Trabajo Parcial

Curso: Big Data

Profesor: Richard Fernando Fernández Vásquez

Alumnos:

Luis Gustavo Becerra Bisso, U20195451

Henry Josue Diaz Huarcaya, U20201C579

Sección: CCA3

Mayo de 2024

Descripción del caso de uso:

Muchos grandes centros comerciales y supermercados obtienen información a través de membresías de los clientes y usan el Algoritmos de clustering de K-means para poder otorgar mejores beneficios para clientes objetivos según variables necesarias como la edad y su salario. El objetivo de la agrupación de K-means es dividir las observaciones en K clústeres y cada una de las observaciones pertenece al clúster que posee la media más cercana.(Run-Qing et al., 2018).
Como gerente quieres crear nuevos beneficios para los clientes del centro comercial y le pides al equipo de marketing que segmento sería el mejor para esto basándose en su edad, su salario anual y el gasto en el centro comercial para lanzar promociones de .

Descripción del conjunto de datos obtenidos de kaggle:

Los datos fueron recolectados de Kaggle y se utilizarán para segmentar los datos y mostrar los beneficiarios de los nuevos beneficios que desean tener.

Los datos del dataset son los siguientes:

1. **CostumerID:** Es una variable única que es asignada a cada cliente del centro comercial.
2. **Género:** Esta variable representa el género del cliente. Es un valor booleano que indica si es femenino o masculino.
3. **Edad:** Esta variable representa la edad de los clientes en el dataset. Es un número entero que indica la edad en años.
4. **Salario**: Representa el salario o ingreso anual de los clientes. Es un valor numérico que indique el salario en dólares.
5. **Puntuación de gasto:** Esta variable podría ser una métrica que mide cuánto gasta un cliente en el centro comercial en comparación con otros clientes. Es un valor numérico que indica la cantidad de dinero gastado durante un período de tiempo específico.

Análisis Exploratorio de los datos (EDA)

- **Inspección de Datos**

  Se observó que los datos contiene 5 columnas, teniendo 4 de ellas con valores enteros y una con valores booleanos, debido a que no se discrimina el uso del género para los beneficios, esta columna no es utilizada para este caso. Además, este dataset no muestra datos faltantes por lo que es más fácil para la etapa de preprocesamiento. Estos datos han sido normalizados utilizando H2OKMeansEstimator antes de la creación del modelo K-means.

- **Preprocesamiento de los Datos**

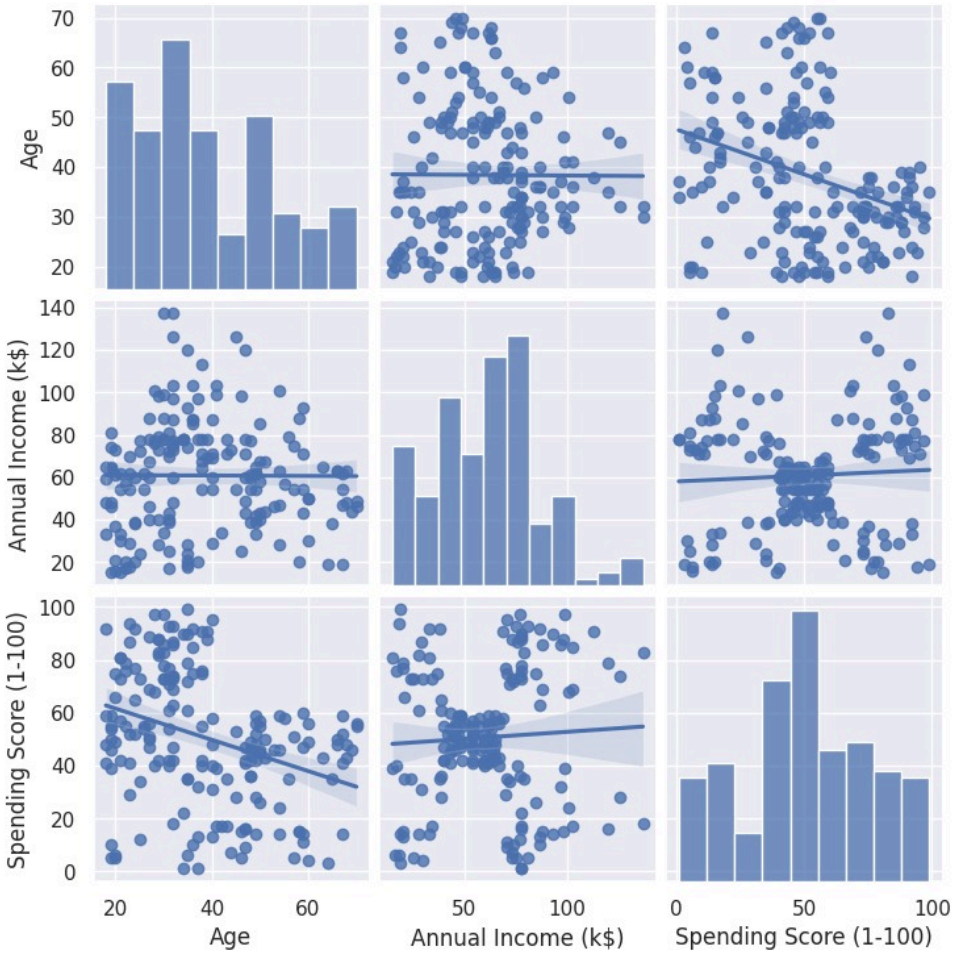
Para esta etapa se ha usado a H2O para las etapas de carga de los datos en su clúster y usamos nuestro predictors necesarios para mostrar en el resultado. En nuestro caso hemos elegido 3: edad, salario anual, puntuación de gasto.

```
predictors = ["Age", "Annual Income (k$)",  "Spending Score (1-100)"]
predictors
```

```
['Age', 'Annual Income (k$)', 'Spending Score (1-100)']
```

- **Análisis Bivariado**

En los resultados se muestra un análisis bivariado ya que se examinan las relaciones entre las variables de Age, Annual Income y Spending Score.



