Universidad Alfonso X El Sabio

Grado en Ingeniería Matemática

Gestión de Datos

Modelo DWH y CLTV Sector Automotriz

Autor:
González García, María

29 de marzo de 2025

Resumen

El presente trabajo expone el diseño e implementación de un ecosistema analítico orientado a la consolidación, modelado y explotación de datos del sector automotriz. El proyecto parte de la extracción de datos desde un entorno Azure SQL, con acceso restringido a lectura, para su transformación y carga en un entorno local basado en SQL Server. A través de un proceso ETL, los datos han sido reestructurados desde un modelo Entidad-Relación (ER) hacia un modelo dimensional en estrella, compuesto por una tabla de hechos y múltiples dimensiones (cliente, producto, tiempo y geografía).

Sobre esta infraestructura, se han generado métricas clave, destacando el cálculo de la probabilidad de abandono (churn) mediante regresión lineal, cuyos coeficientes han sido integrados en una vista consolidada de cliente. Posteriormente, se ha estimado el Customer Lifetime Value (CLTV) utilizando dicha vista, optimizando la segmentación y análisis del ciclo de vida del cliente.

Adicionalmente, se ha aplicado un análisis de componentes principales (PCA) y un modelo de clustering no supervisado (K-Means) para la identificación de segmentos homogéneos. Los resultados han sido integrados en un entorno visual interactivo mediante Power BI y una aplicación desarrollada en Streamlit, facilitando la exploración y análisis estratégico de los perfiles de clientes.

Índice

1. Contexto y Objetivos del Proyecto.

En la industria automotriz, la integración y explotación eficiente de los datos provenientes de los distintos procesos de negocio —producción, comercialización, postventa y fidelización— constituye un factor estratégico clave para la sostenibilidad y la rentabilidad a largo plazo. Sin embargo, la dispersión de la información en múltiples sistemas operacionales (ERP, CRM, logística y postventa), así como la ausencia de un entorno analítico unificado, dificultan la obtención de insights estratégicos y limitan la capacidad de las empresas para anticiparse a las necesidades de sus clientes.

En este contexto, la dispersión de datos, unida a la ausencia de un modelo analítico consolidado, impide el desarrollo de métricas clave como la probabilidad de abandono (churn) o la estimación del Customer Lifetime Value (CLTV), parámetros fundamentales para optimizar las estrategias de retención y maximización del valor de cada cliente.

El proyecto, desarrollado tiene como objetivo principal el diseño e implementación de una infraestructura analítica que permita transformar los datos operacionales en información estructurada, accesible y explotable. Esta infraestructura estará orientada a:

- Consolidación de los datos heterogéneos procedentes de distintas fuentes en un entorno único y accesible.
- Facilitar el análisis del comportamiento de los clientes y la medición de indicadores clave.
- Estimación del CLTV a medio plazo, integrando información histórica de ventas, postventa y perfil demográfico.
- Identificación de segmentos de clientes con características y patrones de comportamiento homogéneos.
- Poner a disposición de la organización herramientas de visualización interactiva que permitan una toma de decisiones informada y basada en datos.

El desarrollo de este proyecto permite, por tanto, cerrar el ciclo de la gestión de datos, desde la captura y transformación de información dispersa hasta su explotación analítica y visual, alineándose con los objetivos estratégicos de fidelización y crecimiento sostenible de la empresa.

2. Metodología y Arquitectura Técnica.

El desarrollo del proyecto se ha abordado mediante una **metodología secuencial** centrada en la implementación de un ecosistema analítico robusto y escalable. El proceso se ha estructurado en distintas fases que cubren desde la identificación de fuentes hasta la visualización avanzada de los resultados, utilizando herramientas especializadas y técnicas de modelado de datos, ingeniería de características y análisis predictivo.

En primer lugar, se llevó a cabo la extracción de los datos desde un entorno **Azure SQL Database** de solo lectura, donde se encontraban distribuidos en distintas tablas asociadas a los sistemas ERP, CRM, logística y postventa. A partir de esta información, se diseñó un modelo dimensional basado en una estructura en estrella (**Star Schema**).

