Momento de Retroalimentación: Módulo 2 Uso de framework o biblioteca de aprendizaje máquina para la implementación de una solución. (Portafolio Implementación)

Introducción:

La base de Datos establece el historial de crédito de diferentes personas, mediante el uso de esta base de datos y un algoritmo de clustering se busca definir diferentes perfiles de clientes que puedan ser usados más adelante por los bancos para establecer diferentes estrategias de abordamiento, cobro o atención a los clientes dependiendo de los perfiles de cada grupo.

Variables:

Los significados de las variables individuales son los siguientes:

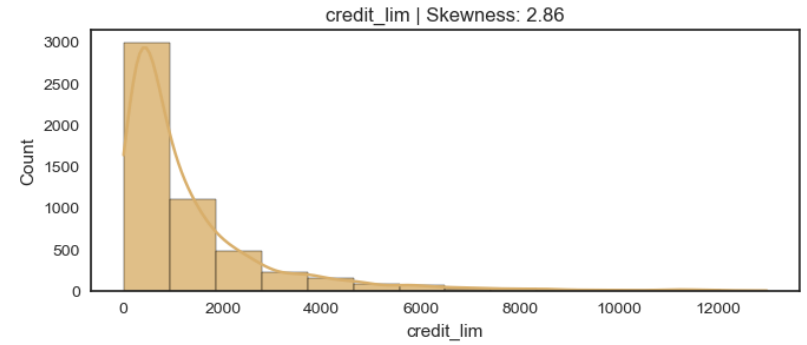
- CUST_ID - Identificación del titular de la tarjeta de crédito.
- BALANCE - Importe del saldo que queda en la cuenta para realizar compras.
- BALANCE_FREQUENCY - Frecuencia de actualización del saldo, puntuada entre 0 y 1 (1 = se actualiza con frecuencia, 0 = no se actualiza con frecuencia).
- PURCHASES - Importe total de las compras realizadas con cargo a la cuenta.
- ONEOFF_PURCHASES - Importe máximo de una sola compra realizada de una sola vez.
- INSTALLMENTS_PURCHASES - Importe de la compra realizada a plazos.
- CASH_ADVANCE - Efectivo entregado por adelantado por el usuario.
- PURCHASES_FREQUENCY - Frecuencia de las compras, puntuada entre 0 y 1 (1 = compras frecuentes, 0 = compras poco frecuentes).
- ONEOFF_PURCHASES_FREQUENCY - Frecuencia de compras únicas realizadas de una sola vez (1 = compras frecuentes, 0 = compras poco frecuentes).
- PURCHASES_INSTALLMENTS_FREQUENCY - Frecuencia de compras realizadas a plazos (1 = se realizan con frecuencia, 0 = no se realizan con frecuencia).
- CASH_ADVANCE_FREQUENCY - Frecuencia con la que se paga en efectivo por adelantado.
- CASH_ADVANCE_TRX - Número de transacciones realizadas con «Efectivo por adelantado».
- PURCHASES_TRX - Número de transacciones de compra realizadas.
- CREDIT_LIMIT - Cantidad máxima de dinero que un titular de tarjeta de crédito puede gastar utilizando su tarjeta de crédito.
- PAYMENTS - Importe de los pagos realizados por el usuario.
- MINIMUM_PAYMENTS - Importe mínimo de pagos realizados por el usuario.
- PRC_FULL_PAYMENT - Porcentaje de pago completo realizado por el usuario.
- TENURE - Tenencia del servicio de tarjeta de crédito para el usuario.

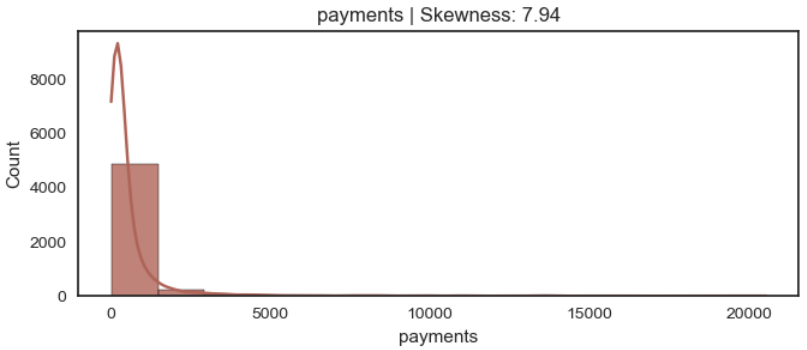
Limpieza de la base de Datos:

Como casi no había valores Nan en las columnas y solo 300 en una columna opte por quitarlos todos ya que esto no iba a afectar el modelo debido a que hay muchos más datos.