La figura muestra existen correlaciones tanto positivas entre el spending Score y Annual Income, relaciones neutras entre el Annual Income y el Age y Negativa entre Age y Spending Score. Las correlaciones positivas y negativas, o cercana al 1 o -1, significan que existe un correlación entre las variables, pero la relación neutral, o cercana a 0, significa que existe una relación débil entre la variables.

- **Visualización de los datos**

Para visualizar los clusters de manera efectiva, se crea un diagrama de dispersión donde:

- Eje X: Representa la edad del cliente.
- Eje Y: Representa el ingreso anual del cliente (en miles de dólares).
- Color de los puntos: Indica el segmento de cluster al que pertenece cada cliente. Se utiliza una paleta de colores "viridis" para diferenciar visualmente los clusters.

El diagrama de dispersión revela patrones interesantes en los datos:

- Cluster 1 (azul claro): Se agrupa a clientes jóvenes con un ingreso anual relativamente bajo.
- Cluster 2 (verde): Se concentra a clientes de edad intermedia con un ingreso anual alto.
- Cluster 3 (amarillo): Comprende a clientes de mayor edad con un ingreso anual moderado.

Modelización

- **Carga de datos:** Usamos h2o.import_file() para cargar los datos en un objeto H2OFrame. Esta es una estructura de datos optimizada para trabajar con grandes conjuntos de datos distribuidos en clústeres de H2O.
- **División de datos:** Se divide los datos en conjuntos de entrenamiento y prueba utilizando split_frame(). Esto se hace de manera que puedas evaluar el rendimiento del modelo en datos no vistos.

- **Entrenamiento del modelo:** Se utiliza H2OKMeansEstimator para crear un modelo de K-Means. Se especifica el número de clústeres (3), se estandariza y se le da una semilla (1234).
- **Predicciones:** Se hacen predicciones sobre los datos de entrenamiento y prueba utilizando prediction_train y prediction_test.
- **Análisis de los resultados:** Se muestra los resultados del modelo, como los segmentos encontrados, la distribución de los datos en los segmentos y las características promedio de cada segmento.

Resultados

El resultado se muestran 3 segmentos con la media de nuestros 3 predictors: Age, Annual Income (k$) y Spending Score (1-100).

segmento	mean_Age	mean_Annual Income (k$)	mean_Spending Score (1-100)
0	52.2568	59.3378	34.9459
1	33.4146	89.2439	76.4146
2	25.2581	43.7419	52.5

Esta tabla muestra:

1. **Segmento 0 (Prudentes):** Estos clientes de mayor edad, ganan un salario anualmente moderado y su puntuación de gasto es la más baja, por eso pueden ser más prudentes en sus gastos y pueden valorar más la calidad y durabilidad de los productos en lugar de la cantidad.
2. **Segmento 1 (Lujosos)**: Estos clientes son jóvenes, son los que más ganan anualmente y su puntuación de gasto es la más alta, por eso puede que estén dispuestos a gastar en productos y servicios de alta gama, y podrían ser objetivo de campañas de marketing enfocadas en experiencias exclusivas y de lujo.
3. **Segmento 2 (Cautelosos):** Estos clientes son más jóvenes y no ganan mucho anualmente y su puntuación de gasto es moderada, por eso puede que estén buscando un buen equilibrio entre la calidad y el precio, y podrían ser sensibles a ofertas y promociones que les permitan maximizar el valor de su dinero.

Conclusiones

En este trabajo, usamos K-means para segmentar clientes del centro comercial en grupos según edad, salario y gastos . Encontramos tres segmentos distintivos con características y comportamientos de compra únicos. Estos resultados permiten personalizar estrategias de marketing y mejorar la experiencia del cliente. Aunque identificamos patrones de comportamiento, hay áreas para mejorar la precisión de la segmentación y la personalización de servicios.

Recomendaciones

Para futuros trabajos, se recomienda evaluar detalladamente las variables utilizadas en el clustering, considerando la inclusión de otras para una segmentación más precisa. Además,

agregar nuevas variables como la retroalimentación de los clientes a través del análisis de texto. También explorar modelos predictivos para anticipar el comportamiento futuro de los clientes y adaptar las estrategias de marketing. Finalmente, realizar análisis comparativos entre diferentes algoritmos de clustering y técnicas de segmentación para evaluar su efectividad en este contexto del centro comercial.

Referencias bibliográficas

Run-Qing Liu, Hong-Lei Mu, & Young-Chan Lee. (2018). Customer Classification and Market Basket Analysis Using K-Means Clustering and Association Rules: Evidence from Distribution Big Data of Korean Retailing Company. Knowledge Management Review, 19(4), 59–76. https://doi.org/10.15813/KMR.2018.19.4.004
Dataset:https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python/data