Índice

Resultado del análisis con k-vecinos más cercanos (informe en español)

Es importante recalcar que las observaciones que se harán en este documento se referencian siempre en **primera instancia** a los datos obtenidos de la página de Kaggle *"Mall Customer Segmentation Data"* del usuario Yanyan Wu. Dicho esto, cualquier fuente externa que ayude al entendimiento de los datos será referenciada en el **último encabezado** para que el lector pueda corroborar su veracidad.

El significado de las columnas en la tabla es:

- **"CustomerID"**: Número del cliente en nuestra base de datos.
- **"Gender"**: Sexo biológico del cliente.
- **"Age"**: Edad del cliente.
- **"Annual Income (k$)"**: Salario anual del cliente en base a mil dólares (1k$ = 1000$).
- **"Spending Score (1-100)"**: Es una métrica que ayuda a las empresas a segmentar a los clientes según sus patrones de gasto y comportamiento de compra donde los valores cercanos a 1 indica un comportamiento de gasto muy bajo, y 100 indica un comportamiento de gasto muy alto.

Tratamiento de los datos antes de su procesamiento

Antes de procesar los datos, debemos **eliminar o ajustar las columnas** que nos pueden causar problemas a la hora de entender la **tendencia del grupo** que compra en este mercado, una columna que **no tiene utilidad** para nuestro análisis es *"CustomerID"*[1], aunque es de utilidad para saber cuántos clientes se están tomando en cuenta en la base de datos (que son 200 clientes), dentro de un análisis más profundo, podemos no tomar en cuenta la columna, podemos ver esto como al llegar el primero a un avión: debes ocupar tu puesto correspondiente y llegar más temprano no influirá la hora en la que despegue de la pista (asegúrate de llevar tu teléfono o tablet para entretenerte mientras el resto sube).

Por otro lado, la columna *"Gender"* especifica el **sexo del cliente** masculino ("Male") y femenino ("Female"), en la página que he creado se muestra con estas palabras su clasificación, pero esta representación de la columna no es útil en el análisis, pues obtendremos un error al procesar este tipo de dato y no números como el resto[2]. Por esto usamos una técnica de la ciencia de datos que se conoce como **codificación** "One Hot", que transforma los datos en caracteres numéricos más simples de entender por el método que **aplicamos con nuestra inteligencia artificial**, la pregunta que nos hacemos es: ¿Esta fila de nuestra tabla de datos es de tipo "Female"?, si es verdadero, sustituiremos la palabra por el número 1, de lo contrario es cero 0.

[1] La columna se elimina antes de que se presente en el análisis, pero está presente en la página de Kaggle.
[2] Se origina el error porque se procesa un tipo de dato string y no int o float, lo que genera errores en su interpretación.

Como queremos comparar variables que están en **diferentes escalas** (como años y miles de dólares o $k\$$ de ingreso anual), usaremos un **método matemático de estandarización** que se conoce como "normalización Z-score", su objetivo principal es centrar los datos en torno a la media y escalarlos según su desviación estándar, de modo que todos los valores tengan una media de 0 y una desviación estándar de 1, se podría decir de manera muy simplificada que la estandarización toma como referencia la media y ajusta los datos para que presenten una menor tasa de variabilidad relativa en torno a la media, lo que hace más fácil para el algoritmo no solo procesar los datos, también visualizar más fácilmente los valores atípicos. Su fórmula matemática es:

