UNIVERSIDAD DE
COSTA RICA

ESTADÍSTICA ACTUARIAL I

CA0303

PROYECTO DE INVESTIGACIÓN
BITÁCORA 1

Estudiantes:
Holmar Adrián Rivera Castellón - B86564
María Paula Jiménez Torres - C04095
Andy Roberto Peralta Duarte - C25827
Dominick Rodríguez Trejos - B76600

Profesor:
Dr. Maikol Solís

3 de abril de 2025

1. Pregunta de investigación

1.1. Definición de la idea

Análisis de la relación entre las características de publicaciones con su popularidad y calidad.

1.2. Conceptualización de la idea

Para conceptualizar la idea de investigación, se procederá a definir cada concepto considerando la definición brindada por la Real Academia Española:

- Análisis:
  1. Distinción y separación de las partes de algo para conocer su composición.
  2. Estudio detallado de algo, especialmente de una obra o de un escrito.
- Relación:
  1. Conexión, correspondencia de algo con otra cosa
  2. Resultado de comparar dos cantidades expresadas en números.
- Característica:
  1. Perteneciente o relativo al carácter.
  2. Dicho de una cualidad: Que da carácter o sirve para distinguir a alguien o algo de sus semejantes.
- Publicación: Escrito impreso, como un libro, una revista, un periódico, etc., que ha sido publicado.
- Popularidad: Aceptación y aplauso que alguien tiene en el pueblo.
- Calidad: Propiedad o conjunto de propiedades inherentes a algo, que permiten juzgar su valor.

1.3. Identificación de tensiones

La medición o interpretación de la popularidad y calidad de una obra literaria es subjetiva, en especial la segunda, pues cada individuo tiene diferentes estándares y expectativas al momento de leer un nuevo libro, y no es enteramente claro cuales rasgos son los que cargan más peso en la calidad de un texto, estos pueden variar dependiendo del tipo de libro, como ejemplo tome una novela, su extensión e historia a contar serían de los elementos claves para evaluar su calidad, en comparación con un texto instructivo, donde la historia es prácticamente inexistente y hay un alto énfasis en formato, como su fuente, espaciado y uso de visualizaciones.

Además, se puede intuir que la región de origen y el contexto social del lector influyen en su opinión acerca de la calidad de una publicación, ya sea por que no logra apreciar el subtexto en

la obra debido a diferencias culturales entre lector y autor, o por las preferencias del lector que han sido formadas por su contexto cultural.

En el caso de la popularidad, aunque pareciera simple cuantificarla, este proceso puede complicarse, es posible determinar la cantidad de lectores que interactúan con una obra, ya sea considerando la cantidad de copias vendidas o, en el caso del presente proyecto, la cantidad de reseñas que ha recibido, pero, estas medidas usualmente no se traducen directamente a popularidad, es correcto decir que la cantidad de ventas o reseñas de un libro equivale al alcance de este, pero no necesariamente se puede concluir que es popular, pues este concepto involucra también la aceptación e interacción con la obra.

Finalmente, se debe recordar que la base de datos a analizar es limitada a la muestra dada, la información a mano no incluye a todos los lectores que interactúan con estas publicaciones, solamente la interacción desde el servicio de Amazon Goodreads, además, no se incluye información de los usuarios, por lo que es posible que la muestra tenga una preferencia por un tipo de publicación, autor o editorial especifica a raíz del contexto social y regional de los lectores.

1.4. Reformulación de la idea en modo pregunta

- ¿De qué manera influyen las características de una publicación, como autor, género y número de páginas, en su popularidad y calidad?
- ¿Cómo cuantificar el nivel de aceptación y valor de una publicación basado en sus características?
- ¿Por qué es importante medir la calidad y popularidad de un escrito con respecto a sus características?
- ¿Cuánto inciden las características de obras previas de un autor en la popularidad y calidad de publicaciones sucesivas?

