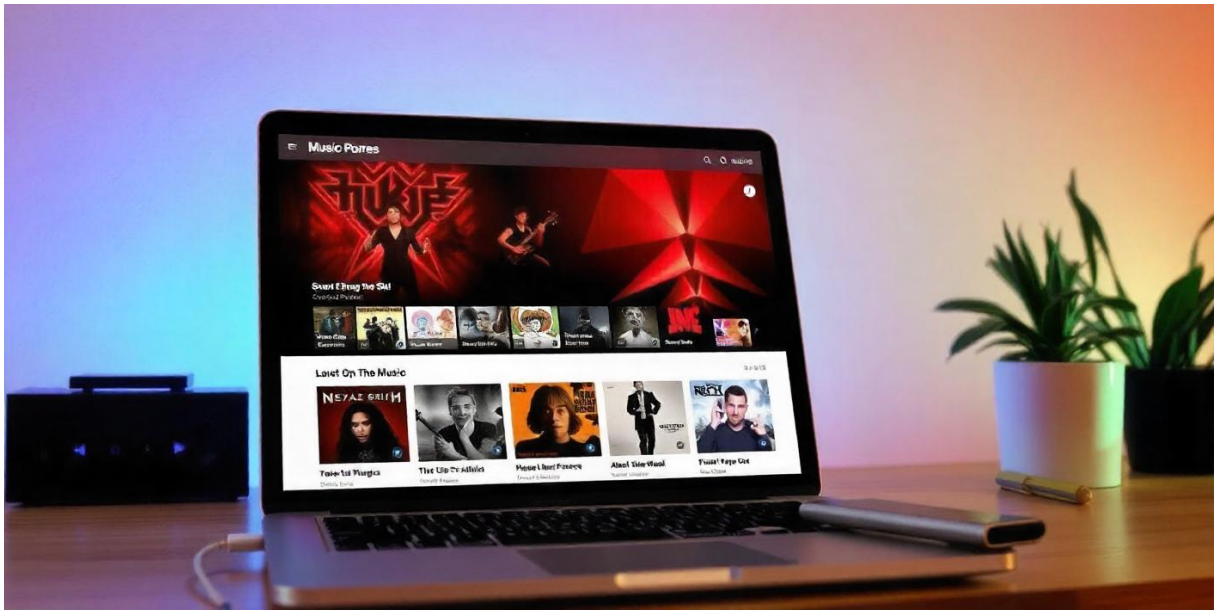
CASO PRÁCTICO 207 ESTADÍSTICA PARA CIENTÍFICO DE DATOS



¿CÓMO APLICAR LA ESTADÍSTICA EN LA INTELIGENCIA DE NEGOCIOS?
CIENCIA DE DATOS PARA NEGOCIOS

Fernando Aleisy González
13 de septiembre de 2024

Índice

Caso 1: Empresa de comercio eléctronico

Una empresa de comercio electrónico está interesada en mejorar su sistema de recomendación de productos para aumentar la tasa de conversión y las ventas. Como experto en Big Data, te han asignado el proyecto de analizar los datos de la empresa y proponer mejoras en el sistema de recomendación.

Figura 1: Representación del concepto Comercio electronico según inteligencia artificail (IA) freepik

1. ¿Cuál es el objetivo principal del análisis de datos en este proyecto?

En este proyecto, el objetivo principal, desde el análisis de datos, es identificar factores calves que permitan personalizar las recomendaciones de manera más efectivas lo que se traduce en un aumento en la tasa de conversión y en las ventas, factores tales como patrones en el comportamiento de los usuarios, preferencias de compra, interacciones con productos.

2. ¿Qué tipo de datos sería relevante recopilar y analizar para mejorar el sistema de recomendación?

Para mejorar un sistema de recomendación en una empresa de comercio electrónico, es fundamental recopilar y analizar varios tipos de datos que permitan entender el comportamiento y las preferencias de los usuarios. Los datos relevantes incluyen:

Datos del usuario

- **Historial de compras**: Productos que el usuario ha comprado anteriormente.
- **Historial de navegación**: Páginas vistas, productos consultados, categorías visitadas.
- **Interacciones con productos**: Clics, tiempo de permanencia en las páginas de productos, productos añadidos al carrito.
- **Preferencias explícitas**: Reseñas, calificaciones, listas de deseos.
- **Datos demográficos**: Edad, género, ubicación, historial de ingresos, etc.
- **Comportamiento en múltiples dispositivos**: Acciones del usuario en diferentes plataformas, como móvil, tablet o pc.

Figura 2: Representación del concepto Datos según inteligencia artificail (IA) freepik

Datos de productos

- **Características del producto**: Descripciones, especificaciones técnicas, precio, imágenes, vídeos, disponibilidad de stock.
- **Categorías y subcategorías**: Relación de productos con categorías específicas para mejorar las recomendaciones dentro de un segmento.
- **Opiniones y reseñas**: Valoraciones de los clientes, comentarios positivos o negativos.
- **Datos de ventas históricas**: Información sobre la popularidad y tendencias de ventas de productos.

Datos de interacción entre usuarios y productos

- **Comportamiento de compra de usuarios similares**: Qué productos compran usuarios con perfiles o comportamientos similares.
- **Secuencia de compras**: Relación entre productos que tienden a comprarse juntos (ejemplo, productos complementarios).
- **Tasa de conversión por recomendación**: Información sobre qué recomendaciones han llevado a compras en el pasado.
- **Abandono de carrito**: Productos que fueron añadidos al carrito pero no comprados, que pueden ser utilizados para nuevas recomendaciones.

Datos contextuales

- **Datos temporales**: Temporada, días festivos, fechas de ofertas, momento del día, frecuencia de interacción con la tienda.
- **Localización geográfica**: Preferencias locales y productos populares en determinadas regiones.
- **Datos de personalización en tiempo real**: Información que permita ajustar recomendaciones en tiempo real según el comportamiento reciente del usuario.

Datos de comportamiento general del sitio

- **Mapas de calor**: Información sobre dónde los usuarios hacen clic o pasan más tiempo en el sitio.
- **Tasa de abandono y navegación**: Secciones del sitio donde los usuarios tienden a abandonar, lo que puede indicar oportunidades para optimizar recomendaciones.

Feedback explícito del sistema de recomendación

- **Interacciones con las recomendaciones**: Qué productos recomendados fueron ignorados o rechazados (ej., un usuario cierra un anuncio o ignora una recomendación).

3. ¿Cómo definirías el problema de recomendación como un problema estadístico?

De acuerdo al tipo de proyectos, se puede sugerir estas etapas como la comprensión del negocio, la captura de datos, el modelado y, finalmente, la aceptación y puesta en producción (Gensollen, 2022). En la etapa de modelado se tiene al problema de recomendación como un problema estadístico, el cual se puede abordar con diferentes herramientas de acuerdo a los tipos de modelados (Flores, 2023).