Una vez definido el modelo lógico, se construyó un pipeline ETL encargado de realizar la extracción, transformación y carga de los datos en un entorno local basado en **SQL Server**. Durante esta fase se llevaron a cabo tareas de limpieza, incluyendo métricas agregadas a nivel de cliente y producto.

Sobre este entorno, se construyó una variable continua de **Churn**, abandono del cliente, utilizando regresión lineal como modelo predictivo. Los coeficientes obtenidos se almacenaron en una tabla auxiliar y fueron aplicados para estimar la probabilidad de abandono de cada cliente en la vista analítica, diseñada como base para el cálculo posterior del **Customer Lifetime Value** (CLTV).

El cálculo del CLTV se realizó considerando múltiples fuentes de ingresos (ventas, postventa, suscripciones, etc.), ajustadas por la probabilidad de retención estimada a partir del modelo de churn. El horizonte de análisis se fijó en cinco años, y se empleó una tasa de descuento acorde al sector para reflejar el valor temporal del dinero.

Con el objetivo de enriquecer el análisis y permitir una segmentación más estratégica, se aplicaron técnicas de reducción de dimensionalidad mediante **Principal Component Analysis** (PCA), seguidas de un modelo de clustering no supervisado basado en **K-Means**. Esto permitió identificar grupos de clientes con patrones de comportamiento y valor similares.

Finalmente, los resultados se integraron en dos entornos visuales: por un lado, un conjunto de dashboards en **Power BI** enfocados al análisis de métricas de negocio y comportamiento de cliente; y por otro, una aplicación interactiva desarrollada en **Streamlit** para la exploración de segmentos, características clave y escenarios de análisis avanzados.

3. Esquema del Flujo de Datos.

El flujo de datos desarrollado tiene como objetivo trasladar información operativa, almacenada en la nube (**Azure SQL Database**), hacia un entorno analítico local (**SQL Server**) donde es consolidada, transformada y preparada para su explotación estratégica. Este proceso se articula mediante un pipeline ETL que automatiza la extracción, limpieza y transformación de los datos, además de estructurarlos bajo un modelo dimensional optimizado para el análisis.

Una vez integrados en el entorno local, los datos son explotados mediante herramientas analíticas como **Power BI**, para la construcción de dashboards estratégicos, y **Streamlit**, a través de una aplicación interactiva que permite el análisis segmentado del CLTV y la identificación de perfiles de cliente.La Figura 1 muestra la arquitectura completa del flujo de datos, desde las fuentes operacionales hasta las herramientas de explotación.