$$Z = \frac{x - \mu}{\sigma}$$

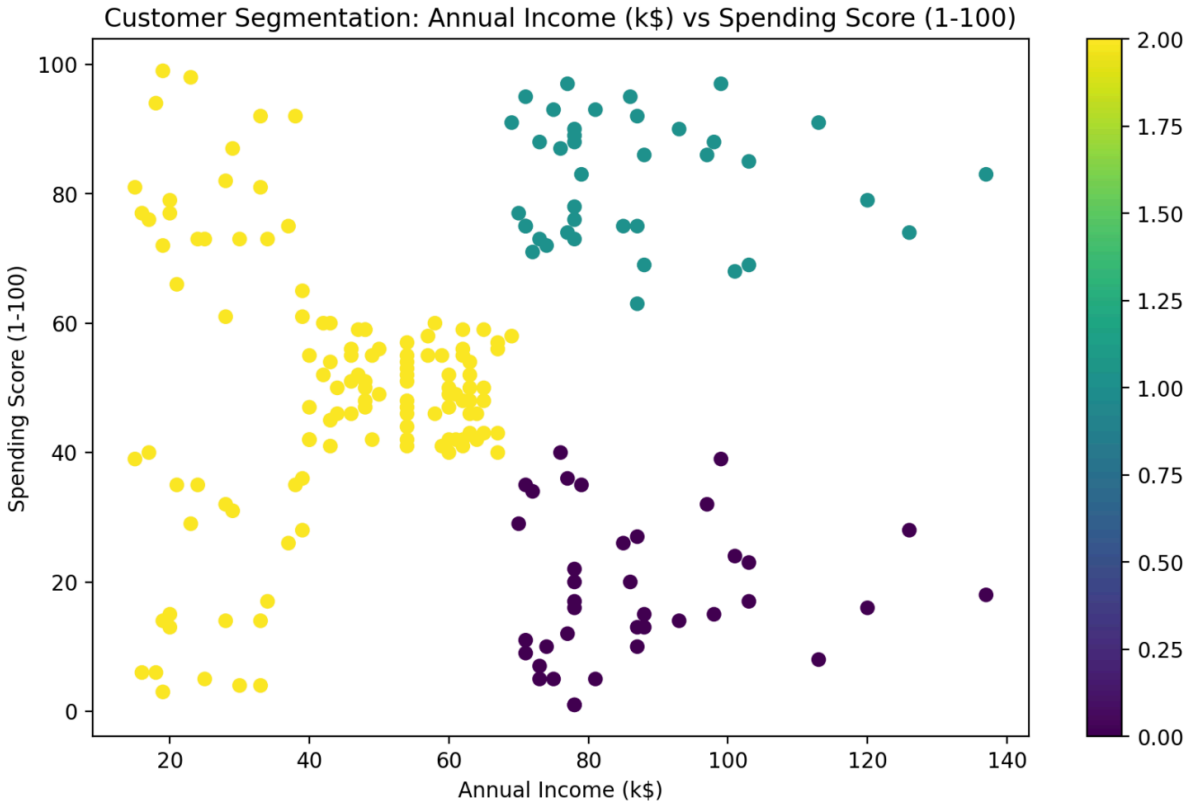
Donde la variable (x) es el valor presente en la tabla a estandarizar, (μ) es la media de la variable o característica, (σ) es la desviación estándar de la variable o característica, por último (Z) es la puntuación Z-score que nos dice cuántas desviaciones estándar está un dato de nuestro conjunto (tener un dato con $Z = 2$ equivale a decir que se desvía de la media en 2 veces la desviación estándar[3]).

Con este preprocesamiento de los datos, nos aseguramos de cumplir con los requisitos de un análisis más sencillo de interpretar y procesar.

