### **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]. Por su parte Inexmoda, una de las principales empresas de este mercado, indica que este mercado ha venido creciendo considerablemente y que el gasto promedio trimestral de las mujeres colombianas es de $300,000 pesos [1].

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) [5]. Los canales de distribución, según el artículo [4] publicado en el periódico El Espectador, “la venta directa o por internet de productos cosméticos tiene una participación de un 30,9% y un 0,2%, respectivamente”. 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?**[6], y la respuesta conduce a que se debe utilizar técnicas basadas de minería de datos para grandes volúmenes de información, es decir Big data.

El Big Data consiste en analizar grandes volúmenes de datos, bien para crear nuevos productos o bien para mejorar en competitividad y productividad. De acuerdo a esto se utilizarán varias etapas para crear valor a esta gran cantidad de datos: Infraestructura, herramienta de acceso, organización (arquitectura), análisis y toma de decisiones[7]. En efecto para aplicar estas etapas se emplea tecnologías que pueden almacenar y administrar esta información como lo es en este caso AWS services, la cual es una colección de servicios de computación en la nube pública que en conjunto forman una plataforma de computación en la nube.[8]

### **Trabajos realizados con metodologías de sistema de recomendación con KNN y ALS**

De acuerdo con Kevin Liao, científico de datos de aprendizaje automático, se conoce que la mayoría de los productos de internet que hoy en día usamos funciona con sistemas de recomendación. Ejemplo YouTube, Netflix, Amazon, Pinterest y otras más plataformas que se basan en este sistema para filtrar millones de contenidos y realizar recomendaciones personalizadas para cada usuario.

Los sistemas de recomendación han sido estudiados y probados para proporcionar efectividad en la compra por internet y favorece en gran tamaño a los usuarios, ventaja que hoy en día el mercado de cosméticos y belleza está alcanzando cada vez más. Kevin propone un prototipo el cuál ha sido viable para realizar sistemas de recomendación, en este caso realiza uno para recomendar películas y en este proyecto lo aplicaremos para productos de belleza.

Los sistemas de recomendación se dividen en tres categorías:

**Sistemas basados en contenido, sistemas de filtrado colaborativo y sistemas híbridos**.

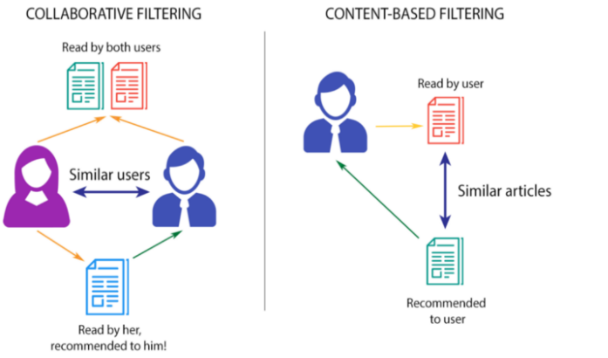


Figura 1: “Descripción general de los sistemas de recomendación”[10]

Enfoque basado en **contenido** utiliza una serie de características discretas de un artículo para recomendar artículos adicionales con propiedades similares.

El enfoque de **filtrado colaborativo** crea un modelo a partir de los comportamientos pasados ​​de un usuario (artículos comprados o seleccionados previamente y / o calificaciones numéricas otorgadas a esos artículos), así como decisiones similares tomadas por otros usuarios. El enfoque híbrido combina los dos enfoques anteriores.

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]

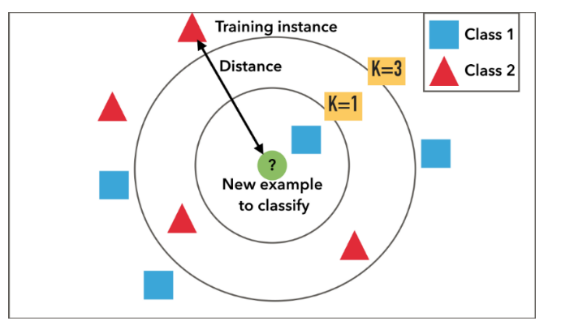


Figura2 : KNN realizando clasificación sobre una nueva muestra[11]

