Tecnológico de Monterrey

Análisis de grandes volúmenes de datos

TC4034.10

Evidencia 1
Proyecto | Base de Datos de Big Data

Equipo 37:

Alejandro Díaz Villagómez	A01276769
Alonso Pedrero Martínez	A01769076
César Iván Pedrero Martínez	A01366501
Emiliano Saucedo Arriola	A01659258

Docentes:
Dr. Iván Olmos Pineda
Mtra. Verónica Sandra Guzmán de Valle
Mtro. Alberto Daniel Salinas Montemayor

Fecha de entrega:
4 de mayo del 2025

Caracterización de la población

La caracterización de la población es una etapa fundamental en el diseño de una estrategia de muestreo robusta. Su propósito es analizar en profundidad las propiedades generales de la población objetivo (P), con el fin de identificar comportamientos, atributos y patrones recurrentes. Para lograrlo, se seleccionan variables de caracterización que permiten representar de manera significativa las diferencias y similitudes dentro de la población.

En este proyecto, la población objetivo está conformada por usuarios de Amazon que han emitido reseñas de productos en la categoría "PC". La base de datos empleada contiene más de 6.9 millones de registros y 16 columnas. A partir de su análisis, se han identificado aquellas variables que contienen información relevante sobre el comportamiento de estos usuarios. La selección se realiza considerando criterios como: valor explicativo o comportamental, dominio acotado o discretizable, presencia significativa de datos (sin valores nulos) y potencial utilidad para segmentar o representar a la población.

La importancia de esta caracterización radica en varios aspectos clave. Por mencionar algunos:

- ***Base para la estrategia de muestreo***: Al comprender las características de la población, podemos diseñar una estrategia de muestreo que garantice la representatividad, determinando tanto el tamaño de muestra adecuado como la técnica más apropiada.
- ***Fundamento para el particionamiento***: Una caracterización adecuada permite identificar segmentos homogéneos dentro de la población, facilitando la implementación de estrategias de particionamiento efectivas que capturen los diferentes comportamientos de interés.
- ***Identificación de variables críticas***: Nos permite distinguir qué variables tienen mayor influencia en el análisis de sentimientos y patrones en las reseñas, orientando el enfoque analítico posterior.
- ***Detección de patrones y anomalías***: Facilita la identificación de comportamientos típicos y atípicos que podrían requerir tratamiento especial durante el análisis.

Variables no consideradas para la caracterización

De nuestro conjunto de datos original, las siguientes variables no serán consideradas para la caracterización de la población por las razones que se detallan:

Nombre	Comentarios
customer_id	Aunque es una variable numérica con 4,055,429 valores únicos, funciona como un identificador y no caracteriza comportamientos colectivos de la población. Si bien podría utilizarse para análisis de clientes recurrentes en etapas posteriores, no aporta información directa sobre patrones de comportamiento en las reseñas.
product_parent	Con 382,279 valores únicos, actúa principalmente como un identificador de

	agrupación de productos similares. No caracteriza directamente el comportamiento de las reseñas ni aporta información relevante para el análisis de sentimientos, por lo que no se considera para la caracterización inicial de la población.
marketplace	Solo contiene un valor único ("US"), por lo que no aporta variabilidad para la caracterización. Aunque define el contexto geográfico de todas las reseñas (mercado estadounidense), no proporciona información sobre comportamientos variables dentro de la población
product_category	Contiene un único valor ("PC"), por lo que no aporta variabilidad para la caracterización. Aunque contextualiza el segmento de mercado estudiado (productos electrónicos/computadoras), al ser constante no permite identificar patrones diferenciales dentro de la población.
review_id	Identificador único para cada reseña (6,906,564 valores únicos). No caracteriza comportamientos colectivos sino que funciona como clave primaria del dataset.
product_id	Identificador único de producto con 441,903 valores distintos. Aunque podría ser útil para análisis a nivel de producto específico, su alta cardinalidad lo hace inadecuado para la caracterización general de la población.
product_title	Descripción textual del producto con 385,803 valores únicos. Su alta cardinalidad y naturaleza de texto libre lo hacen inadecuado para caracterización estadística inicial, aunque podría ser valioso para análisis de texto posteriores.
review_headline	Títulos de las reseñas con 3,389,509 valores únicos. Aunque los más frecuentes ("Five Stars", "Four Stars") ofrecen cierta información sobre tendencias generales, su alta cardinalidad y naturaleza de texto libre lo hacen más adecuado para análisis de texto específicos que para caracterización estadística.
review_body	Contenido principal de las reseñas con 6,384,933 valores únicos. Es el núcleo para el análisis de sentimiento futuro, pero no es adecuado para la caracterización estadística inicial debido a su naturaleza de texto libre y extremadamente alta cardinalidad.

Variables relevantes para la caracterización

Las siguientes variables han sido identificadas como relevantes para caracterizar los patrones de comportamiento en las reseñas de Amazon:

Nombre	Dominio (posibles valores)	Descripción estadística	Comentarios adicionales
star_rating	{1, 2, 3, 4, 5}	Media: 4.09, DE: 1.36, Mediana: 5, Min: 1,	Variable central para medir la satisfacción del cliente. La distribución muestra un fuerte sesgo positivo (mediana=5), lo que indica que la mayoría de los usuarios tienden a dar

		Max: 5, V. únicos: 5	calificaciones altas. Este sesgo es un comportamiento típico en plataformas de reseñas y debe considerarse al diseñar la estrategia de muestreo para garantizar representatividad de todas las valoraciones.
helpful_votes	Enteros ≥ 0	Media: 1.48, DE: 1.49, Mediana: 0, Min: 0, Max: 621, V. únicos: 1169	Indicador clave del impacto social de una reseña. La distribución es extremadamente asimétrica (long-tailed): la mayoría de reseñas reciben 0 votos útiles, mientras que muy pocas reciben gran cantidad. Esta variable puede ser crucial para identificar reseñas influyentes vs. ignoradas. La presencia de 1169 valores únicos sugiere patrones potencialmente significativos en el comportamiento de voto de la comunidad.
total_votes	Enteros ≥ 0	Media: 1.96, DE: 3.07, Mediana: 0, Min: 0, Max: 836, V. únicos: 1216	Complementa la información de helpful_votes proporcionando el contexto total de la visibilidad de la reseña. Su distribución también es altamente asimétrica. La relación entre helpful_votes y total_votes permite calcular un ratio de utilidad percibida. La presencia de 1216 valores únicos sugiere una amplia variabilidad en el nivel de interacción que generan las distintas reseñas.
sentiment	{0, 1}	0: negativo, 1: positivo. Media: 0.76, DE: 0.42, V. únicos: 2	Variable objetivo binaria donde 1 representa sentimiento positivo y 0 negativo. La proporción de sentimientos positivos (76%) es consistente con la tendencia de calificaciones altas en star_rating, lo que sugiere una buena correlación entre valoración numérica y análisis textual. Esta variable es fundamental para el análisis de sentimientos y la identificación de patrones en las opiniones.
verified_purchase	{'Y', 'N'}	Y: 6,047,075 (87.6%) N: 859,489 (12.4%) V. únicos: 2	Indicador crítico de la credibilidad de la reseña. La gran mayoría (87.6%) son compras verificadas, lo que sugiere un alto nivel de autenticidad en el conjunto de datos. Esta variable permite distinguir entre opiniones de compradores reales versus posibles reseñas sesgadas o falsas, siendo fundamental para evaluar la fiabilidad de los sentimientos expresados.
vine	{'Y', 'N'}	N: 6,870,336 (99.5%) Y: 36,228 (0.5%) V. únicos: 2	Identifica reseñas del programa Vine (reseñadores incentivados). Solo el 0.5% de las reseñas provienen de este programa, lo que indica que la gran mayoría son orgánicas. Esta variable permite analizar posibles diferencias en patrones de evaluación entre reseñadores regulares y los del programa especial, ayudando a identificar potenciales sesgos.