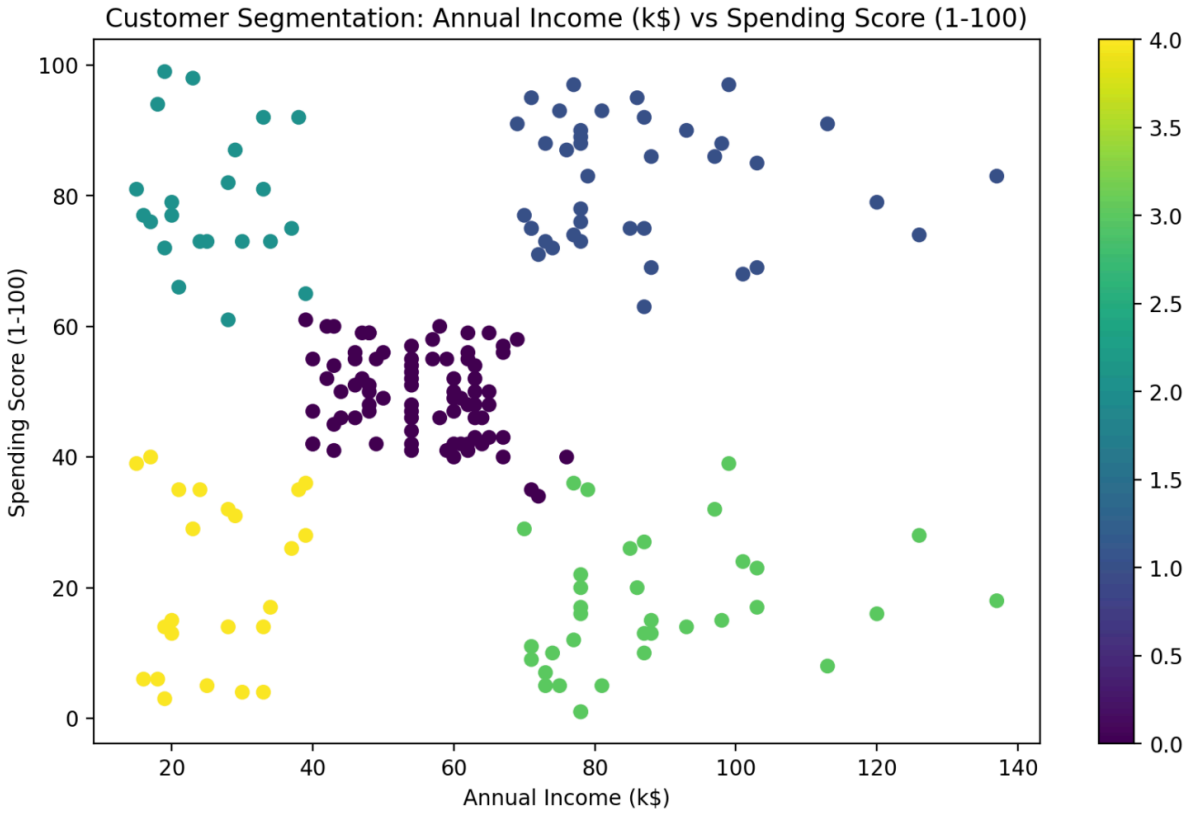
Agrupamiento de k-vecinos

Elegí esta técnica de **aprendizaje automático no supervisado de la ciencia de datos** porque nos **ayuda a visualizar** como agrupar conjuntos de datos por sus relaciones comunes y menos visibles para un humano, por ende, nos permite ver mejor los datos que **compartan características** con otros **puntos vecinos** (k-vecinos más cercanos o K-means). Al aplicar este método con el número de grupos a 3 ("Number of Clusters"), que es la configuración predeterminada, y solo nos fijamos en el ingreso anual ("Annual Income") y puntos de gasto ("Spending Score"), nos encontramos con este gráfico:

[3] Recomiendo investigar estos términos para no alargar el informe.



Como podemos ver en el gráfico, se representan tres grupos que se agrupan en base a colores para señalar el comportamiento común entre estos, no obstante, el grupo amarillo es muy generalizado, por lo que aumentaremos el número de grupos a 5 (“Number of Clusters”), con esto podremos obtener grupos más categorizados:



Excelente, ahora está mucho más claro las categorías que poseemos en este análisis en base a estas variables, una vista sobre la gráfica muestra como poseemos un grupo promedio en morado que parece concentrar la mayoría de nuestros clientes, que engloba un ingreso anual entre 40 y casi 80 mil dólares anuales con puntos de gasto entre 40 y 60, estos son clientes muy importantes, pues no solo compran en un promedio estable, también, de media, poseen ingresos que permiten que esta actividad sea sostenible en el tiempo (se recomienda tener en cuenta el salario mínimo antes de sacar conclusiones), en este documento se tratará el salario mínimo por ley a 7.25$ por hora o 63510$ que representaremos de forma aproximada como 63k$ anuales[4], aunque en otros estados y en tierra federal puede ser diferente, muchos estados tienen este salario mínimo.