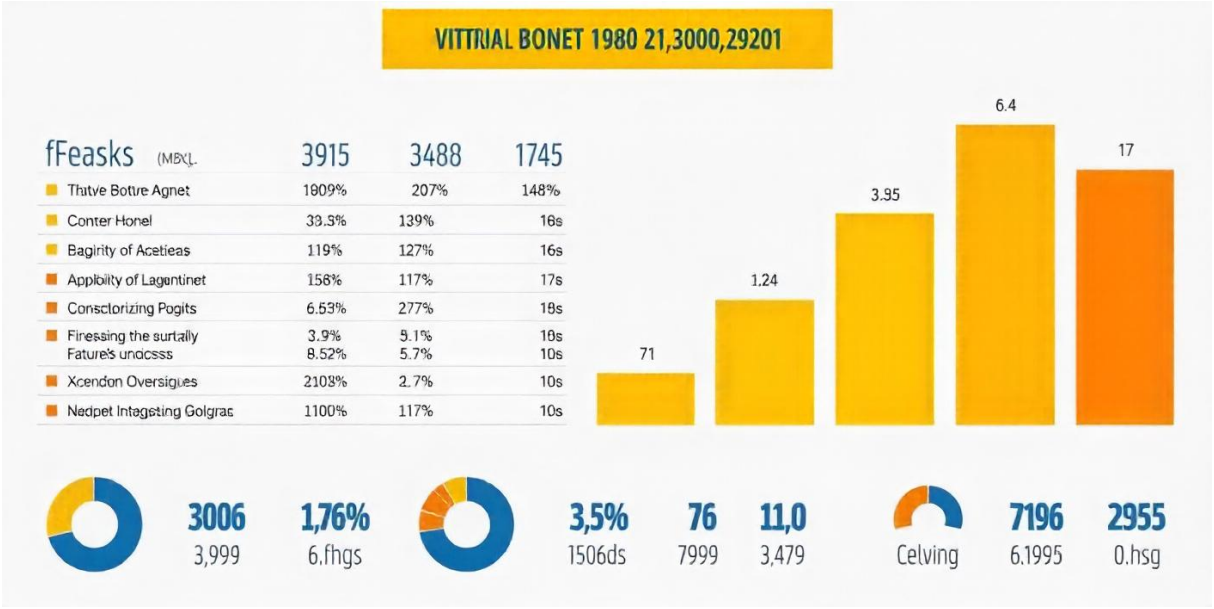


Figura 3: Representación del concepto Datos según inteligencia artificail (IA) freepik

Problema de predicción

Se busca predecir una variable de interés (por ejemplo, la probabilidad de que un usuario elija o compre un producto) en función de múltiples variables predictoras (como el historial de compras, las interacciones anteriores con productos, y características del usuario y del producto). Desde esta perspectiva, el problema de recomendación puede verse como una estimación de la función:

$$\widehat{y} = f(x_1, x_2, ..., x_n)$$

Donde:

$\widehat{y}$ es la predicción (la probabilidad de que el usuario compre o interactúe con un producto).

$x_1, x_2, ..., x_n$ son las variables predictoras que incluyen datos sobre el usuario, el producto y el contexto.

Problema de clasificación

En este tipo de problemas el objetivo es asignar a cada usuario una categoría o clase que corresponda a sus preferencias de productos. El sistema intentará clasificar a los usuarios en base a productos que podrían ser de su interés, según etiquetas o características.

- **Etiquetas:** Productos o categorías de productos que son relevantes o no para un usuario.
- **Características:** Datos demográficos, comportamiento de navegación, historial de compras, entre otros. Así, el problema se convierte en predecir la clase correcta (producto o categoría) para un usuario basado en un conjunto de características.

Problema de estimación de probabilidad condicional

El sistema de recomendación puede modelarse como un problema de estimación de la probabilidad condicional de que un usuario u elija o compre un producto p, dado un conjunto de información $I(u,p)$, que incluye los comportamientos pasados, datos demográficos y características del producto:

$$P(p|I(u,p))$$

El objetivo es maximizar esta probabilidad para recomendar los productos con mayor probabilidad de ser seleccionados por el usuario.

Problema de matriz de descomposición (factorización)

En muchos sistemas de recomendación, especialmente aquellos basados en collaborative filtering (Torre de Silva Fuentes, 2023), el problema se puede ver como una descomposición matricial. Se tiene una matriz R, donde las filas corresponden a usuarios y las columnas a productos. Cada elemento r_{ij} de la matriz indica la interacción entre el usuario i y el producto j (como una calificación, compra, o clic). La tarea es completar la matriz prediciendo los valores faltantes $\widehat{r}_{ij}$ en base a patrones subyacentes. Matemáticamente, el problema puede expresarse como:

$$R \approx U \cdot P^T$$

Donde:

U es una matriz de características de los usuarios.

P es una matriz de características de los productos. La multiplicación de

R es la matriz aproximada al producto $U \cdot P^T$ que permite predecir interacciones faltantes entre usuarios y productos.

Problema de optimización

En términos de un problema de optimización, el sistema de recomendación puede buscar maximizar una función objetivo, como la tasa de conversión o la satisfacción del usuario. Esto implica encontrar los parámetros óptimos que mejoren las recomendaciones y, por ende, los resultados de negocio.

$$Maximizar: \ f(r) = \sum_{i=1}^{n} P(i|D\&C)$$

Ecuación que se traduce mejorar el sistema de recomendaciones a partir de la probabilidad de que se compre un producto en especifico i dada la mejor combinación de datos de usuarios y contexto $D\&C$.

4. ¿Qué métricas o indicadores utilizarías para evaluar la efectividad del sistema de recomendación actual?

Cómo el objetivo es mejorar su sistema de recomendación de productos para aumentar la tasa de conversión y las ventas, entonces utilizaría las métricas de impacto en las ventas y conversión (Gómez-Zorrilla & Piña, 2022):

Tasa de conversión

Proporción de usuarios que realizaron una compra después de interactuar con una recomendación. Esta métrica indica el impacto directo del sistema en las ventas.

$$TC = \frac{UCR}{URR}$$

Figura 4: Representación del concepto Métricas de negocio según inteligencia artificail (IA) Microsof designer

Donde:

TC es la tasa de conversión.

UCR es la cantidad de usuarios que compraron a partir de una recomendación.

URR es la cantidad de usuarios que recibieron recomendaciones.

Valor promedio de la transacción (AOV - Average Order Value)

Mide el valor promedio de las compras realizadas por los usuarios que interactúan con las recomendaciones, para determinar si el sistema está sugiriendo productos que incrementen el gasto.

$$AOV = \frac{TVR}{N}$$

Donde:

AOV es el valor promedio de la transacción

TVR es el total de ventas generadas por las transacciones.

N el número de transacciones.

Tasa de clics (CTR - Click-Through Rate)

Proporción de usuarios que hacen clic en un producto recomendado sobre el total de usuarios que recibieron recomendaciones.

$$CTR = \frac{CPR}{URR}$$

Donde:

TCR es la tasa de clic en productos recomendados

CPR es la cantidad de clic en productos recomendados.

URR es la cantidad de usuarios que recibieron recomendaciones.

5. ¿Qué técnicas estadísticas podrías aplicar para analizar los datos y encontrar patrones en el comportamiento de los usuarios?

Para analizar los datos y encontrar patrones en el comportamiento de los usuarios, se pueden aplicar varias técnicas estadísticas que ayudan a descubrir relaciones y tendencias en los datos. Estas técnicas se pueden agrupar en métodos de análisis descriptivo, inferencial, predictivo y de segmentación. A continuación, te presento algunas técnicas clave:

Análisis Descriptivo: Este tipo de análisis ayuda a resumir y visualizar el comportamiento de los usuarios utilizando estadísticas básicas.

- **Medidas de tendencia central y dispersión**: Promedios, medianas, moda, desviación estándar, y percentiles. Esto te permite entender características generales de la población de usuarios, como el gasto promedio por transacción o el tiempo medio de interacción.
- **Distribución de frecuencias**: Crear histogramas o gráficos de barras para visualizar la frecuencia con la que los usuarios realizan ciertas acciones, como visitas, compras, o clics en productos recomendados.
- **Tablas de contingencia**: Para explorar la relación entre diferentes variables, como la frecuencia de compra y el género, o la categoría de productos preferidos y la edad de los usuarios.