1.5. Argumentación de la pregunta

1. **Pregunta** ¿De qué manera influyen las características de una publicación, como autor, género y número de páginas, en su popularidad y calidad?
   Contraargumentos
   Lógica La popularidad y calidad son conceptos altamente subjetivos que cambian entre persona y persona.
   Ética Diferentes ediciones pueden presentar características físicas que cambien el número de páginas así como diferentes autores y géneros acaparan cierta población que podría no ser comparable.
   Emocional Siempre habrá aquellos que mantengan obstinadamante su opinión acerca de una publicación independientemente de la opinión popular.
   Argumentos
   Lógica El alcance y calidad percibida de un libro estarán siempre ligados a múltiples factores que deben ser tomados en cuenta para un análisis adecuado.
   Ética Analizando un amplio rango de publicaciones se obtendrán resultados que apelen a un público más amplio.

Emocional La información ayudará a lectores y escritores a tener acceso a una selección y público, respectivamente, más amplia.
Concluya Saber cómo influyen las características de un libro o publicación a su éxito requiere la cuidadosa selección de estos así como una muestra adecuada que refleje la realidad de las personas.

2. **Pregunta** ¿Cómo cuantificar el nivel de aceptación y valor de una publicación basado en sus características?
   Contraargumentos
   Lógica No se cuenta con el total de reseñas y no se sabe qué tan significativa es la muestra disponible.
   Ética Para los datos es imposible verificar cuántas valuaciones son "reales" (hechas por personas de verdad).
   Emocional La métrica podría ser de bajo interés para aquellos con disposiciones a obras en específico.
   Argumentos
   Lógica Bajo un método adecuado se puede contar con una herramienta para calcular el impacto cultural y científico de una obra.
   Ética Sobresaldrían aquellos libros con un valor social más alto.
   Emocional Una métrica común facilitará la transición de lectores entre géneros.
   Concluya Poder cuantificar el valor y popularidad de un libro permitirá identificar aquellos de alta importancia para promoverlos entre la sociedad.

3. **Pregunta** ¿Por qué es importante medir la calidad y popularidad de un escrito con respecto a sus características?
   Contraargumentos
   Lógica El éxito de un libro incluye más que solo sus propias características como lo son los gustos y preferencias de las personas y las costumbres del grupo al cual está dirigido.
   Ética Actualmente se puede obtener información de otros medios además del impreso.
   Emocional Métodos de puntuación tan sencillos como un número del 0 al 5 no explican la relación única que puede tener cada lector con el libro.
   Argumentos
   Lógica Identificar las obras mayor impacto social es importante para una valoración más objetiva de su alcance e influencia.
   Ética Las reseñas deben comprender personas de diferentes edades y grupos sociales.
   Emocional Conocer de manera previa el valor de un libro puede ayudarle al lector a conocer si debería o no invertir su tiempo en él.
   Concluya Evaluar de manera más justa un libro se vuelve difícil debido a la subjetividad de los conceptos de popularidad y calidad. Para calcular la importancia social es necesario un punto de referencia común.

4. **Pregunta** ¿Cuánto inciden las características de obras previas de un autor en la popularidad y calidad de publicaciones sucesivas?

Contraargumentos
Lógica Un autor preferiría no verse confinado dentro de un solo género o estilo aún si fue popular en este.
Ética Este no es siempre el caso, no es posible saber de manera directa los sentimientos del creador.
Emocional Los lectores también tienden a quedarse con los autores que les parecen buenos independientemente del género que decidan escribir y viceversa.
Argumentos
Lógica Mantener cierta consistencia permitiría a un escritor mantener sus lectores.
Ética Se deben analizar tantos trabajos de un autor como sea posible.
Emocional Si se lee una serie de libros las personas estarán más dispuestas a seguir leyendo al mismo autor.
Concluya Un artista se verá siempre en la situación de desafiarse a sí mismo en cuanto a sus obras por lo que el contenido y naturaleza de ellas podría variar en gran medida y así la opinión de las personas.

1.6. Argumentación a través de datos

- Fuente de información: Base de datos pública en Kaggle ”Goodreads-books-with-genres” por el usuario Middlelight.
- Contexto temporal y espacial de los datos: Años 2006-2022, internacional.
- Facilidad de obtener la información: La base de datos es altamente accesible ya que se puede descargar directamente de Kaggle, un sitio web gratuito.
- Población de estudio: Libros publicados por una variedad de editoriales.
- Muestra observada: Más de 10 000 publicaciones distintas.
- Unidad estadística o individuos: Libros con reseñas en el sitio web Goodreads.
- Descripción de las variables de la tabla: La tabla cuenta con las siguientes variables.