Ahora se sabe que KNN funciona muy bien para aplicarlo a productos de belleza. En documento “Recommender Systems Based on Collaborative Filtering Using Review Texts—A Survey” se expone que mediane 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.[11]

Dado que la base de datos que se emplea para el desarrollo del proyecto se utiliza el filtrado colaborativo basados en KNN, como también se usó pankja Kumar en su artículo “KNN Based Collaborative Filtering In Python using Surprise”[12]

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 ayuda a entender que el sistema de recomendación se puede mejorar mediante factorización de matrices implementada en Spark.

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

El sesgo de popularidad se refiere al sistema que recomienda los artículos con más interacciones sin ninguna personalización. El problema de inicio en frío del elemento 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 luego el problema de escalabilidad es cuando falta de capacidad para 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 plante 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]

Otra metodología utilizada para construir los sistemas de recomendación es **Descomposición de valor singular (SVD)**, este 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 al reducir la dimensión espacial de la dimensión N a la dimensión K (donde K< N). En sistemas de recomendación se utiliza la en 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. Los elementos de la matriz son las calificaciones que los usuarios otorgan a los elementos.

La factorización de esta matriz se realiza mediante la descomposición de valores singulares. Encuentra factores de matrices a partir de la factorización de una matriz de alto nivel (usuario-elemento-calificación). 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 género de la música. La SVD disminuye la dimensión de la matriz de utilidad A al extraer sus factores latentes. Mapea a cada usuario y cada elemento en un espacio latente r- dimensional.[15]

### **Metodología para el desarrollo del Sistema de recomendación con aplicaciones KNN, SVD en Python y ALS escalado en Spark**

**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 KNN y SVD en Python y ALS en 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.

**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 de SRD para seleccionar el mejor sistema de recomendación que se ajusta a la realidad que se requiere del sector.

### **Análisis de los datos**

La base de datos utilizada para el presente proyecto cuenta con 5,115,666 de reviews realizados por 2,816,378 de consumidores para 588 mil productos de belleza. El tamaño del archivo con los datos en formato parquet pesa 782 Mb, se tendrá un uso de memoria de 663.5 Mb.

A continuación, se definen cada una de las variables que se utilizaron:

* marketplace: es el país donde se hizo la compra, en nuestro caso solo tenemos información de USA (string)
* customer\_id: Identificador aleatorio que se puede utilizar para identificar reviews hechas por un solo cliente. (string)
* review\_id: Es el id de las opiniones del cliente. (string)
* product\_id: Id del producto al que se le hizo la opinión/review. (string)
* product\_parent: Random identifier that can be used to aggregate reviews for the same product (string)
* product\_title: Título del producto (string)
* product\_category: categoría del producto (string)
* star\_rating: Calificación del review de 1 a 5 estrellas (string)
* helpful\_votes: Número de votos que ayudaron. (string)
* total\_votes: Número total de votos del review (string)
* verified\_purchase: Si el review es sobre una compra verificada. (string)

RMSE :

DATOS NORMALIZADOS

SESGO

(pendiente incluir hallazgos)