review_date	Fechas entre 1999 y 2015	V. únicos: 5,848	Variable temporal que permite analizar patrones y tendencias a lo largo del tiempo. Con 5,848 valores únicos (fechas distintas), esta variable es esencial para identificar estacionalidad, evolución de sentimientos, o respuestas a eventos externos. La distribución por fechas muestra cierta concentración en períodos específicos, lo que podría indicar comportamientos estacionales en las reseñas.

Selección de variables para el particionamiento

Aunque las siete variables identificadas previamente son relevantes para la caracterización de la población, para la etapa de particionamiento seleccionaremos específicamente las siguientes variables: star_rating, verified_purchase y vine. Esta selección se justifica por las siguientes razones:

a) Relevancia directa para el objetivo del proyecto:

- star_rating es la expresión numérica directa de la satisfacción del cliente, estrechamente relacionada con nuestro objetivo de analizar sentimientos.
- verified_purchase es un indicador crucial de la credibilidad de las reseñas, permitiendo diferenciar entre opiniones potencialmente más confiables y aquellas que podrían tener otros motivos.
- vine señala si la reseña fue realizada en el marco del programa Vine, donde Amazon invita a ciertos usuarios a escribir reseñas sobre productos gratuitos. Esta variable permite explorar cómo las reseñas patrocinadas difieren de las reseñas orgánicas, ofreciendo un ángulo crucial para el análisis de sesgos o patrones incentivados

b) Balance entre representatividad y complejidad:

- Utilizar las siete variables generaría una explosión combinatoria de particiones, dificultando tanto el análisis como la interpretación por lo que sólo se ocupan las únicas variables categóricas que tienen diferencia en sus valores.
- La combinación de estas tres variables específicas permite crear segmentos interpretables y significativos sin fragmentar excesivamente la población.

c) Alineación con preguntas de investigación:

- Estas variables permiten responder preguntas fundamentales como:
    - ¿Cómo se relaciona la credibilidad de la reseña con la valoración otorgada?
    - ¿Cómo la valoración otorgada varía con el título y la descripción de una reseña del producto?

d) Equilibrio en los tamaños de partición:

- Aunque verified_purchase tiene un desbalance (87.6% vs 12.4%), sigue ofreciendo particiones con suficientes datos para análisis estadísticamente significativos.
- vine representa una minoría (0.52% "Y"), pero sus valores siguen siendo relevantes por el efecto cualitativo que pueden tener en el análisis, especialmente al estudiar la influencia del patrocinio en las opiniones de los usuarios

Particionamiento

El particionamiento constituye una etapa crítica en la construcción de una muestra representativa (M), ya que permite dividir la población (P) en subconjuntos homogéneos a partir de las variables de caracterización seleccionadas. Esta división posibilita una mejor cobertura de la variabilidad presente en los datos, asegurando que la muestra no esté sesgada hacia ningún grupo específico.

El propósito del particionamiento es, por tanto, estructurar la base de datos (D) en bloques significativos, dentro de los cuales se pueden aplicar posteriormente técnicas de muestreo adecuadas para extraer datos con un nivel aceptable de representatividad. La correcta elección de variables para el particionamiento evita combinaciones redundantes o poco informativas, y permite generar reglas sólidas y reproducibles para la extracción de datos.

Probabilidades individuales de ocurrencia

- ***Probabilidades de star_rating:***

star_rating	probability
1	0.1099
2	0.0526
3	0.0748
4	0.1699
5	0.5928

- ***Probabilidades de verified_purchase:***

verified_purchase	probability
N	0.1244
Y	0.8756

- ***Probabilidades de vine:***

vine	probability
Y	0.0052
N	0.9948

Combinaciones de particionamiento

El espacio de particionamiento se obtiene mediante el producto cartesiano de los dominios de ambas variables, generando un total de $5 \times 2 \times 2 = 20$ combinaciones posibles.

Las probabilidades y el tamaño estimado de cada partición fueron calculados bajo la hipótesis de independencia entre star_rating, verified_purchase y vine. Los valores numéricos corresponden a estimaciones teóricas sobre una base de ~6.9 millones de registros. Las frecuencias reales pueden diferir ligeramente.

Partition Key	Partition rule	Probability	Partition rows
R5_VPY_VN	5 star * Yes verified purchase * No vine	0.5163565905	3566249.839
R4_VPY_VN	4 star * Yes verified purchase * No vine	0.1479908649	1022108.38
R3_VPY_VN	3 star * Yes verified purchase * No vine	0.06515430662	449992.3886
R2_VPY_VN	2 star * Yes verified purchase * No vine	0.04581706589	316438.4978
R1_VPY_VY	1 star * Yes verified purchase * No vine	0.09572805211	661151.9185
R5_VPY_VY	5 star * Yes verified purchase * Yes vine	0.002699089536	18641.43462
R4_VPY_VY	4 star * Yes verified purchase * Yes vine	0.000773575088	5342.745854
R3_VPY_VY	3 star * Yes verified purchase * Yes vine	0.000340573376	2352.191818
R2_VPY_VY	2 star * Yes verified purchase * Yes vine	0.000239494112	1654.081412
R1_VPY_VY	1 star * Yes verified purchase * Yes vine	0.000500387888	3455.960973
R5_VPN_VN	5 star * No verified purchase * No vine	0.07336084954	506671.4024
R4_VPN_VN	4 star * No verified purchase * No vine	0.02102565509	145215.0325
R3_VPN_VN	3 star * No verified purchase * No vine	0.009256733376	63932.22149
R2_VPN_VN	2 star * No verified purchase * No vine	0.006509414112	44957.68517
R1_VPN_VN	1 star * No verified purchase * No vine	0.01360046789	93932.5019
R5_VPN_VY	5 star * No verified purchase * Yes vine	0.000383470464	2648.463302
R4_VPN_VY	4 star * No verified purchase * Yes vine	0.000109904912	759.0653086
R3_VPN_VY	3 star * No verified purchase * Yes vine	0.000048386624	334.1853154
R2_VPN_VY	2 star * No verified purchase * Yes vine	0.000034025888	235.0019731
R1_VPN_VY	1 star * No verified purchase * Yes vine	0.000071092112	491.0022214

Partition Probability spreadsheet

Jupyter Notebook

Con base en las reglas de particionamiento definidas en la sección anterior, se desarrolló un cuaderno Jupyter en el que se implementa el proceso de filtrado y extracción de submuestras a partir de la base de datos original D, utilizando PySpark como motor de procesamiento.

El notebook completo junto con el conjunto de datos procesado, ha sido subido a Google Drive y está disponible a través del siguiente enlace:

- *Evidencia1_Equipo37.ipynb*

El análisis de las particiones generadas revela patrones significativos en la distribución de las reseñas de Amazon para productos de la categoría PC. La estrategia de particionamiento basada en star_rating, verified_purchase y vine ha permitido segmentar efectivamente la población en subconjuntos con características distintivas y volúmenes de datos representativos.

Los resultados muestran un fuerte desequilibrio en la distribución, siendo la partición R5_VPY_VN (reseñas con 5 estrellas, compra verificada y no provenientes del programa Vine) la más numerosa, con 3,679,909 registros, lo cual representa más del 50% del total del conjunto de datos. Esta concentración confirma la tendencia observada durante la etapa de caracterización: la mayoría de las reseñas en esta categoría son positivas y provienen de compradores verificados no vinculados al programa Vine.

