retrasos.

1. Extracción y almacenamiento de los datos en AWS

Se tuvo una ejecución del 100%, donde participaron 2 integrantes del equipo quienes se encargaron de diseñar la arquitectura e implementarla en S3 donde se almacenó el conjunto de datos, allí se creó el clúster en EMR para el procesamiento de los datos y se preparó el ambiente para los lenguajes de Python y Spark. No se evidenciaron retrasos y se ejecutó en el tiempo planteado.

1. Transformación y limpieza de datos

En esta sección se tuvo una ejecución de 100% con varias iteraciones lo cual extendió el plazo de ejecución con 1 día de desviación. Se tuvo en cuenta la necesidad de limpieza y transformación de variables para preparar la capa semántica donde se alimentará los modelos. En este plan estuvieron a cargo 2 integrantes del equipo.

1. Etapa de exploración de métricas, estructuración de sistemas de recomendación.

En cuanto a la construcción e identificación de mejores modelos de base para el desarrollo del sistema de recomendación propuesto se tuvo una ejecución del 100% sin iteraciones, sin retrasos. En el proceso de validaciones de resultados de los modelos y comparación entre ellos se tuvieron en cuenta varias iteraciones que extendieron el plazo de entrega, teniendo una desviación de 2 días con dos integrantes encargados del proceso.

1. Etapa final del modelo de SDR con KNN, Baseline, SVD y ALS

Gracias a la etapa anterior se genera dos modelos finales de KNN, Baseline, SVD y ALS en sistemas de recomendación, se tuvo una ejecución del 100% pero con varias iteraciones para afinar el modelo final con pruebas y validaciones de resultados. Tuvo una desviación de 3 días con respecto a lo planteado con dos integrantes encargados del proceso.

1. Etapa de documentación final y entrega

Cada etapa tuvo su documentación respectiva y almacenamiento en el repositorio en Github por lo tanto se tuvo una ejecución del 100% sin desviaciones. Se construyó una presentación de entrega la cual tuvo una desviación de 1 día. Se trabajo en esta etapa con todos los integrantes del equipo.

Al principio del planteamiento de plan de ejecución se estimaron las iteraciones que se podrían tener en cada etapa gracias a esto se ejecutó el total del plan en 100% con promedio de 2 días de desviación según lo acordado.

Las lecciones aprendidas en la ejecución del proyecto para mitigar las desviaciones es reforzar la comunicación entre los miembros del equipo en cuanto a funciones reiterativas que ya se habían hecho para no perder tiempo en trabajos ya realizados.

[Link para tener una amplia visión del plan ejecutado](https://eafit.sharepoint.com/sites/ProyectoMinera/Documentos%20compartidos/Forms/AllItems.aspx?RootFolder=%2fsites%2fProyectoMinera%2fDocumentos%20compartidos%2fGeneral%2f4%2e%20Plantilla%20diagrama%20Gantt&FolderCTID=0x012000DAADB7E50628854C9AA8502DD330C1CD)

### **9. Implicaciones éticas**

En relación a las implicaciones éticas del proyecto, en la etapa de recolección de datos se asume que estos fueron recolectados por Amazon cumpliendo con los términos y condiciones acordados con los usuarios y que no se está infringiendo en ningún momento aspectos que vayan en contra de la privacidad y anonimidad de cada uno de éstos. En relación a la etapa de desarrollo se utilizan tanto datos como programas libres de uso y para la etapa de implementación, dado que la recomendación de estos productos no es sensible al público no tendrá ningún impacto o consecuencia negativa en la sociedad. Finalmente resaltar que los algoritmos utilizados en este proyecto son modelos estadísticos que no son de tipo “Caja Negra” y que sus conclusiones podrán ser abiertamente explicados en caso de llegar a ser necesario

### **10. Aspectos legales y comerciales**

Los datos utilizados en el desarrollo del presente proyecto son open data de la empresa Amazon, estos son de libre uso y no representa ningún tipo de limitantes en cuanto a exponer los resultados encontrados.

El potencial comercial del proyecto es muy grande en el mercado colombiano, adicional a lo significativo volumen de ventas que actualmente representa, la visión que se tiene para el año 2032 en el Plan de Negocios del sector Cosméticos y Aseo, un programa liderado por Colombia Productiva, es ser un país “Líder en el continente en producción y exportación de cosméticos y productos de aseo del hogar de alta calidad con base en ingredientes naturales”.

### **11. Entregables y descripción**

En el repositorio de [Github](https://github.com/Proy-Maest/SDR-mineria-de-datos) se encuentra ubicado la documentación de los códigos y la argumentación teórica del trabajo.

El objetivo de tenerlo en este sitio es compartir conocimiento sobre la propuesta de realizar sistemas de recomendación con diferentes metodologías estadísticas y matemáticas y cuando se tienen grandes cantidades de datos.

Se evidencian varias carpetas que contienen el desarrollo del proyecto

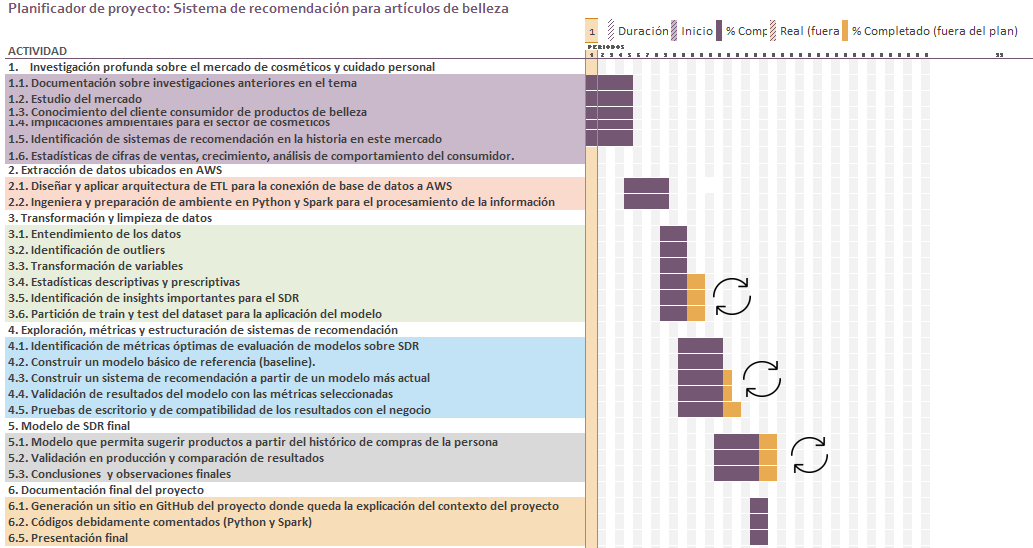
1. Trabajo final
2. Fuente de datos AWS
3. Arquitectura de ETL conexión con AWS
4. Modelamiento: Notebook Python y PySpark

### **12. Conclusiones y trabajo futuro**

* Para crear un buen sistema de recomendación el tratamiento de datos es fundamental para obtener una buena estimación, en este caso se tiene ratings de 1 a 5 y esto genera complicaciones en la matriz de utilidad generando ceros en los datos faltantes y afectando así el rating de calificación.
* Se tuvo que realizar ciertas normalizaciones con el fin de que el cero sea un valor neutro en la calificación.
* Evaluando los resultados de RMSE y de tiempo de entrenamiento del modelo, se considera que la mejor metodología para desarrollar el sistema de recomendación escalable es un modelo de ALS desarrollado en Spark
* Para el tratamiento de grandes volúmenes es recomendable utilizar herramientas que provee una plataforma como AWS, ya que para este proyecto nos facilitó la distribución de los datos de una manera óptima.
* Desde la base de datos se podría realizar un desarrollo de minería de texto con el fin de extraer insights en los comentarios que dejan las personas en la plataforma ya que se podría alimentar la base datos y probablemente mejore la predicción en el sistema de recomendación y un mejor conocimiento del consumidor.
* Se utiliza un solo parámetro como un primer acercamiento de experimentación y luego se intenta realizar diferentes iteraciones con diferente parámetro logrando finalmente mejorar el RMSE a 0.98, concluyendo así que los parámetros son significativos al ejecutar el modelo.

### **Anexo**

### Diagrama de Gantt – Plan de desarrollo



### **Referencias bibliográficas**

[1] “Informe del sector cosmético”, 2019, Prensa Inexmoda

[2] “Cosméticos, un mercado que movió el año pasado US$3.572 millones en Colombia”, Ximena González, diario La Republica, 2019

[3] “Con 74%, el precio es el principal determinante en la compra de cosméticos”, sectorial, análisis, monitoreo y evaluación de sectores, 2019

[4] “Mercado de belleza en Colombia mueve US$3,000 millones al año” El Espectador, 2019.

[5] Srivastava, A., Bala, P. K., & Kumar, B. (2020). “New perspectives on gray sheep behavior in E-commerce recommendations. Journal of Retailing and Consumer Services”

[6] Walek, B., & Fojtik, V. (2020). “A hybrid recommender system for recommending relevant movies using an expert system. Expert Systems With Applications”

[9] “Prototyping a Recommender System Step by Step Part 1: KNN Item-Based Collaborative Filtering”, Kevin Liao, 2018

[10] “Advanced Collaborative Filtering and Image-based Recommender Systems”, Bowen Zhou, 2017

[11] “ Recommnder Systems Based on Collaborative Filtering Using Review Texts—A Survey”, Mehdi Srif, 2020

[12] “KNN Based Collaborative Filtering In Python using Surprise” Pankaj kumar, 2019

[13] “Informe del Sector Cosmético – junio 2019”, sala de prensa Moda, cluster Bogota cosméticos

[14] “Prototyping a Recommender System Step by Step Part 2: Alternating Least Square (ALS) Matrix Factorization in Collaborative Filtering”, kevin liao, 2018

[15] “Singular Value Decomposition (SVD), its application in Recomender System”,DR. Vaibhav Kumar,2020.

[16]Subías, M. H., Laverón, M. M., & Molina, A. M. U. (2018). Online recommendation systems in the Spanish Audiovisual market: comparative analysis between Atresmedia, Movistar+ and Netflix. Universia Business Review, 60, 54–89. <https://doi-org.ezproxy.eafit.edu.co/10.3232/UBR.2018.V15.N4.02>

[17]Maheswari, M., Geetha, S., & Selva, kumar, S. (2019). Adaptable and proficient Hellinger Coefficient Based Collaborative Filtering for recommendation system. Cluster Computing, 22(5), 12325–12338. https://doi-org.ezproxy.eafit.edu.co/10.1007/s10586-017-1616-7