    1. **Title**: título del libro.
    2. **Author**: autor o autores del libro.
    3. **average_rating**: puntuación promedio en el sitio web, entre 0 y 5, del libro.
    4. **language_code**: lenguaje del libro.
    5. **num_pages**: número de páginas del libro.
    6. **ratings_count**: cantidad total de reseñas que ha recibido el libro.
    7. **text_reviews_count**: cantidad de reseñas de las totales que incluyen un texto crítico.
    8. **publication_date**: fecha de publicación del libro.
    9. **publisher**: editor del libro.

10. **genre**: género literario del libro.

La base de datos cuenta con más de 10000 observaciones que constan de libros con reseñas en el sitio web Goodreads, provee el nombre, autor, editorial, número de páginas, lenguaje en el cuál está escrito, fecha de publicación, género así como información de las reseñas: cantidad y valuación promedio. La cantidad de reseñas será la de más interés para el presente trabajo pues se usará para medir la popularidad y calidad, y cómo influyen los demás factores en estas.

2. Revisión bibliográfica

2.1. Construcción de fichas de literatura

Ficha de Literatura 1

Encabezado	**Contenido**
Título:	Reader and Author Gender and Genre in Goodreads
Autor(es):	Mike Thelwall
Año:	2019
Nombre del tema:	¿Cómo las preferencias de género del lector y el autor influyen en las evaluaciones de libros dentro de géneros específicos en Goodreads?
Cronológica:	2017
Metodológica:	Análisis de varianza.
Temática:	Análisis de datos.
Teórica:	Sociología de la literatura y dinámicas digitales en páginas de reseña.
Resumen en una oración:	El artículo analiza cómo el género y preferencias afectan evaluaciones literarias en Goodreads.
Argumento central:	Las preferencias de género del autor y del lector tienen un impacto significativo en las evaluaciones de libros dentro de géneros específicos, mostrando sesgos de género. Se explora cómo estas dinámicas pueden influir en el consumo literario en plataformas digitales.
Problemas con el argumento o el tema:	El género de autores y reseñadores se estimó utilizando listas de nombres comunes con asociación de género en inglés, lo que pudo ocasionar errores en la determinación de género.
Resumen en un párrafo:	Este artículo analiza cómo las diferencias de género influyen en las evaluaciones de libros dentro de géneros específicos en Goodreads. Utilizando un enfoque cuantitativo, se recopiló una muestra aleatoria de más de 500,000 páginas de libros para evaluar patrones de calificaciones y reseñas de género. La metodología incluyó la extracción de datos web, el análisis de nombres para determinar el género de autores y reseñadores, y el estudio de las frecuencias de palabras en reseñas para identificar temas recurrentes. Los resultados mostraron que los lectores tienden a preferir libros de autores de su mismo género. También se observó que las reseñas reflejan temas de género, como relaciones y romance.

Ficha de Literatura 2

Encabezado	**Contenido**
Título:	Understanding Book Popularity on Goodreads.
Autor(es):	Suman Kalyan Maity, Ayush Kumar, Ankan Mullick, Vishnu Choudhary, Animesh Mukherjee.
Año:	2018.
Nombre del tema:	¿Cuáles factores contribuyen a la popularidad de los libros en Goodreads?
Cronológica:	2018.
Metodológica:	Análisis de correlación y modelos de regresión.
Temática:	Análisis predictivo.
Teórica:	Estudio de las influencias en la popularidad literaria online.
Resumen en una oración:	Investigación de factores en la popularidad de libros en Goodreads.
Argumento central:	La predicción de la popularidad de los libros en Goodreads utilizando métricas basadas en el comportamiento de los usuarios y las características de los autores.
Problemas con el argumento o el tema:	Factores como las calificaciones y reseñas del usuario pueden estar sesgados según la popularidad inicial del libro, reduciendo la capacidad del modelo para predecir el éxito de libros menos conocidos.
Resumen en un párrafo:	El documento analiza la predicción de la popularidad de los libros en Goodreads mediante el uso de métricas relacionadas con el comportamiento de los usuarios y las características de los autores. Identifica factores clave como la diversidad de estanterías, calificaciones, reseñas, y el prestigio del autor. Utilizando técnicas de regresión el modelo predice la cantidad de votos con una correlación de 0.61 y un RMSE de 1.25, mostrando que los factores de interacción de los usuarios son los más determinantes. Además, se categorizaron los libros en géneros específicos para mejorar la precisión de las predicciones. La investigación muestra cómo los datos sociotécnicos pueden ayudar a comprender y anticipar tendencias de popularidad.

Ficha de Literatura 3

Encabezado	**Contenido**
Título:	Exploring Goodreads Reviews for Book Impact assessment.
Autor(es):	Wang Kai, Liu Xiaojuan, y Han Yutong.
Año:	2019.
Nombre del tema:	¿Cómo las reseñas en línea, específicamente de Goodreads, pueden ser útiles para evaluar el impacto de los libros académicos?
Cronológica:	2019
Metodológica:	Análisis de correlación, prueba chi-cuadrado, análisis de polaridad del sentimiento y frecuencia de términos inversa del documento.
Temática:	Análisis de datos para bibliometría.
Teórica:	Minería de datos textuales y valoración de impacto académico.
Resumen en una oración:	El estudio analiza reseñas de Goodreads como métricas complementarias del impacto de libros académicos.
Argumento central:	Explorar cómo las reseñas en línea de Goodreads pueden usarse como indicadores complementarios para evaluar el impacto de los libros académicos.
Problemas con el argumento o el tema:	Las reseñas pueden ser muy subjetivas y varían dependiendo del contexto de los usuarios, lo que dificulta para interpretarlas como datos confiables.
Resumen en un párrafo:	El estudio analiza cómo las reseñas de Goodreads pueden servir como indicadores de impacto de libros, considerando factores como las disciplinas, los roles de los reseñadores y las emociones expresadas. Utilizando análisis de correlación para evaluar la relación entre las calificaciones de Goodreads y las citas académicas, y pruebas chi-cuadrado para identificar diferencias significativas entre los comportamientos de los reseñadores. También utilizó técnicas de análisis de sentimientos con TextBlob para calcular la polaridad emocional de las reseñas (-1 a 1) y el algoritmo TF-IDF para identificar palabras clave. Los resultados mostraron que las reseñas en línea ofrecen una perspectiva complementaria a las métricas tradicionales como las citas académicas, especialmente en disciplinas como artes y humanidades, destacando su utilidad en evaluaciones de impacto interdisciplinarias.