### **Uso de herramientas de Big Data**

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

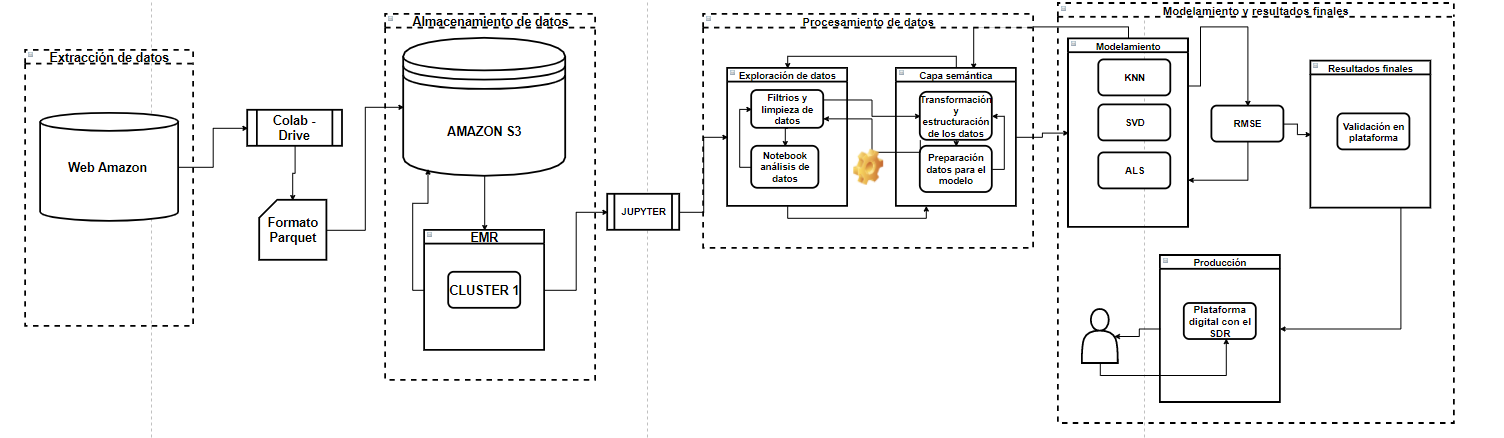


Figura3: Arquitectura ETL AWS

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 generan tres nuevos notebooks, uno para KNN, SVD y otro para 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 selecciona quien será el sistema que consuma el 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.

### **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, SVD y ALS

Gracias a la etapa anterior se genera dos modelos finales de KNN, 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)

### **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

### **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.

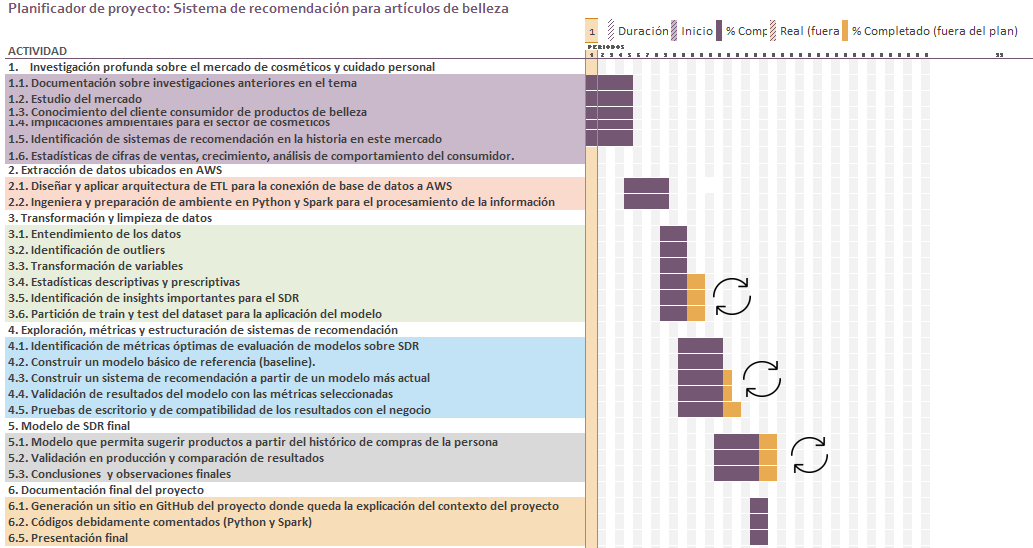
El potencial comercial del proyecto es muy grande en el mercado colombiano, adicional a lo significativo volumen de ventas que actualmente representa, la visión que se tiene para el año 2032 en el Plan de Negocios del sector Cosméticos y Aseo, un programa liderado por Colombia Productiva, es ser un país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

### **Entregables y descripción**

### **Conclusiones y trabajo futuro**

### **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”

[7] What is “Big Data”, 2020, Tomas Olarte Hernandez, Minería de Datos, EAFIT

[8] Almacenamiento y recuperación de la información en AWS, Edwin Nelson Montoya Múnera,2019 EAFIT

[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] “ Recommender 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