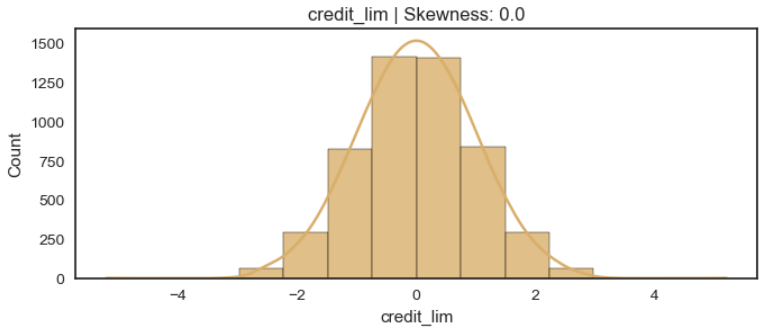
Después se hizo la visualización a través de boxplots para quitar valores que puedan ser outliers que entorpecen el modelo.
Al ver la distribución de las variables podemos aplicarle una transformación a algunas para normalizar un poco su distribución y reducir el impacto de los outliers.

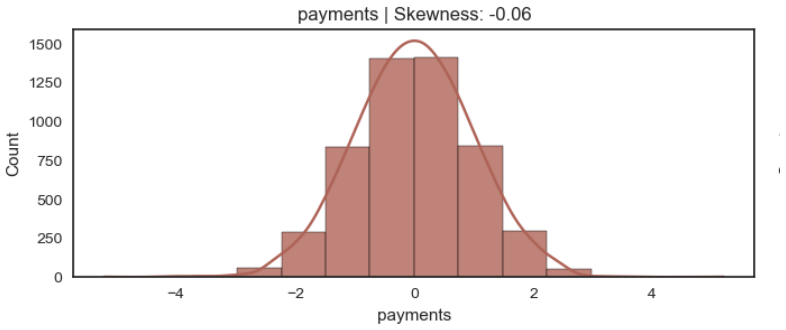
ANTES:





DESPUÉS:



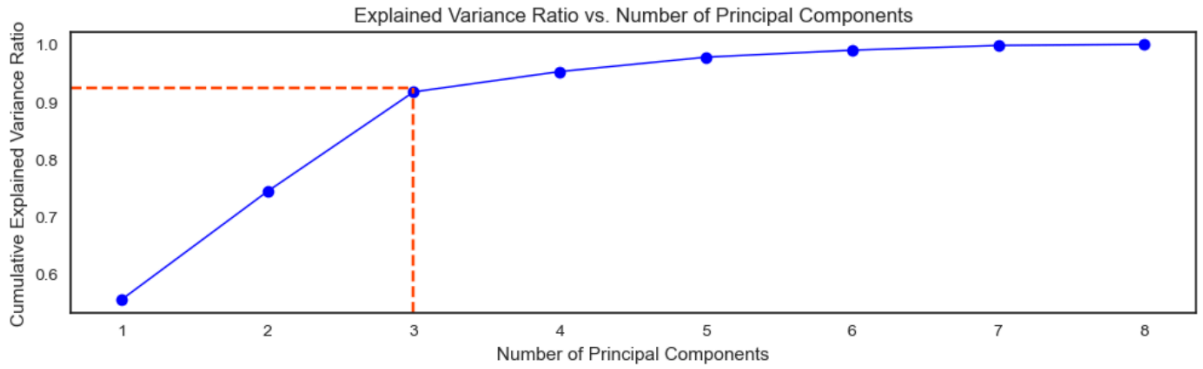


Por último se usa la matriz de correlación para eliminar algunas variables que no aportan mucha información al modelo.

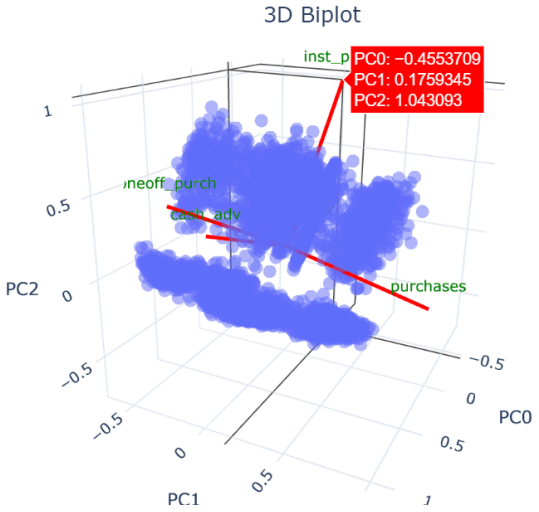
Modelo de Clustering:

PCA:

El PCA (Análisis de Componentes Principales) tiene varias ventajas importantes: reduce la dimensionalidad de los datos, elimina redundancias causadas por la correlación entre variables, y facilita la visualización de datos complejos al convertirlos en nuevas variables llamadas componentes principales. Además, mejora la eficiencia de los modelos de machine learning al reducir el número de variables, disminuyendo el riesgo de overfitting.