En el extremo opuesto, se encuentran particiones como R1_VPN_VY o R2_VPN_VY, con apenas 705 y 1,634 registros respectivamente. Estas particiones representan casos atípicos o de baja ocurrencia (por ejemplo, usuarios del programa Vine con reseñas negativas no verificadas), lo cual plantea desafíos para el muestreo representativo y sugiere la necesidad de técnicas que aseguren equilibrio estadístico sin sacrificar la diversidad de comportamiento.

A pesar del fuerte sesgo hacia las valoraciones positivas, la inclusión de la variable vine en el particionamiento permite capturar con mayor fidelidad la estructura contextual de las reseñas, diferenciando entre opiniones orgánicas y reseñas incentivadas. Este diseño evita la unificación artificial de contextos distintos, lo que incrementa la calidad del análisis posterior.

El particionamiento realizado sienta las bases para aplicar estrategias de muestreo estratificado que permitan conservar la heterogeneidad de la población sin perder representatividad, atendiendo tanto a las clases dominantes como a las minoritarias de manera controlada.

Técnica de muestreo

Para la recuperación de instancias a partir de cada partición generada, se empleó una técnica de muestreo estratificado, aplicada sobre la variable de salida sentiment. Esta técnica consiste en dividir previamente los datos en estratos homogéneos según las clases de dicha variable (por ejemplo, positivo, neutro y negativo), y extraer de cada estrato una fracción determinada de registros de manera proporcional a su tamaño original dentro de la partición.

En este caso, se configuró una fracción de muestreo del 20% por clase de sentimiento. Esto permite conservar la distribución de clases presente en cada partición, garantizando que el conjunto resultante sea representativo desde el punto de vista estadístico y útil para el posterior entrenamiento o evaluación de modelos de aprendizaje supervisado. Además, para evitar distorsiones derivadas de tamaños de muestra extremadamente reducidos, se estableció un umbral mínimo (min_rows = 50) que descarta automáticamente aquellas particiones con un volumen insuficiente de datos. Este control asegura que cada subconjunto utilizado en análisis o modelado cuente con una base mínima de observaciones que permita obtener estimaciones robustas.

El riesgo de sesgo es mínimo gracias al uso del muestreo estratificado por clase de sentimiento dentro de cada partición. Esta técnica es particularmente efectiva para evitar sesgos derivados de clases desbalanceadas, un problema común en datasets de opinión o revisión. Al preservar la proporción original de cada clase dentro de cada subconjunto muestreado, se mitiga el sesgo hacia clases mayoritarias y se asegura que el modelo tenga exposición equitativa a la variabilidad de los datos. Aunando a lo anterior, al aplicar el muestreo de forma separada en cada partición (y no sobre el conjunto total), se mantiene la variabilidad contextual asociada a combinaciones particulares de star_rating, verified_purchase y vine, lo cual añade riqueza informativa a las muestras y evita unificación artificial de contextos distintos.

El único riesgo potencial de sesgo proviene de las particiones descartadas por baja frecuencia. Sin embargo, esta exclusión es necesaria para asegurar estabilidad estadística en las muestras y se documenta de manera trazable, de forma que el analista tenga control sobre qué subconjuntos fueron omitidos.

La incertidumbre estadística asociada al proceso de muestreo es baja en las particiones con alto volumen de datos, ya que el tamaño de muestra obtenido (20% por clase) es suficientemente grande para representar con precisión la distribución subyacente de sentimientos. Para estas particiones, se espera que las métricas estimadas (por ejemplo, proporciones, medias, varianzas) tengan márgenes de error reducidos.

En cambio, para particiones con un volumen más moderado (por ejemplo, entre 1,000 y 10,000 registros), la incertidumbre relativa aumenta, aunque sigue siendo controlada por el muestreo estratificado. El umbral mínimo de 50 observaciones se definió precisamente para evitar escenarios de alta varianza o resultados inestables debido a tamaños de muestra insuficientes.

En conclusión, la estrategia implementada logra un equilibrio entre representatividad, control de sesgos y minimización de incertidumbre, adaptándose dinámicamente al volumen de datos en cada partición.