3. Construcción de la UVE de Gowin

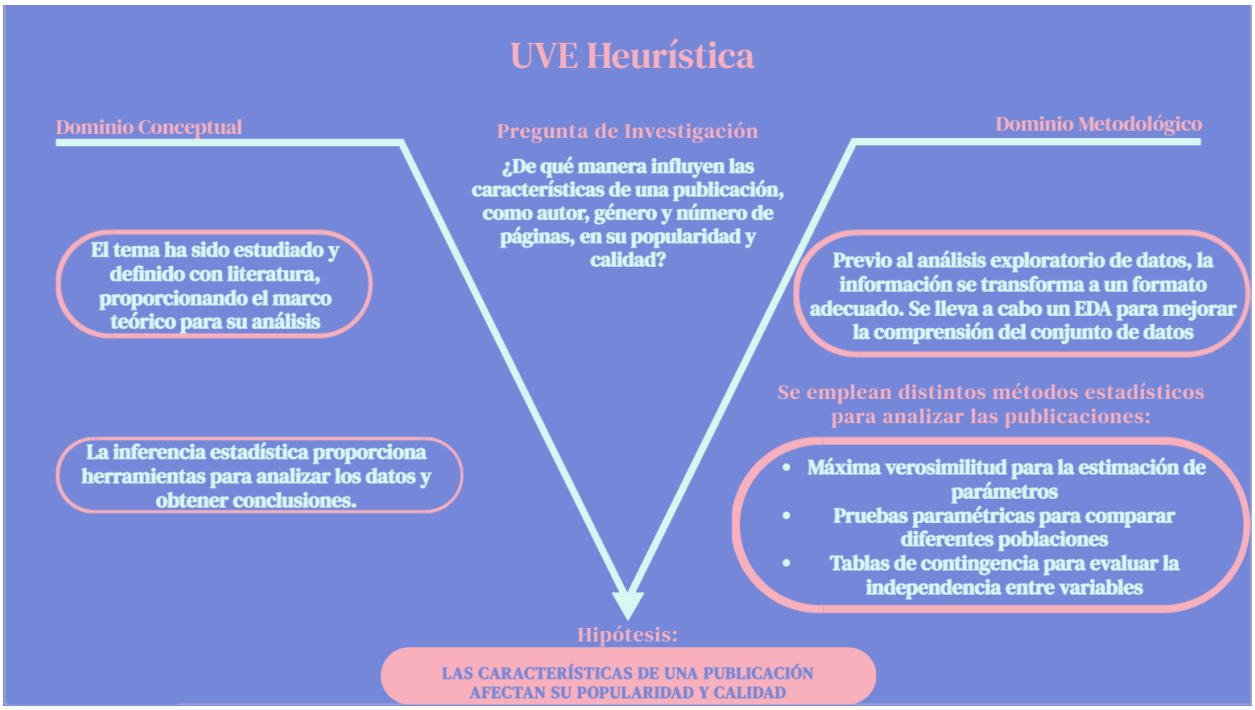

Figura 1: V Heurística

4. Parte de escritura

Con la llegada de la era digital, los hábitos de lectura han evolucionado. Las plataformas de reseñas en línea han adquirido un papel central en la evaluación de la popularidad y calidad de los libros, lo que ha generado un creciente interés en el análisis del comportamiento de los lectores en entornos digitales.

Desde la ciencia de datos, surge una nueva área de estudio dedicada a comprender estos patrones. ¿Son ciertos lectores más propensos a leer o calificar positivamente determinados tipos de textos? ¿Existen características que hacen que algunas obras sean más aclamadas y reconocidas? Son incógnitas como estas las que guían el presente trabajo.

Plataformas como Goodreads han permitido a los usuarios compartir opiniones, otorgar calificaciones y generar datos sobre el impacto de las obras literarias en distintos contextos. En este sentido, el análisis de métricas asociadas a la interacción de los usuarios, como el número de calificaciones, reseñas y la puntuación promedio, se ha convertido en un campo de investigación relevante para el estudio de tendencias literarias.

Diversos estudios han abordado la relación entre las reseñas y la evaluación de los libros. Por ejemplo, Wang, Liu y Han (2019) analizaron factores como las disciplinas, los roles de los reseñadores y las emociones expresadas en las críticas. Sus hallazgos demostraron que las reseñas en línea ofrecen una perspectiva complementaria a las métricas tradicionales, como las citas

académicas, especialmente en disciplinas como las artes y las humanidades, destacando su utilidad en evaluaciones interdisciplinarias.

Por otro lado, Thelwall (2019) examinó la influencia del género en la evaluación de libros dentro de Goodreads. Su estudio reveló que los lectores tienden a preferir obras de autores de su mismo género. Finalmente, Maity et al. (2018) exploraron los factores que contribuyen a la popularidad de los libros en Goodreads, identificando patrones de comportamiento de los usuarios y características de los autores mediante modelos de regresión y análisis de correlación. Estos estudios muestran cómo los enfoques técnicos pueden ayudar a comprender y predecir tendencias de popularidad.

En este contexto, el análisis de datos se ha convertido en una herramienta clave para comprender la calidad y la popularidad de los libros, permitiendo interpretar mejor la información generada en plataformas digitales. Este estudio se propone analizar las métricas cuantitativas asociadas a las calificaciones y reseñas, como su volumen y promedio, para comprender su relación con la percepción de popularidad y calidad de los libros mediante métodos de inferencia estadística. Con ello, se busca contribuir al conocimiento sobre la influencia de estos factores en la formación de opiniones y en la toma de decisiones de los lectores en entornos digitales.