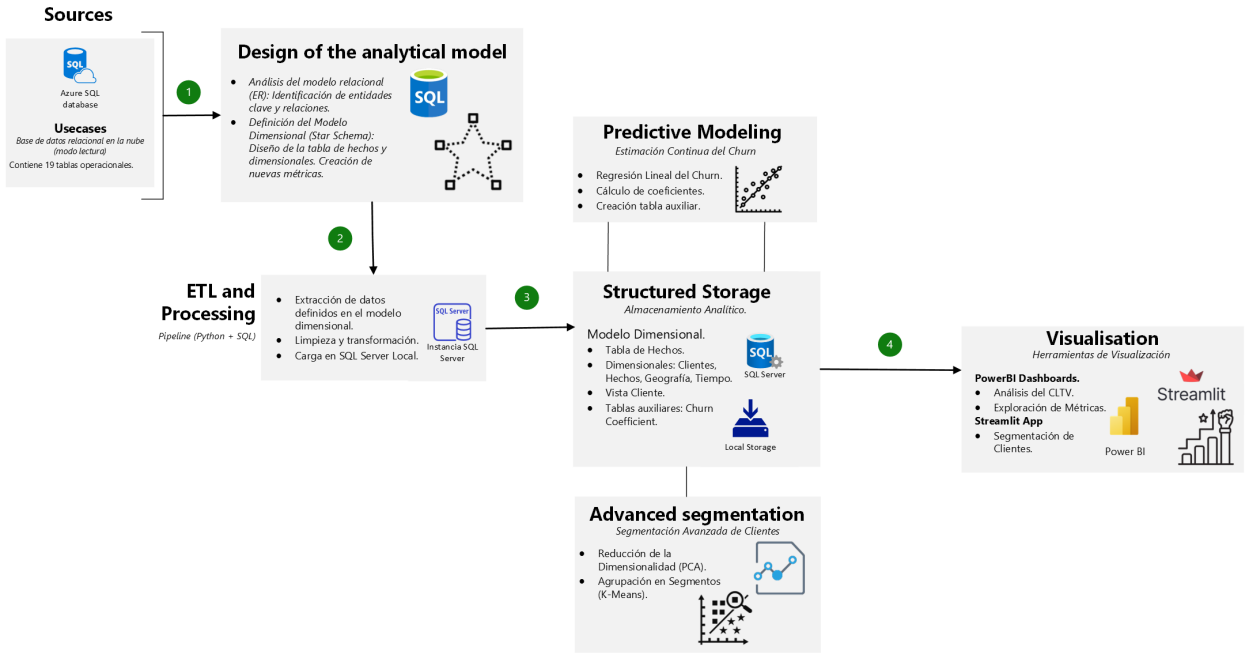


Figura 1: Esquema del flujo de datos: de Azure SQL a entorno analítico local.

Repositorio técnico del proyecto.

Todo el desarrollo técnico de este proyecto —incluyendo consultas SQL, notebooks de análisis, esquemas preliminares, pruebas de modelado y documentación auxiliar— se encuentra disponible en el siguiente repositorio: `https://github.com/mgonzalz/gd_modelos/tree/01_modelo-relacional`.

4. Modelado de Datos: Modelo Entidad-Relación y Modelo Dimensional.

4.1. Modelo Entidad-Relación (ER).

El punto de partida para la construcción del entorno analítico ha sido un modelo Entidad-Relación (ER) procedente de los sistemas operacionales de la organización, desplegado en el entorno Azure SQL Database. Este modelo representa la estructura lógica de los datos y las relaciones existentes entre las distintas entidades que conforman el ecosistema de negocio.

El modelo ER integra información procedente de los sistemas ERP, CRM, logística y postventa, estructurada en torno a un conjunto de tablas interrelacionadas. Las entidades principales incluyen datos de ventas, clientes, productos, revisiones postventa, logística y datos geográficos, así como tablas auxiliares de referencia como forma de pago, motivo de venta o tipologías de producto.

Las relaciones entre entidades se han establecido de forma rigurosa, asegurando la integridad referencial mediante claves primarias y foráneas. En términos de cardinalidad, predominan las relaciones uno a muchos (1:N), donde una única instancia de una entidad principal puede estar asociada a múltiples registros en las entidades dependientes (por ejemplo, un cliente puede tener varias ventas asociadas).

El diagrama Entidad-Relación original, que recoge la estructura lógica de los datos operacionales utilizada como base para el desarrollo del modelo analítico, se presenta en la Figura 2, ubicada en la página siguiente.

4.2. Modelo Dimensional.

El análisis del modelo ER permitió identificar las entidades clave y establecer la trazabilidad de los datos necesarios para el análisis estratégico. No obstante, su estructura relacional, orientada a la operativa diaria, no resultaba adecuada para su explotación analítica, motivando la necesidad de su transformación hacia un modelo dimensional optimizado para la consulta y el análisis. Es por ello que se necesita un **Modelo Dimensional** bajo un esquema en estrella (**Star Schema**).

El diseño dimensional se articula en torno a una tabla de hechos denominada **Fact Sales**, que consolida las métricas transaccionales y operacionales relevantes para el análisis estratégico. Dicha tabla incluye variables relacionadas con ventas, postventa y comportamiento del cliente, así como la métrica calculada de probabilidad de abandono (**Churn**) y el margen económico asociado a cada transacción.

Criterio aplicado para la definición de Churn.