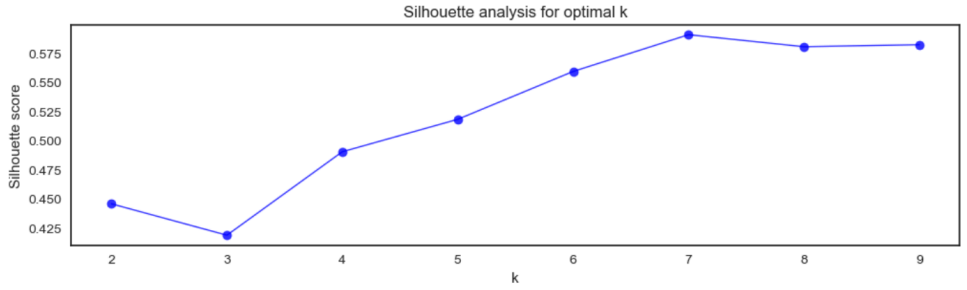
Aquí podemos ver que usar 3 componentes principales es la mejor opción porque explican el 92.5% de la variabilidad de los datos, capturando la mayor parte de la información esencial. Aunque usar solo 2 componentes simplificaría aún más el modelo, perdería una fracción importante de la varianza, por lo que con 3 componentes se logra un buen equilibrio entre simplicidad y retención de información clave sin añadir demasiada complejidad.



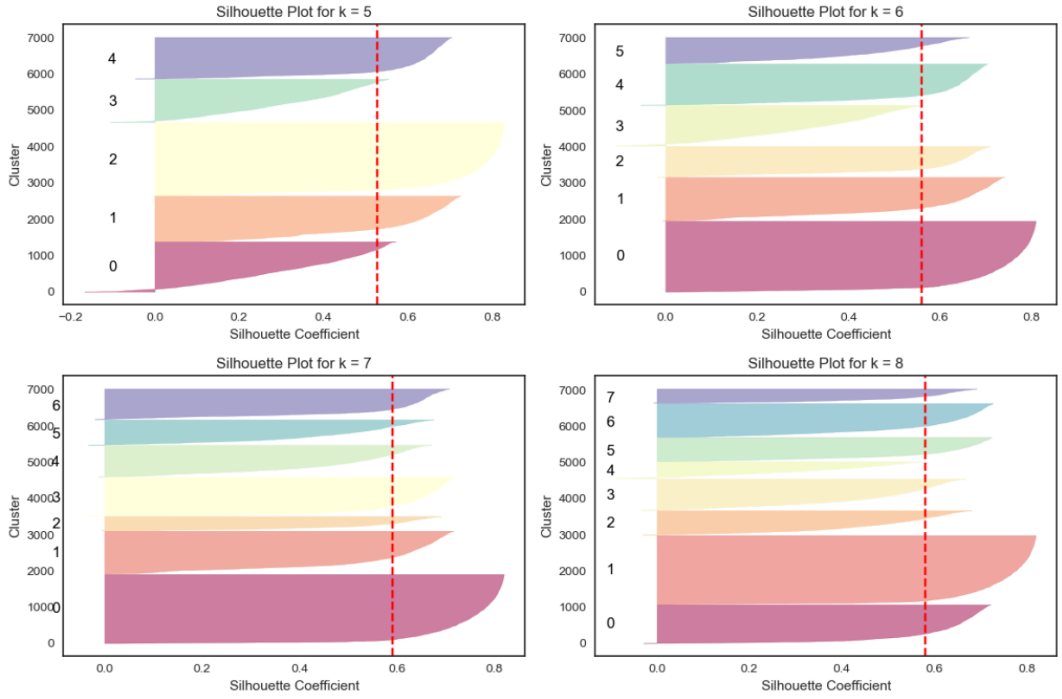
Aquí podemos ver cómo se distribuyen las variables que conocemos dentro de los componentes que usamos en el modelo.

Coeficiente de Silhouette:

El coeficiente de Silhouette es una métrica utilizada para evaluar la calidad de un modelo de clustering, midiendo qué tan bien se agrupan los puntos dentro de su clúster y qué tan distintos están de otros clústeres. El valor de Silhouette varía entre -1 y 1; un valor cercano a 1 indica que los puntos están bien agrupados, mientras que valores cercanos a 0 sugieren que los puntos están en los límites de los clústeres y un valor negativo indica mala asignación de clústeres.



Aquí podemos ver que el número de clusters óptimo puede ir de 5 a 8, siendo 7 el óptimo en cuanto al coeficiente de Silhouette, aunque se podrían considerar usar menos si no se quisieran tantos clusters pero se estuviera dispuesto a sacrificar un poco la óptima agrupación.



Aquí podemos ver dependiendo del número de agrupaciones que hagamos cuales cumplen con un valor “mínimo” del coeficiente como para ser considerados una buena agrupación.

Al final asi quedo el modelo: