Predicción de las preferencias de consumo de café: insights sobre el *Great American Coffee Taste Test Breakdown*

Entrega final

Autoras

Clara Bernhardt

María Teresa Laszeski

Rosario Luque

Profesores

María Noelia Castillo

Ignacio Spiousas

Asignatura

Ciencia de Datos

Semestre y año de presentación

2° semestre 2024

1. Introducción

El café es una de las bebidas más consumidas en el mundo, y en los Estados Unidos, su importancia cultural y económica lo convierte en un elemento central de la vida cotidiana. Más de dos tercios de los adultos estadounidenses consumen café a diario, lo que lo posiciona como un mercado altamente competitivo y en constante evolución. Sin embargo, detrás de esta popularidad se encuentran patrones complejos de consumo influenciados por factores como la edad, el género, los ingresos y las preferencias culturales. Estos factores demográficos no sólo determinan qué tipos de café son preferidos, sino también cómo, cuándo y dónde se consume.

Esta investigación se centra en analizar cómo las características individuales afectan las decisiones de compra de los consumidores en el contexto del *Great American Coffee Taste Test Breakdown*, un conjunto de datos que permite desentrañar las preferencias y comportamientos del consumidor estadounidense. Las preguntas clave que guían este estudio son: ¿Cómo varían las preferencias de café según las características sociodemográficas de los individuos?; ¿Cómo impactan los hábitos de consumo de café en las preferencias de distintas variedades del producto?; y, por último, ¿es posible predecir la preferencia por alguna variedad de café a partir de las variables sociodemográficas y de hábitos de consumo del producto*?*

La motivación detrás de este análisis radica en la creciente necesidad de las empresas por comprender mejor a sus consumidores para desarrollar productos y campañas más personalizadas. Asimismo, si se identifican preferencias particulares entre los jóvenes, como bebidas más azucaradas o frías, esto podría influir en las estrategias para captar este segmento específico.

La presente propuesta de investigación tiene como objetivo la predicción de la preferencia de los individuos por distintas variedades de café a partir de sus características

sociodemográficas y sus hábitos de consumo del producto. En primer lugar, se hará un relevamiento de la bibliografía pertinente al estudio de las preferencias de consumo de café. En segundo lugar, se describirá la base de datos que se utilizará en la propuesta, con el fin de entender de qué tipo de variables se disponen y de caracterizar el fenómeno que se quiere predecir. En tercer lugar, se propondrá un análisis descriptivo de las variables de interés, con el fin de conocer la distribución de los datos, como así también una análisis exploratorio a partir de PCA para comprender la variabilidad que potencialmente podría explicar cada variable. En cuarto lugar, se propondrán los modelos de Naïve Bayes y Random Forest para realizar la predicción, a la vez que distintas métricas para evaluar la *performance* de cada modelo. Finalmente, se concluirá la propuesta con los resultados que se esperarían obtener y las potenciales limitaciones que podría enfrentar el análisis propuesto.

2. Literatura previa

El análisis del consumo de café ha capturado la atención de múltiples disciplinas, desde el Marketing hasta la Economía del Comportamiento. Sin embargo, muchos estudios se centran en aspectos específicos sin abordar de manera integral las interacciones entre factores demográficos y preferencias de consumo. Este apartado tiene como objetivo recopilar los hallazgos relevantes de la literatura que sirven como base para el análisis posterior.

En primer lugar, el estudio de Acuña (2012) utilizó técnicas de análisis conjunto para evaluar el comportamiento y las preferencias de los consumidores, estimando cómo valoran diferentes características de productos y servicios. Este enfoque permite identificar patrones de consumo y factores determinantes, como la demografía, relevantes para analizar cómo variables como la edad, ingresos o género impactan en las elecciones de café. Por ejemplo, estas herramientas pueden ser adaptadas para segmentar consumidores según sus hábitos, tal como el tipo de café (negro, con leche) y su frecuencia de consumo.

En segundo lugar, se puede rescatar la investigación de Campos Trigoso et al. (2021) que analizaron la sostenibilidad en la producción de café utilizando marcos como SAFA (Sustainability Assessment of Food and Agriculture Systems) y evaluaciones de ciclo de vida. Si bien este trabajo se enfoca en la cadena de suministro, es relevante porque el interés por la sostenibilidad ha influido en la percepción y demanda de café, especialmente entre grupos demográficos específicos. Por ejemplo, consumidores jóvenes o de altos ingresos podrían priorizar productos éticos, afectando sus preferencias.

En tercer lugar, hay que tener en cuenta los impactos del café en la salud y el consumo habitual. Talero et al. (2019) realizaron una revisión sistemática sobre los efectos del café en la salud cardiovascular, identificando relaciones dosis-respuesta. Esto resalta el impacto del café en la rutina diaria y su relevancia cultural. Este aspecto es clave para contextualizar preguntas de investigación, como la relación entre patrones de consumo y edad, género o lugar de consumo (hogar versus cafeterías).

Por último, en la investigación de Bastos Osorio et al. (2019), analizaron tendencias globales y prácticas disruptivas en la cultura del café. Explican cómo la cultura del café ha evolucionado globalmente, destacando el surgimiento de innovaciones que agregan valor tanto a productores como a consumidores. Estas transformaciones han redefinido las expectativas y hábitos de los consumidores en diferentes mercados, justificando explorar cómo factores como ingresos y geografía impactan las preferencias en el mercado estadounidense.

La presente investigación se apoya en estas perspectivas para explorar cómo variables demográficas moldean las decisiones de consumo de café en los EE.UU., un mercado diverso y competitivo, al analizar un conjunto de datos reciente y completo como el del *Great American Coffee Taste Test Breakdown*. Además se busca explorar el potencial de modelos predictivos para predecir la preferencia de los individuos por distintas variedades de café a

partir de sus características sociodemográficas y sus hábitos de consumo de café, llenando así un vacío en la literatura existente. Se espera generar información útil para personalizar productos y diseñar estrategias de marketing efectivas, contribuyendo al avance en la comprensión del comportamiento del consumidor.

3. Base de datos: "*The Great American Coffee Taste Test"*

La base de datos *The Great American Coffee Taste Test* (Hoffman, 2023) ofrece una visión profunda y multidimensional de las preferencias y hábitos de consumo de café en los Estados Unidos. Este conjunto de datos se compone de 53 variables que tienen potencial para comprender el consumo de café en la cultura estadounidense.

La estructura de los datos, que comprende 49 variables categóricas y 5 strings, permite un análisis variado que va más allá de las simples estadísticas de consumo. Esta combinación de tipos de datos facilita la exploración de relaciones complejas entre diferentes aspectos del consumo de café, desde factores demográficos hasta preferencias sensoriales específicas.

Las variables demográficas incluidas, como edad, género y nivel educativo, permiten segmentar y analizar las tendencias de consumo a través de diferentes grupos poblacionales. Esto es particularmente valioso para comprender cómo las preferencias de café varían entre generaciones o niveles socioeconómicos. Adicionalmente, la base contiene datos acerca de los ingresos del hogar en categorías que permiten analizar cómo el poder adquisitivo influye en los hábitos de consumo de café.

Los hábitos de consumo detallados en la base de datos ofrecen una visión única de cómo el café se integra en la vida cotidiana de los estadounidenses. La información sobre frecuencia de consumo, lugares preferidos para beber café y métodos de preparación favoritos proporciona insights relevantes para la industria cafetera, desde productores hasta minoristas.

Un aspecto interesante de esta base de datos es la inclusión de preferencias de sabor detalladas. Las variables que cubren la intensidad del café, el nivel de tostado preferido y las adiciones comunes (como leche o azúcar) permiten un análisis profundo de las preferencias gustativas de los consumidores estadounidenses. Esta información es crucial para los tostadores y fabricantes de café que buscan adaptar sus productos a las preferencias del mercado.

La dimensión económica del consumo de café también está bien representada en los datos. Las variables relacionadas con el gasto mensual en café y la disposición a pagar por diferentes calidades o experiencias de café ofrecen una perspectiva sobre el valor percibido del café en la sociedad estadounidense y las oportunidades de mercado para estos productos.

La base de datos incluye evaluaciones comparativas de cuatro muestras de café específicas: A (Light roast, washed process), B (Medium roast), C (Dark roast) y D (Light roast, natural process). Estas evaluaciones, que incluyen calificaciones de amargor, acidez y preferencia personal, proporcionan datos concretos sobre las respuestas sensoriales de los consumidores a diferentes perfiles de café. Aunque las evaluaciones de amargor, acidez y preferencia personal para las muestras de café, así como la autoevaluación de experiencia en café, se presentan como variables categóricas ordinales, pueden considerarse fundamentalmente numéricas. Estas variables utilizan escalas discretas (del 1 al 5 y del 1 al 10 respectivamente) que representan un continuo subyacente de intensidad o nivel.

Además, la base de datos incorpora variables únicas que reflejan tendencias contemporáneas, como el trabajo desde casa, lo que permite analizar cómo los cambios en los patrones de trabajo afectan el consumo de café.

4. Metodología

Previo a cualquier tipo de análisis, el primer paso será la limpieza y la preparación de los datos. Para ello, se hará un análisis de los *missing values*, lo cual determinará qué columnas

con muchos valores faltantes sean descartadas desde un principio. Además, de las columnas preservadas serán eliminados los *missing values* restantes. Luego, se corroborará que las variables numéricas no tomen valores negativos y se utilizará el método de la Desviación Absoluta de la Mediana (MAD) para identificar outliers debido a su robustez (Rousseeuw, 1990), fijando un umbral de 3.5 y etiquetando las filas sin outliers con un 0 y las que tenían outliers con un 1.

En segundo lugar, se realizará un análisis descriptivo y exploratorio de las variables de interés, con el fin de responder a las preguntas sobre la caracterización de los patrones de consumo por variables de género, edad e ingreso. Para ello, se creará una nueva variable denominada "Preferencia" que funcionará como indicador de la elección de la muestra preferida de las cuatro disponibles (A, B, C y D). Esta variable categórica surge de la necesidad de resumir la información de las columnas "prefer_abc", que solo denota la preferencia por las variedades A, B o C, y "prefer_ad", que solo señala si se prefiere A o D, en una sola, para de esta forma resumir la información sobre la preferencia de las muestras de café en una sola variable. Luego, se realizarán histogramas entre "Preferencia" y las variables seleccionadas a priori como predictoras del consumo, con el fin de observar la dispersión de los datos. Además, se acompañarán por medidas estadísticas descriptivas relevantes, como la media, mediana y la varianza de los datos. Luego, por medio de un Análisis de Componentes Principales se hará una primera aproximación al estudio de la variabilidad explicada por cada una de las variables preservadas luego de la limpieza. Esto permitiría observar, por ejemplo, una alta multicolinealidad entre las variables que luego podría afectar el análisis predictivo. Además, para evaluar este último factor se realizará una matriz de correlaciones entre las variables numéricas del data set.

Para responder a la pregunta de investigación referida a la predicción de las preferencias de café se aplicará Naïve Bayes, un modelo de clasificación que se basa en la

regla de Bayes, y Random Forest, un método de *ensamble* que hace uso de árboles de decisión, para predecir la variable "Preferencia", la cual funcionará como *proxy* de lo que se quiere medir. La selección de estos modelos fue realizada luego de la revisión del estudio de Ossani et al. (2021), quienes evalúaron la precisión de diversos modelos de clasificación supervisada para la tarea de clasificar 4 variedades de café de especialidad a partir de sus características, lo cual es similar a la tarea que se busca realizar en el presente estudio. Sin embargo, aquí serán incluídas no solo variables referidas a las propiedades sensoriales de las variedades de cafés sino también variables sociodemográficas de los participantes del estudio y de preferencias de consumo del producto en cuestión.

Por un lado, el algoritmo de Naive Bayes evalúa cuánto aporta cada predictor en la clasificación de la variable dependiente y, con base en los datos de entrenamiento, en el presente estudio elegiría la función de densidad de cada predictor *j*-ésimo para cada una de las 4 categorías a ser predecidas (A, B, C y D). Si bien relaja el supuesto de la normalidad de los predictores considera la no dependencia entre los atributos. Por el otro, Random Forest es un método de *ensamble* que crea B árboles con m predictores *bootstrappeados*, lo cual hace que $m<p$ predictores originales. En este trabajo, la cantidad de *m* predictores seleccionados por *leave one out cross validation* y se crearán B = 100 árboles.

Finalmente, para evaluar y comparar la *performance* de ambos modelos se reportarán la matriz de confusión, la curva ROC, los valores de AUC y de Accuracy para cada modelo.

5. Conclusión

A partir de la implementación de los métodos predictivos Naïve Bayes y Random Forest se esperaría identificar patrones clave que expliquen cómo las características sociodemográficas y los hábitos de consumo influyen en la preferencia por diferentes variedades de café. Naïve Bayes proporcionará una visión probabilística de la relación entre los predictores y las categorías de preferencia, mientras que Random Forest permitirá capturar relaciones no

lineales y complejas entre las variables. Además, se esperaría que Random Forest performance mejor que Naïve Bayes debido a que el supuesto de independencia entre los predictores sería difícil de garantizar.

Asimismo, el proyecto enfrenta varias limitaciones. En primer lugar, la representatividad de la muestra es cuestionable, dado que quienes contestaron la encuesta podrían ser considerados como una submuestra especializada en el consumo de café, lo que podría restringir la generalización de los resultados. En segundo lugar, ninguno de los métodos de selección vistos en el curso se consideraron pertinentes para seleccionar variables categóricas para Naïve Bayes, lo que podría afectar la precisión y robustez del modelo debido a la inclusión de variables irrelevantes o altamente correlacionadas.

Además, las variables de la encuesta podrían generar complicaciones. Por un lado, el desequilibrio en las clases de la variable objetivo (si existiera) podría influir negativamente en el rendimiento de los modelos, particularmente en métricas como el AUC y la Accuracy. Finalmente, la dependencia de datos categóricos y ordinales en la base puede limitar la capacidad de los modelos para capturar ciertas sutilezas en las preferencias de los consumidores.

En conclusión, aunque el análisis propuesto tiene un sólido marco metodológico y potencial para generar valor, estas limitaciones deben ser abordadas en estudios futuros para garantizar resultados más robustos y generalizables.

Referencias

Acuña, O.A. (2012). Método de valoración de preferencias "Análisis conjunto" una revisión de literatura. Research Papers in Economics.

Bastos Osorio, L.M., Salazar Escalante, R.Y., Mora Carvajal, C., y Duarte Cristancho, M. (2019). Análisis de las tendencias en la producción y el consumo de café a nivel internacional. Visión Internacional (Cúcuta).

Campos Trigoso, J.A., Murga Valderrama, N.L., Rituay Trujillo, P.A., y García Rosero, L.M. (2021). Sostenibilidad del café: revisión sistemática de la literatura. Revista Venezolana de Gerencia.

Hoffmann, J. (2023). The Great American Coffee Taste Test. https://www.kaggle.com/datasets/jackogozaly/the-great-american-coffee-taste-test

Ossani, P. C., Rossoni, D. F., Cirillo, M. Â., y Borém, F. M. (2021). Classification of specialty coffees using machine learning techniques. Research, Society and Development, 10(5), e13110514732-e13110514732. DOI: 10.33448/rsd-v10i5.14732.

Rousseeuw, P. J. (1990). Robust estimation and identifying outliers. Handbook of statistical methods for engineers and scientists, 16, 16-11.

Talero, L.H., Peñaloza, M., Gutiérrez, V., y Castillo, J.S. (2019). “Efecto del consumo habitual de café en la salud cardiovascular de la población adulta: protocolo de una revisión de revisiones sistemáticas de la literatura”. Universitas Médica.