Análisis de correlación: Este análisis ayuda a identificar relaciones entre diferentes variables.

- **Correlación de Pearson**: Mide la relación lineal entre dos variables cuantitativas. Por ejemplo, la correlación entre el tiempo de permanencia en la web y la probabilidad de compra.
- **Correlación de Spearman**: Para medir relaciones monotónicas entre variables que no necesariamente siguen una distribución normal, como la relación entre el número de visitas y la clasificación de productos.

Análisis de regresión El análisis de regresión permite modelar la relación entre una variable dependiente y una o más variables independientes.

- **Regresión lineal**: Para modelar la relación entre variables cuantitativas, como predecir el valor promedio de la compra basado en el número de productos vistos por el usuario.
- **Regresión logística**: Para modelar la probabilidad de que un usuario realice una compra (variable binaria: compra/no compra) en función de variables predictoras, como el tiempo en la página, la ubicación geográfica, o el historial de compras.
- **Regresión multinomial**: Para predecir la probabilidad de que un usuario elija entre múltiples opciones (por ejemplo, elegir entre varias categorías de productos).

Modelos predictivos: Se utilizan para predecir el comportamiento futuro de los usuarios en función de sus interacciones pasadas.

- **Árboles de decisión**: Permiten clasificar a los usuarios en función de características clave para predecir comportamientos como la probabilidad de compra o la categoría de producto preferida. Estos modelos son fáciles de interpretar y pueden capturar relaciones no lineales.
- **Random Forest**: Un conjunto de árboles de decisión que mejora la precisión y la generalización. Es útil para identificar las características más importantes que influyen en las decisiones de compra o en las interacciones de los usuarios.
- **Máquinas de soporte vectorial (SVM)**: Una técnica para clasificar o predecir si un usuario comprará o no un producto, basado en características como el historial de clics, la duración de la sesión, etc.

Técnicas de segmentación: Estas técnicas permiten agrupar usuarios en segmentos o clusters con comportamientos similares, lo que facilita personalizar las recomendaciones.

- **Análisis de clusters (K-means, DBSCAN)**: Agrupa a los usuarios en segmentos con características similares, como comportamiento de compra, frecuencia de visitas, o categorías de productos preferidas. Esto es útil para personalizar recomendaciones para diferentes grupos.
- **Análisis de componentes principales (PCA)**: Una técnica de reducción de dimensionalidad que puede identificar patrones en grandes volúmenes de datos. Ayuda a reducir la complejidad de los datos y a identificar las variables más importantes para el comportamiento del usuario.
- **Análisis de cohortes**: Agrupa a los usuarios según el tiempo o las acciones realizadas, lo que permite analizar su comportamiento a lo largo del tiempo y detectar tendencias o patrones de comportamiento, como el ciclo de vida del cliente.

Modelos de series temporales: Útil para analizar patrones de comportamiento en función del tiempo.

- **Modelos ARIMA**: Se pueden aplicar para predecir ventas o la demanda de productos a lo largo del tiempo, basándose en patrones históricos.
- **Descomposición de series temporales**: Para analizar la tendencia, la estacionalidad y los componentes residuales en los datos de compra y actividad de los usuarios a lo largo del tiempo.

Análisis de mercado (Reglas de asociación): Este tipo de análisis busca descubrir relaciones entre los productos que los usuarios tienden a comprar juntos.

- **Algoritmo Apriori**: Identifica conjuntos de productos que suelen comprarse juntos, como las reglas de asociación del tipo "si el usuario compra A, es probable que también compre B" (también llamado "basket analysis" o análisis de cesta de compra).
- **Análisis de secuencia**: Identifica secuencias comunes de acciones que los usuarios realizan, como los productos que suelen comprar después de ver ciertos productos, lo que puede ayudar a mejorar las recomendaciones.

Modelos de filtrado colaborativo: Este es un enfoque clásico de los sistemas de recomendación que se basa en el comportamiento colectivo de los usuarios.

- **Filtrado colaborativo basado en usuarios**: Identifica usuarios con comportamientos similares y les recomienda productos basados en lo que otros usuarios con gustos parecidos han comprado o calificado positivamente.
- **Filtrado colaborativo basado en ítem**: Recomendaciones basadas en la similitud entre productos. Se recomienda a un usuario productos que son similares a aquellos con los que ha interactuado previamente.

Modelos de factorización matricial: Utilizados para sistemas de recomendación, permiten descomponer matrices de interacciones usuario-producto para identificar patrones latentes de preferencias.

- **SVD (Singular Value Decomposition)**: Es una técnica matemática utilizada para factorizar la matriz de interacciones usuario-producto, reduciendo la dimensionalidad del problema y descubriendo patrones ocultos.

Análisis de supervivencia: Esta técnica evalúa la "vida útil" de un cliente o usuario, y se usa para predecir eventos como el abandono de la plataforma.

- **Modelos de supervivencia**: Predicen el tiempo que un usuario permanecerá activo en el sitio web antes de abandonar o hacer una compra, lo que permite tomar acciones para retener usuarios en riesgo de abandonar.

6. ¿Cómo podrías utilizar la estadística descriptiva para obtener información sobre los productos más populares, las preferencias de los usuarios, etc.?

Se puede usar la **moda** para identificar los productos que más compran los usuarios, esto a partir de una tabla de frecuencia y/o frecuencia relativa. Con la **moda** y los **cuartiles** o la **desviación estándar** se puede determinar un rango de precios preferidos por los usuarios, con cierto nivel de significancia. Se puede identificar tendencias temporales utilizando **series temporales** o estableciendo los productos que más se compran según periodos específicos tales como mensual, trimestral o semestral. Finalmente, también se puede realizar una segmentación básica de los usuarios y establecer el producto que más se vende según cada tipo de usuario.

7. ¿Cómo podrías aplicar técnicas de inferencia estadística para validar hipótesis sobre la efectividad de diferentes algoritmos de recomendación?

Formular las hipótesis:

H_0 (hipótesis nula): No hay diferencia significativa en la efectividad entre los algoritmos de recomendación.

H_1 (hipótesis alternativa): Existe una diferencia significativa en la efectividad entre al menos dos algoritmos de recomendación.

Diseño experimental:

Implementar un A/B testing, donde cada variante representa un algoritmo de recomendación diferente. Se debe asignar aleatoriamente usuarios a cada grupo para evitar sesgos y definir métricas de éxito claras, como tasa de conversión, ingresos por usuario o engagement.

Durante este experimento (A/B testing), se medirían las métricas clave, como:

- Tasa de conversión.
- Valor promedio de las transacciones.
- Interacciones con los productos (clics, agregados al carrito).
- Recompras o lealtad del cliente.
- Los datos recolectados de ambos grupos serán la base para el análisis inferencial.

Recolección de datos:

Registrar las interacciones de los usuarios con las recomendaciones y recopilar datos sobre compras, clics, tiempo de sesión, etc.

Análisis estadístico:

Utilizar *ANOVA* (análisis de varianza) para comparar múltiples algoritmos simultáneamente. Si solo se comparan dos algoritmos, se puede usar una prueba *t de Student*. Calcular el *p-valor* para determinar la significancia estadística.

Validación cruzada:

Del análisis estadístico se podría implementar técnicas de validación cruzada para evaluar la generalización de los resultados.

Interpretación de resultados:

Si $p < 0{,}05$ (nivel de significancia común), rechazar H_0 y concluir que hay diferencias significativas. Realizar pruebas post-hoc (como Tukey HSD) para identificar qué algoritmos difieren entre sí.