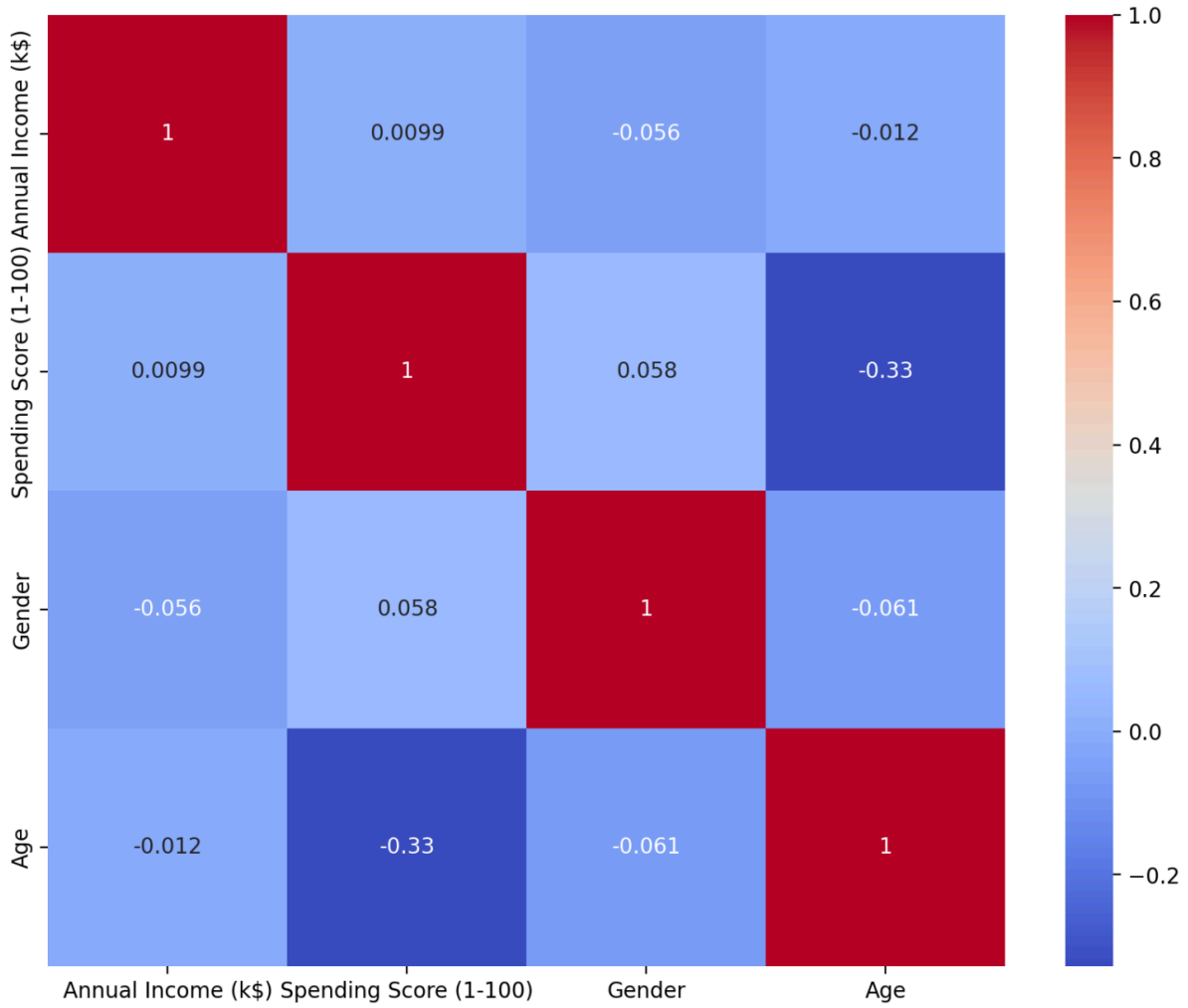
Analizando el resto de grupos, el grupo amarillo (izquierda abajo) tiene una tendencia sostenible que se ajusta en términos de cuanto compra y gana anualmente, a diferencia del grupo turquesa (izquierda arriba) que compran demasiado sin tener en cuenta su salario anual, lo que deriva en una tendencia inestable. Por otro lado, se ha logrado atraer la atención de compradores que ganan por encima de 70k$, aunque tenemos que hacer una investigación más profunda en el grupo verde (derecha abajo) para observar los productos eligen y su precio correspondiente en una tendencia semanal o mensual (lo que nos permitirá ganar más dinero ofreciéndole los mejores productos en el momento que más lo necesiten), esto es fácil de entender si pensamos en un caso más evidente, si tenemos un producto que se compra en un mes del año, preparamos no solo un mayor volumen de estos productos, también pensamos en crear nuevas variaciones del producto para ver si

[4] Las referencias del salario mínimo en Estados Unidos (se asume, ya que no se referencia en el documento original de Kaggle), se encuentran al final del documento y son 2, 3, y 4.

podemos tener un nuevo producto que compita con el estándar (preferiblemente mejor para los clientes y por lo tanto, que lo compren más), en caso donde haya menor demanda, se bajará la oferta del producto y sus variaciones.

Correlación entre variables y su matriz

Es importante señalar que saber la correlación o la influencia de una variable sobre otra es crucial para la deducción del comportamiento de la población que estamos estudiando, para esto, usaremos lo que denominamos una matriz de correlación, que es una herramienta matemática donde cada valor en la matriz representa el coeficiente de correlación entre dos variables específicas, este valor es la relación lineal entre variables. Si tomamos la matriz de correlación para nuestras variables, obtendremos el siguiente gráfico:



Como podemos observar, se divide en filas (horizontal) y columnas (vertical), si vemos una fila marcada por una de las variables, podemos encontrar la relación con otra variable vista en una columna, por ejemplo: si vemos la relación entre las mismas variables (mismo número de fila y columna), veremos que la relación es 1, tiene sentido pues tiene sentido consigo misma.

Por otro lado, si se aproxima a cero o está lejos de 1 y -1, quiere decir que no posee ninguna relación con la variable, si nos fijamos en la matriz, casi todas son cero, por lo que en el grupo que estamos estudiando, no influye demasiado entre ellas.

Aquí está presente una correlación negativa entre la edad (“Age”) y los puntos de gastos (“Spending Score (1-100)”), la correlación lineal es muy débil pero notable, lo que nos marca que las personas mayores suelen gastar menos al comprar[5].

Conclusión

En conclusión, el análisis presentado proporciona una visión detallada del comportamiento de los clientes en función de su edad, género, ingreso anual y patrones de gasto. Al preprocesar los datos de manera adecuada y aplicar técnicas de agrupamiento y análisis de correlación, se ha logrado identificar grupos específicos de clientes y sus tendencias de compra. Estos hallazgos no solo facilitan la segmentación de los clientes para fines comerciales, sino que también permiten una mejor comprensión de las dinámicas de consumo dentro del mercado.

Con base en estos resultados, las empresas pueden diseñar estrategias de marketing más efectivas, optimizar la oferta de productos y mejorar la satisfacción del cliente, alineando mejor sus servicios con las necesidades y comportamientos observados en cada grupo de clientes. Por tanto, este análisis es una herramienta valiosa para la toma de decisiones informadas y estratégicas en el ámbito comercial.

Referencias utilizadas

Este documento usa los recursos de:

1. “Mall Customer Segmentation Data” | Yanyan Wu | Kaggle.
2. “What are the annual earnings for a full-time minimum wage worker?” | Center for Poverty and Inequality Research.
3. “State Minimum Wage Laws” | U.S. Department of Labor | State Minimum Wage Laws.
4. “Minimum Wage” | U.S. Department of Labor | Minimum Wage.

[5] Recordemos: una correlación débil.