8. ¿Qué enfoques de modelado estadístico podrías utilizar para mejorar la precisión de las recomendaciones?

Filtrado Colaborativo (Collaborative Filtering):

Si el modelo de recomendaciones es basado en usuarios o en ítem, entonces se podría utilizar la técnica de *factorización de matrices* (Bojorque Chasi, 2020) o de *SVD* (Descomposición en Valores Singulares).

Modelos de Factores Latentes:

Utiliza técnicas como *LDA* (Latent Dirichlet Allocation) para descubrir temas ocultos en las preferencias de los usuarios, lo que permite capturar relaciones más complejas y abstractas entre usuarios y productos.

Regresión Logística:

Predice la probabilidad de que un usuario compre un producto específico, por lo que es útil para modelar decisiones binarias (comprar o no comprar).

K vecinos más cercanos:

Se trabajó con las preferencias de un usuario y una cantidad significativa de productos de distintos modelos y funcionalidades que son identificadas mediante variables como, color, marca, modelo, precio con los que se calcula la distancia y generar "N" recomendaciones más cercanas a los gustos del cliente (Guevara-Fernandez & Coral-Ygnacio, 2023).

Árboles de Decisión y Bosques Aleatorios:

Capturan relaciones no lineales en los datos. Los bosques aleatorios son especialmente útiles para manejar grandes conjuntos de datos con muchas características.

Gradient Boosting Machines (GBM):

Algoritmos como *XGBoost* o *LightGBM* para mejorar la precisión de las predicciones, por lo que son excelentes para capturar patrones complejos en los datos.

Redes Neuronales y Deep Learning:

Se puede utilizar *autoencoders* para reducción de dimensionalidad y extracción de características y *Redes neuronales profundas* para modelar relaciones altamente no lineales.

Modelos de Series Temporales:

Los modelos del tipo *ARIMA*, *SARIMA* se utilizan para capturar patrones estacionales en las preferencias de los usuarios. Son Útiles para predecir tendencias futuras en las preferencias de productos.

Modelos Bayesianos:

Incorporan incertidumbre y conocimiento previo en las predicciones. Son Útiles cuando se tienen datos limitados sobre nuevos usuarios o productos.

9. ¿Cómo podrías utilizar la validación cruzada y otras técnicas de validación para evaluar y comparar diferentes modelos de recomendación?

Preparación de datos: Limpiar y pre procesar los datos. Dividir en conjuntos de entrenamiento, validación y prueba.

Selección de métricas: Definir métricas relevantes, por ejemplo, NDCG, MAP, Recall@K. Asegurar que las métricas se alineen con los objetivos de negocio.

Implementación de modelos: Entrenar múltiples modelos usando validación cruzada. Registrar el rendimiento en cada fold.

Análisis estadístico: Realizar pruebas de hipótesis (por ejemplo, t-test pareado) para comparar modelos. Calcular intervalos de confianza para las métricas de rendimiento.

Ajuste de hiperparámetros: Usar validación cruzada anidada para ajustar hiperparámetros sin sobre ajuste.

Evaluación final: Evaluar los mejores modelos en el conjunto de holdout.

Monitoreo continuo: Implementar validación online y monitorear el rendimiento en producción.

10. ¿Qué desafíos podrías enfrentar al analizar los datos de la empresa y cómo los abordarías estadísticamente?

Figura 5: Representación del concepto Desafío según inteligencia artificail (IA) freepik

Datos Dispersos: En sistemas de recomendación, la mayoría de los usuarios interactúan solo con una pequeña fracción de los productos disponibles.

Utilizaría:

- Factorización de matrices para reducir la dimensionalidad.
- Técnicas de regularización para evitar el sobreajuste.
- Modelos de embeddings para capturar relaciones latentes.
- Algoritmos de vecinos más cercanos con medidas de similitud ajustadas.

Arranque en Frío: Nuevos usuarios o productos sin historial de interacciones.

Utilizaría:

- Modelos híbridos que incorporen características de contenido.
- Técnicas de transferencia de aprendizaje desde dominios similares.
- Estrategias de exploración-explotación (por ejemplo, bandidos multi-brazo).
- Análisis de características demográficas y de comportamiento para inferir preferencias iniciales.

Sesgos en los Datos: Sesgos de selección, popularidad o presentación en las interacciones registradas.

Utilizaría:

- Técnicas de muestreo estratificado para equilibrar la representación.
- Métodos de corrección de sesgo, como Inverse Propensity Scoring.

- Análisis causal para distinguir entre correlación y causalidad.
- Experimentos A/B cuidadosamente diseñados para evaluar el impacto real de las recomendaciones.

Escalabilidad: Manejar grandes volúmenes de datos y actualizar modelos en tiempo real.

Utilizaría:

- Algoritmos de aprendizaje online para actualizaciones incrementales.
- Técnicas de muestreo para entrenar en subconjuntos representativos.
- Implementación de modelos distribuidos utilizando frameworks como Spark.
- Uso de aproximaciones eficientes para cálculos costosos (por ejemplo, LSH para búsqueda de vecinos).

Temporalidad y Cambios en las Preferencias: Las preferencias de los usuarios y la relevancia de los productos cambian con el tiempo.

Utilizaría:

- Modelos de series temporales para capturar tendencias y estacionalidad.
- Técnicas de olvido exponencial para dar más peso a interacciones recientes.
- Detección de cambios de concepto para identificar cuando las preferencias cambian significativamente.
- Análisis de cohortes para entender cómo evolucionan las preferencias de diferentes grupos de usuarios.

Interpretabilidad: Modelos complejos pueden ser difíciles de interpretar y explicar.

Utilizaría:

- Uso de modelos interpretables como árboles de decisión cuando sea posible.
- Técnicas de explicabilidad post-hoc como SHAP values.
- Análisis de importancia de características para entender qué factores influyen más en las recomendaciones.
- Desarrollo de interfaces de usuario que proporcionen explicaciones claras de las recomendaciones.

Aborde estadístico de forma general

Para abordar estos desafíos de manera efectiva, utilizaría el siguiente plan de acción:

- Análisis exploratorio de datos exhaustivo para identificar la naturaleza y extensión de cada desafío en nuestro conjunto de datos específico.
- Desarrollo de un pipeline de preprocesamiento robusto que aborde los problemas de calidad de datos, sesgos y escalabilidad.
- Implementación de un conjunto de modelos que aborden diferentes aspectos de los desafíos (por ejemplo, modelos híbridos para el arranque en frío, modelos temporales para cambios en preferencias). Diseño de un marco de evaluación multidimensional que capture todos los aspectos relevantes del rendimiento del sistema.
- Implementación de un sistema de monitoreo continuo para detectar y adaptarse a cambios en los patrones de datos y comportamiento de usuarios.
- Colaboración estrecha con los equipos de negocio y UX para asegurar que las soluciones técnicas se alineen con los objetivos de negocio y la experiencia del usuario.

Caso 2: Empresa de streaming de música en línea

Una empresa de streaming de música en línea está interesada en mejorar la experiencia del usuario y retener a sus clientes. Como experto en Big Data, te han asignado el proyecto de analizar los datos de la empresa y proponer mejoras en la recomendación de canciones. La empresa tiene como objetivo principal aumentar el tiempo de reproducción de música de sus usuarios y reducir la tasa de abandono

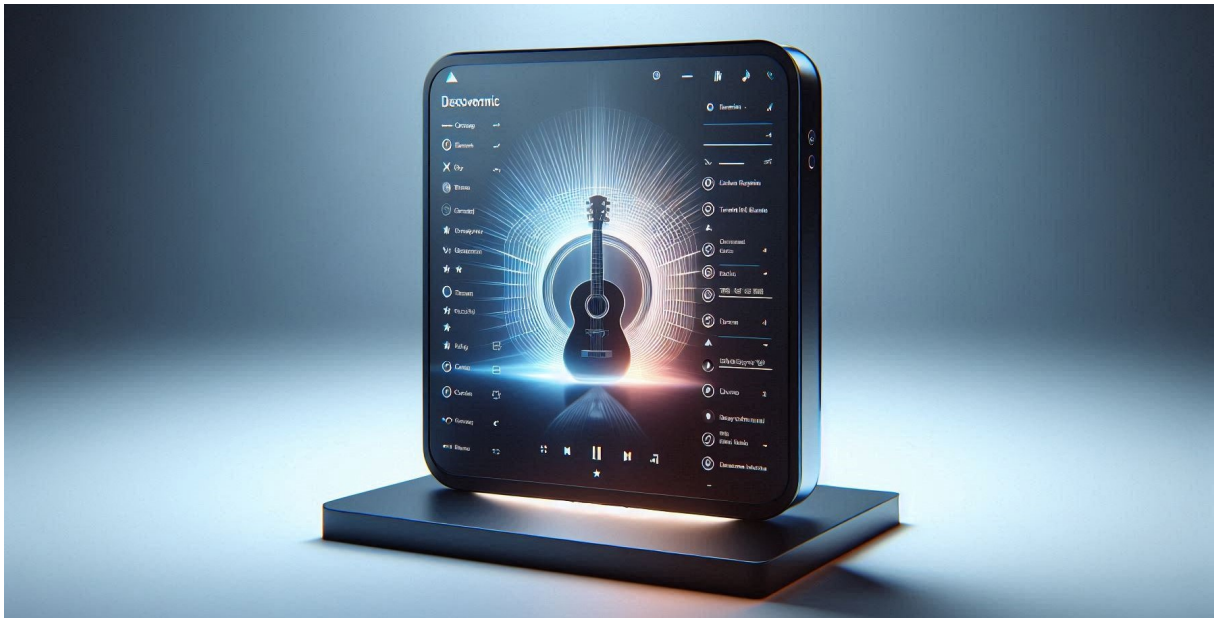


Figura 6: Representación del concepto Música en línea según inteligencia artificail (IA) Microsof designer

1. ¿Cuál es el objetivo principal del análisis de datos en este proyecto?

El objetivo principal del análisis de datos, en este proyecto, es identificar patrones claves en el comportamiento de escucha de los usuarios que nos permitan mejorar significativamente la personalización y eficacia de las recomendaciones musicales. En otras palabras, el fin último es desarrollar un sistema de recomendación más preciso y adaptativo que no solo sugiera canciones que el usuario probablemente disfrute, sino que también optimice la secuencia y el momento de estas recomendaciones. Con esto, buscamos aumentar el tiempo total de reproducción por usuario y, consecuentemente, reducir la tasa de abandono de la plataforma, mejorando así la retención de clientes y la satisfacción general del usuario con el servicio de streaming musical.

2. ¿Qué tipo de datos sería relevante recopilar y analizar para mejorar el sistema de recomendación?

Para mejorar el sistema de recomendación de una empresa de streaming de música en línea, es crucial recopilar y analizar los siguientes tipos de datos:

Datos del usuario

Historial de reproducción: Canciones, álbumes y playlists escuchados. - **Comportamiento de navegación:** Artistas explorados, géneros visitados, búsquedas realizadas.

- **Interacciones con contenido:** Likes, skips, adiciones a playlists personales.
- **Preferencias explícitas:** Calificaciones de canciones, seguimiento de artistas.
- **Datos demográficos:** Edad, género, ubicación.
- **Patrones de uso:** Momentos del día/semana de escucha, duración de sesiones.

Datos de contenido musical

- **Metadatos de canciones:** Género, artista, álbum, año de lanzamiento, duración.
- **Características acústicas:** Tempo, ritmo, instrumentación, energía, valencia.
- **Letras:** Temas, idioma, complejidad lingüística.
- **Popularidad:** Número de reproducciones, tendencias actuales.

Datos de interacción usuario-música

- **Secuencias de reproducción:** Transiciones comunes entre canciones/géneros.
- **Contexto de escucha:** Playlists donde se incluye una canción, mood asociado.
- **Descubrimiento musical:** Nuevos artistas o géneros explorados por el usuario.

Datos contextuales

- **Temporales:** Estación del año, día de la semana, hora del día.
- **Dispositivo:** Tipo de dispositivo usado (móvil, desktop, smart speaker).
- **Actividad:** Modo de reproducción (aleatorio, repetición, radio basada en artista).
- **Interacciones con recomendaciones:** Aceptación o rechazo de sugerencias. Tiempo de escucha: Duración de reproducción de canciones recomendadas.

Datos sociales

- **Conexiones:** Amigos en la plataforma, influencers musicales seguidos.
- **Compartir:** Canciones o playlists compartidas en redes sociales.

3. ¿Cómo definirías el problema de recomendación como un problema estadístico?

De acuerdo al tipo de proyectos, se puede sugerir estas etapas como la comprensión del negocio, la captura de datos, el modelado y, finalmente, la aceptación y puesta en producción (Gensollen, 2022). En la etapa de modelado se tiene al problema de recomendación como un problema estadístico, el cual se puede abordar con diferentes herramientas de acuerdo a los tipos de modelados (Flores, 2023).

Problema de predicción

Se busca predecir una variable de interés (por ejemplo, la probabilidad de que un usuario elija una canción) en función de múltiples variables predictoras (como el historial de canciones escuchadas, las interacciones anteriores con productos, y características del usuario y del producto).

Desde esta perspectiva, el problema de recomendación puede verse como una probabilidad condicional a partir de una matriz de transición o una estimación de una función:

- **Como probabilidad condicional**

Dada una matriz de transición normalizada de dimensiones NxN, donde N es la cantidad de canciones que ha escuchado el usuario:

$$P(e_{ij}) = P(j|i)$$

Donde i y j son canción que ha escuchado el usuario y $P(e_{ij})$ es la probabilidad de que se dé la transición entre la canción i a la canción j, en otras palabras, es la probabilidad de que el usuario escuche la canción j dado que ha escuchado la canción i.

Figura 7: Representación del concepto Datos según inteligencia artificail (IA) freepik

- **Como estimación de una función**

También, se podría calcular la probabilidad de que el usuario escuche una canción en función de múltiples categorías o variables, como la hora del día, el día de la semana, si se es día de descanso o no, el tiempo que lleva sin escuchar la canción, entre otras. Funcionalmente se pude representar como:

$$\widehat{y} = f(x_1, x_2, ..., x_n)$$

Donde:

$\widehat{y}$ es la predicción (la probabilidad de que el usuario compre o interactúe con un producto).

$x_1, x_2, ..., x_n$ son las variables predictoras que incluyen datos sobre el usuario, el producto y el contexto.

Problema de clasificación

Se podría clasificar las canciones según los géneros y a los usuarios según los tipos de canciones que escuchan con mayor frecuencia, algunos usuarios podrían encontrarse en la frontera entre dos o tres tipos de canciones al igual que algunas canciones podrían estar en la frontera entre dos o más géneros musicales. El sistema intentará clasificar a los usuarios en base a productos que podrían ser de su interés, según etiquetas o características.

- **Etiquetas:** Productos o categorías de productos que son relevantes o no para un usuario.
- **Características:** Datos demográficos, comportamiento de navegación, historial de canciones, entre otros. Así, el problema se convierte en predecir la clase correcta (categoría músical) para un usuario basado en un conjunto de características.

Problema de optimización

En términos de un problema de optimización, el sistema de recomendación puede buscar maximizar una función objetivo, como la tasa de conversión (escuchar una canción cuando se le recomienda, cuando la ve en las sugerencias) o la satisfacción del usuario. Esto implica encontrar los parámetros óptimos que mejoren las recomendaciones y, por ende, los resultados de negocio.

$$Maximizar: \ f(r) = \sum_{i=1}^{n} P(i|D\&C)$$

Ecuación que se traduce mejorar el sistema de recomendaciones a partir de la probabilidad de que se compre un producto en especifico i dada la mejor combinación de datos de usuarios y contexto $D\&C$.

4. ¿Qué métricas o indicadores utilizarías para evaluar la efectividad del sistema de recomendación actual?

Figura 8: Representación del concepto Métricas de negocio según inteligencia artificail (IA) Microsof designer

Cómo el objetivo es mejorar su sistema de recomendación de productos para aumentar la tasa de conversión y las ventas, entonces utilizaría las métricas de impacto en las ventas y conversión (Gómez-Zorrilla & Piña, 2022):

Tasa de conversión

Proporción de usuarios que escucharon una canción después de interactuar con una recomendación. Esta métrica indica el impacto directo del sistema en las ventas.

$$TC = \frac{UER}{URR}$$

Donde:

TC es la tasa de conversión.

UER es la cantidad de usuarios que escucharon una canción a partir de una recomendación.

URR es la cantidad de usuarios que recibieron recomendaciones de canciones.

Tasa de clics (CTR - Click-Through Rate)

Proporción de usuarios que hacen clic en el enlace de la canción recomendada sobre el total de usuarios que recibieron estas recomendaciones.

$$CTR = \frac{CER}{URR}$$

Donde:

TCR es la tasa de clic en productos recomendados

CER es la cantidad de clic en productos recomendados.

URR es la cantidad de usuarios que recibieron recomendaciones.

5. ¿Qué técnicas estadísticas podrías aplicar para analizar los datos y encontrar patrones en el comportamiento de los usuarios?

Para analizar los datos y encontrar patrones en el comportamiento de los usuarios, se pueden aplicar varias técnicas estadísticas que ayudan a descubrir relaciones y tendencias en los datos. Estas técnicas se pueden agrupar en métodos de análisis descriptivo, inferencial, predictivo y de segmentación. A continuación, se ven algunas técnicas clave:

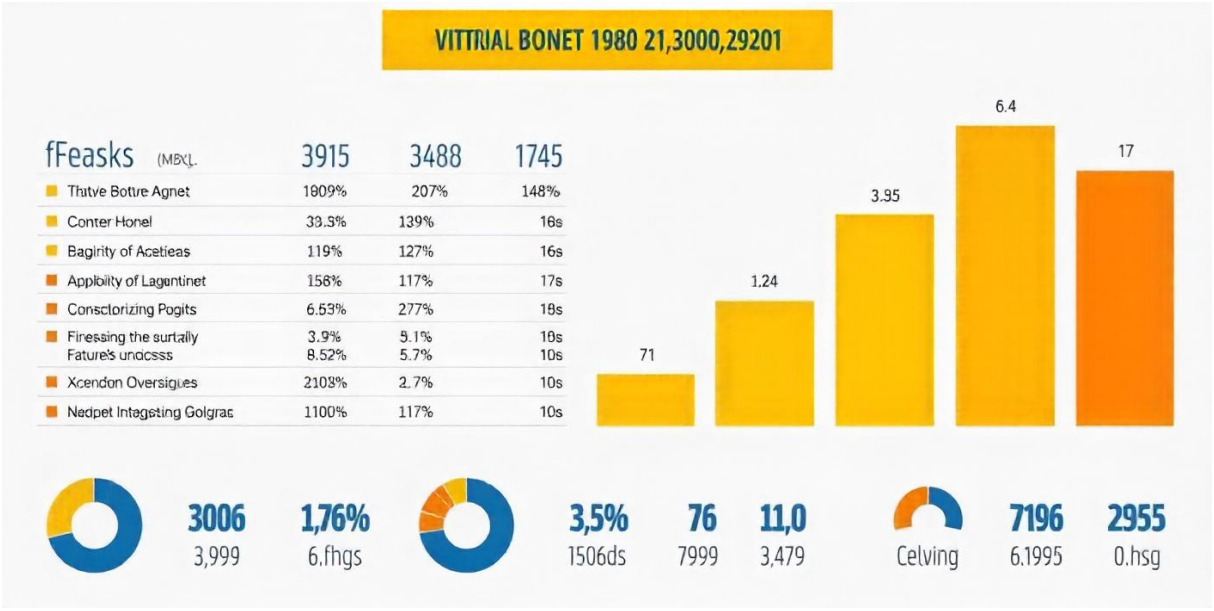


Figura 9: Representación del concepto Datos según inteligencia artificail (IA) freepik

Análisis Descriptivo: Este tipo de análisis ayuda a resumir y visualizar el comportamiento de los usuarios utilizando estadísticas básicas.

- **Medidas de tendencia central y dispersión**: Promedios, medianas, moda, desviación estándar, y percentiles. Esto te permite entender características generales de la población de usuarios, como la cantidad de canciones que escucha o el tiempo medio que dedica a escuchar canciones en la plataforma.
- **Distribución de frecuencias**: Crear histogramas o gráficos de barras para visualizar la frecuencia con la que los usuarios realizan ciertas acciones, como visitas, escuchar una canción, o clics en enlaces a canciones recomendadas.
- **Tablas de contingencia**: Para explorar la relación entre diferentes variables, como la frecuencia de con que se entra en la plataforma, el tiempo que dedica a escuchar canciones, el género de la persona, el género de las canciones y la edad de los usuarios.

Análisis de correlación: Este análisis ayuda a identificar relaciones entre diferentes variables.

- **Correlación de Pearson**: Mide la relación lineal entre dos variables cuantitativas. Por ejemplo, la correlación entre el tiempo de permanencia en la plataforma y la probabilidad de escuchar una canción.
- **Correlación de Spearman**: Para medir relaciones monotónicas entre variables que no necesariamente siguen una distribución normal, como la relación entre el número de visitas y la clasificación de las canciones escuchadas (género).
- **Regresión lineal**: Para modelar la relación entre variables cuantitativas, como predecir el valor promedio de la compra basado en el número de productos vistos por el usuario.
- **Regresión logística**: Para modelar la probabilidad de que un usuario una canción (variable binaria: escuchar/no escuchar) en función de variables predictoras, como el tiempo en la plataforma, la ubicación geográfica, o el historial de canciones escuchadas.

- **Regresión multinomial**: Para predecir la probabilidad de que un usuario elija entre múltiples opciones (por ejemplo, elegir entre varias categorías de canciones).

Análisis de regresión El análisis de regresión permite modelar la relación entre una variable dependiente y una o más variables independientes.

Modelos predictivos: Se utilizan para predecir el comportamiento futuro de los usuarios en función de sus interacciones pasadas (canciones escuchadas).

- **Árboles de decisión**: Permiten clasificar a los usuarios en función de características clave para predecir comportamientos como la probabilidad de escuchar una canción o el género musical preferido. Estos modelos son fáciles de interpretar y pueden capturar relaciones no lineales.
- **Random Forest**: Un conjunto de árboles de decisión que mejora la precisión y la generalización. Es útil para identificar las características más importantes que influyen en las decisiones de escuchar una canción o en las interacciones de los usuarios.
- **Máquinas de soporte vectorial (SVM)**: Una técnica para clasificar o predecir si un usuario escuchará o no una canción, basado en características como el historial de clics, la duración de la sesión, etc.

Técnicas de segmentación: Estas técnicas permiten agrupar usuarios en segmentos o clusters con comportamientos similares, lo que facilita personalizar las recomendaciones.

- **Análisis de clusters (K-means, DBSCAN)**: Agrupa a los usuarios en segmentos con características similares, como frecuencia de visitas, o género de las canciones preferidas. Esto es útil para personalizar recomendaciones para diferentes grupos.
- **Análisis de componentes principales (PCA)**: Una técnica de reducción de dimensionalidad que puede identificar patrones en grandes volúmenes de datos. Ayuda a reducir la complejidad de los datos y a identificar las variables más importantes para el comportamiento del usuario.
- **Análisis de cohortes**: Agrupa a los usuarios según el tiempo o las acciones realizadas, lo que permite analizar su comportamiento a lo largo del tiempo y detectar tendencias o patrones de comportamiento, como el ciclo de vida del cliente.

Modelos de series temporales: Útil para analizar patrones de comportamiento en función del tiempo.

- **Modelos ARIMA**: Se pueden aplicar para predecir canciones escuchadas a lo largo del tiempo, basándose en patrones históricos.
- **Descomposición de series temporales**: Para analizar la tendencia, la estacionalidad y los componentes residuales en los datos de canciones escuchadas y actividad de los usuarios a lo largo del tiempo.

Análisis de mercado (Reglas de asociación): Este tipo de análisis busca descubrir relaciones entre los generos de canciones que los usuarios tienden a escuchar en conjunto.

- **Análisis de secuencia**: Identifica secuencias comunes de acciones que los usuarios realizan, como las canciones que suelen escuchar después de ver la imagen de ciertas canciones, lo que puede ayudar a mejorar las recomendaciones.

Modelos de filtrado colaborativo: Este es un enfoque clásico de los sistemas de recomendación que se basa en el comportamiento colectivo de los usuarios (Jiménez González & Martínez Tomás, 2024).

- **Filtrado colaborativo basado en usuarios**: Identifica usuarios con comportamientos similares y les recomienda productos basados en lo que otros usuarios con gustos parecidos han comprado o calificado positivamente.

- **Filtrado colaborativo basado en ítem (en la canción)**: Recomendaciones basadas en la similitud entre canciones. Se recomienda a un usuario canciones que son similares a aquellas con los que ha interactuado previamente.

Modelos de factorización matricial: Utilizados para sistemas de recomendación, permiten descomponer matrices de interacciones usuario-producto (usuario-canción) para identificar patrones latentes de preferencias.

- **SVD (Singular Value Decomposition)**: Es una técnica matemática utilizada para factorizar la matriz de interacciones usuario-producto, reduciendo la dimensionalidad del problema y descubriendo patrones ocultos.

Análisis de supervivencia: Esta técnica evalúa la "vida útil" de un cliente o usuario, y se usa para predecir eventos como el abandono de la plataforma.

- **Modelos de supervivencia**: Predicen el tiempo que un usuario permanecerá activo en el sitio web antes de abandonar o escuchar una canción, lo que permite tomar acciones para retener usuarios en riesgo de abandonar.

6. ¿Cómo podrías utilizar la estadística descriptiva para obtener información sobre los productos más populares, las preferencias de los usuarios, etc.?, l

Se puede usar la **moda** para identificar los productos que más compran los usuarios, esto a partir de una tabla de frecuencia y/o frecuencia relativa. Con la **moda** y los **cuartiles** o la **desviación estándar** se puede determinar un rango de precios preferidos por los usuarios, con cierto nivel de significancia. Se puede identificar tendencias temporales utilizando **series temporales** o estableciendo los productos que más se compran según periodos específicos tales como mensual, trimestral o semestral. Finalmente, también se puede realizar una segmentación básica de los usuarios y establecer el producto que más se vende según cada tipo de usuario.

7. ¿Cómo podrías aplicar técnicas de inferencia estadística para validar hipótesis sobre la efectividad de diferentes algoritmos de recomendación?

Formular las hipótesis:

H_0 (hipótesis nula): No hay diferencia significativa en la efectividad entre los algoritmos de recomendación.

H_1 (hipótesis alternativa): Existe una diferencia significativa en la efectividad entre al menos dos algoritmos de recomendación.

Diseño experimental:

Implementar un A/B testing, donde cada variante representa un algoritmo de recomendación diferente. Se debe asignar aleatoriamente usuarios a cada grupo para evitar sesgos y definir métricas de éxito claras, como tasa de conversión.

Durante este experimento (A/B testing), se medirían las métricas clave, como:

- Tasa de conversión.
- Interacciones con los enlaces a las canciones.
- La veces que el usuario escucha una canción.
- Los datos recolectados de ambos grupos serán la base para el análisis inferencial.

Recolección de datos:

Registrar las interacciones de los usuarios con las recomendaciones y recopilar datos sobre canciones escuchadas, clics en enlaces de canciones recomendadas, tiempo de sesión, etc.

Análisis estadístico:

Utilizar *ANOVA* (análisis de varianza) para comparar múltiples algoritmos simultáneamente. Si solo se comparan dos algoritmos, se puede usar una prueba *t de Student*. Calcular el *p-valor* para determinar la significancia estadística.

Validación cruzada:

Del análisis estadístico se podría implementar técnicas de validación cruzada para evaluar la generalización de los resultados.

Interpretación de resultados:

Si $p < 0{,}05$ (nivel de significancia común), rechazar H_0 y concluir que hay diferencias significativas. Realizar pruebas post-hoc (como Tukey HSD) para identificar qué algoritmos difieren entre sí.

8. ¿Qué enfoques de modelado estadístico podrías utilizar para mejorar la precisión de las recomendaciones?

Filtrado Colaborativo (Collaborative Filtering):

Si el modelo de recomendaciones es basado en usuarios o en canciones (ítems), entonces se podría utilizar la técnica de *factorización de matrices* o de *SVD* (Descomposición en Valores Singulares).

Modelos de Factores Latentes:

Utiliza técnicas como *LDA* (Latent Dirichlet Allocation) para descubrir temas ocultos en las preferencias de los usuarios, lo que permite capturar relaciones más complejas y abstractas entre usuarios y canciones.

Regresión Logística:

Predice la probabilidad de que un usuario escuche una canción en específica, por lo que es útil para modelar decisiones binarias (escuchar o no escuchar).

Árboles de Decisión y Bosques Aleatorios:

Capturan relaciones no lineales en los datos. Los bosques aleatorios son especialmente útiles para manejar grandes conjuntos de datos con muchas características.

Gradient Boosting Machines (GBM):

Algoritmos como *XGBoost* o *LightGBM* para mejorar la precisión de las predicciones, por lo que son excelentes para capturar patrones complejos en los datos.

Redes Neuronales y Deep Learning:

Se puede utilizar *autoencoders* para reducción de dimensionalidad y extracción de características y *Redes neuronales profundas* para modelar relaciones altamente no lineales (Vega Moreno, 2021).

Modelos de Series Temporales:

Los modelos del tipo *ARIMA*, *SARIMA* se utilizan para capturar patrones estacionales en las preferencias de los usuarios. Son Útiles para predecir tendencias futuras en las preferencias de de las canciones.

Modelos Bayesianos:

Incorporan incertidumbre y conocimiento previo en las predicciones. Son Útiles cuando se tienen datos limitados sobre nuevos usuarios o canciones.

9. ¿Cómo podrías utilizar la validación cruzada y otras técnicas de validación para evaluar y comparar diferentes modelos de recomendación?

Preparación de datos: Limpiar y pre procesar los datos. Dividir en conjuntos de entrenamiento, validación y prueba.

Selección de métricas: Definir métricas relevantes, por ejemplo, NDCG, MAP, Recall@K. Asegurar que las métricas se alineen con los objetivos de negocio.

Implementación de modelos: Entrenar múltiples modelos usando validación cruzada. Registrar el rendimiento en cada fold.

Análisis estadístico: Realizar pruebas de hipótesis (por ejemplo, t-test pareado) para comparar modelos. Calcular intervalos de confianza para las métricas de rendimiento.

Ajuste de hiperparámetros: Usar validación cruzada anidada para ajustar hiperparámetros sin sobre ajuste.

Evaluación final: Evaluar los mejores modelos en el conjunto de holdout.

Monitoreo continuo: Implementar validación online y monitorear el rendimiento en producción.

10. ¿Qué desafíos podrías enfrentar al analizar los datos de la empresa y cómo los abordarías estadísticamente?

Datos Dispersos: En sistemas de recomendación, la mayoría de los usuarios interactúan solo con una pequeña fracción de las canciones disponibles.

Utilizaría:

- Factorización de matrices para reducir la dimensionalidad.
- Técnicas de regularización para evitar el sobreajuste.
- Modelos de embeddings para capturar relaciones latentes.
- Algoritmos de vecinos más cercanos con medidas de similitud ajustadas.

Figura 10: Representación del concepto Desafío según inteligencia artificail (IA) freepik

Arranque en Frío: Nuevos usuarios o productos sin historial de interacciones.

Utilizaría:

- Modelos híbridos que incorporen características de contenido.
- Técnicas de transferencia de aprendizaje desde dominios similares.

- Estrategias de exploración-explotación (por ejemplo, bandidos multi-brazo).
- Análisis de características demográficas y de comportamiento para inferir preferencias iniciales.

Sesgos en los Datos: Sesgos de selección, popularidad o presentación en las interacciones registradas.

Utilizaría:

- Técnicas de muestreo estratificado para equilibrar la representación.
- Métodos de corrección de sesgo, como Inverse Propensity Scoring.
- Análisis causal para distinguir entre correlación y causalidad.
- Experimentos A/B cuidadosamente diseñados para evaluar el impacto real de las recomendaciones.

Escalabilidad: Manejar grandes volúmenes de datos y actualizar modelos en tiempo real.

Utilizaría:

- Algoritmos de aprendizaje online para actualizaciones incrementales.
- Técnicas de muestreo para entrenar en subconjuntos representativos.
- Implementación de modelos distribuidos utilizando frameworks como Spark.
- Uso de aproximaciones eficientes para cálculos costosos (por ejemplo, LSH para búsqueda de vecinos).

Temporalidad y Cambios en las Preferencias: Las preferencias de los usuarios y la relevancia de las canciones cambian con el tiempo.

Utilizaría:

- Modelos de series temporales para capturar tendencias y estacionalidad.
- Técnicas de olvido exponencial para dar más peso a interacciones recientes.
- Detección de cambios de concepto para identificar cuando las preferencias cambian significativamente.
- Análisis de cohortes para entender cómo evolucionan las preferencias de diferentes grupos de usuarios.

Interpretabilidad: Modelos complejos pueden ser difíciles de interpretar y explicar.

Utilizaría:

- Uso de modelos interpretables como árboles de decisión cuando sea posible.
- Técnicas de explicabilidad post-hoc como SHAP values.
- Análisis de importancia de características para entender qué factores influyen más en las recomendaciones.
- Desarrollo de interfaces de usuario que proporcionen explicaciones claras de las recomendaciones.

Aborde estadístico de forma general

Para abordar estos desafíos de manera efectiva, utilizaría el siguiente plan de acción:

- Análisis exploratorio de datos exhaustivo para identificar la naturaleza y extensión de cada desafío en nuestro conjunto de datos específico.

- Desarrollo de un pipeline de preprocesamiento robusto que aborde los problemas de calidad de datos, sesgos y escalabilidad.
- Implementación de un conjunto de modelos que aborden diferentes aspectos de los desafíos (por ejemplo, modelos híbridos para el arranque en frío, modelos temporales para cambios en preferencias). Diseño de un marco de evaluación multidimensional que capture todos los aspectos relevantes del rendimiento del sistema.
- Implementación de un sistema de monitoreo continuo para detectar y adaptarse a cambios en los patrones de datos y comportamiento de usuarios.
- Colaboración estrecha con los equipos de negocio y UX para asegurar que las soluciones técnicas se alineen con los objetivos de negocio y la experiencia del usuario.

Referencias

Barragán, M. S., Chanchí, G. G. E., & Campo, W. M. Y. (2020). Sistema de recomendación para contenidos musicales basado en el análisis afectivo del contexto social. *Revista Ibérica de Sistemas y Tecnologias de Información*, *39*, 100–113.

Bojorque Chasi, R. X. (2020). *Clustering de sistemas de recomendación mediante técnicas de factorization matricial* [PhD thesis]. Universidad Politeçnica de Madrid.

Crisostomo Madueño, O., & Garavito Cruzado, M. J. (2021). *Sistema web para el proceso de ventas por delivery en la empresa la carpita SAC*.

Flores, M. G. (2023). Sistemas de recomendación: Análisis e implementación de modelos principales. *Cartografías Del Sur. Revista de Ciencias, Artes y Tecnología*, *18*.

Fonseca, B. B., & Cornelio, O. M. (2022). Sistemas de recomendación para la toma de decisiones. Estado del arte: Sistemas de recomendación para la toma de decisiones. *UNESUM-Ciencias. Revista Científica Multidisciplinaria*, *6*(1), 149–164.

Gensollen, C. R. C. (2022). Big data en el mundo del retail: Segmentación de clientes y sistema de recomendación en una cadena de supermercados de europa. *Ingeniería Industrial*, 189–216.

Gómez-Zorrilla, J., & Piña, D. S. (2022). *Guía práctica de analítica digital: ROI, KPI y métricas. Cómo medir y optimizar tu estrategia digital para potenciar tu negocio.* LID Editorial.

Guevara-Fernandez, A., & Coral-Ygnacio, M. A. (2023). Sistema de recomendación de artículos de línea blanca basado en el algoritmo KNN. *Revista Científica de Sistemas e Informática*, *3*(2), e557–e557.

Jiménez González, L., & Martínez Tomás, L. (2024). *Recomendación personalizada de canciones*.

Pajuelo Holguera, F. (2021). *Sistemas de recomendación basados en filtrado colaborativo: Aceleración mediante computación reconfigurable y aplicaciones predictivas sensoriales*.

Torre de Silva Fuentes, M. (2023). *Efecto de las campañas de marketing sesgadas en algoritmos de collaborative filtering usados en sistemas de recomendación*.

Vega Moreno, B. D. (2021). *Diseño y desarrollo de un sistema de recomendación basado en filtrado colaborativo utilizando datos secuenciales mediante redes neuronales recurrentes* [{B.S.} thesis].

Zelcer, M. (2023). Sistemas de recomendación en plataformas de streaming audiovisual: Las lógicas de los algoritmos. *Mídia E Cotidiano*, *17*(2).