La variable Churn fue definida a partir de un enfoque basado en la **temporalidad de la última revisión** registrada por cliente, estableciendo un umbral de 400 días como criterio discriminante. Así, se consideró que un cliente permanecía activo si había realizado una revisión dentro de ese período, mientras que un lapso superior fue interpretado como señal de abandono. Para los casos en los que no se disponía de datos sobre revisiones, se aplicó un criterio compensatorio sustentado en la **antigüedad del vehículo**: si el automóvil tenía menos de un año desde su matriculación, se asumió que el cliente aún no había tenido oportunidad de efectuar una revisión, clasificándolo como no Churn; por el contrario, si el vehículo superaba dicho umbral,indicativo de deserción. Finalmente, en los registros residuales en los que no se contaba ni con datos de revisiones ni con información sobre la antigüedad del vehículo, se asignó por defecto la condición de Churn, dada su baja incidencia.

Figura 2: Modelo Entidad-Relación (ER) del sistema fuente en Azure SQL

Complementando a la tabla de hechos, se han definido cuatro **tablas de dimensiones** que proporcionan el contexto necesario para la desagregación de los datos y el análisis multiaxial. Estas dimensiones permiten enriquecer la interpretación de las métricas almacenadas en **Fact Sales**, facilitando la construcción de consultas analíticas flexibles y eficientes. A continuación, se describen en detalle:

- **Customer Dimension:** Contiene información clave sobre el perfil del cliente, incluyendo variables demográficas como la edad y el código postal, así como atributos socioeconómicos como la renta media estimada y la clasificación Mosaic. Estos atributos permiten segmentar la base de clientes por nivel adquisitivo, localización o estilo de vida.
- **Product Dimension:** Agrupa las características técnicas y comerciales de los productos vendidos. Incluye campos como el tipo de combustible, equipamiento, potencia, márgenes unitarios, costes asociados al transporte y campañas de marketing. Esta dimensión permite analizar la rentabilidad por categoría de producto, evaluar la relación entre ciertas configuraciones y la retención de clientes, y detectar patrones de consumo asociados a determinadas gamas.
- **Geo Dimension:** Recoge la estructura organizativa y territorial de los puntos de venta. Incluye información como el identificador de tienda, su descripción, la provincia correspondiente y la zona comercial asociada. Esta dimensión facilita el análisis geográfico del negocio.
- **Time Dimension:** Incorpora una descomposición de la variable temporal a partir de la fecha de la transacción. Se incluyen atributos como el día, mes, año, semana del año, día de la semana, e indicadores booleanos para identificar festivos, fines de semana o días laborales.

Todas las dimensiones se han vinculado a la tabla de hechos mediante claves foráneas, garantizando la integridad referencial y mejorando el rendimiento de las consultas en el **entorno local de SQL Server**. La Figura 3 presenta el diseño conceptual del modelo dimensional.

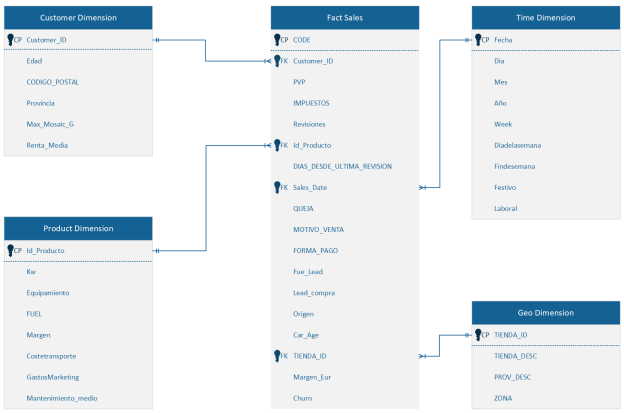


Figura 3: Esquema Dimensional en Estrella implementado en el entorno analítico local.

Disponibilidad de consultas y código fuente.

Todas las consultas SQL desarrolladas para la extracción, transformación y carga de los datos, así como los scripts utilizados en la construcción del modelo dimensional, se encuentra disponible en el **código fuente** del proyecto, ubicado en la carpeta `databases/dwh`.

5. Análisis del CLTV y Construcción de la Visión Cliente.

Una vez consolidado el modelo dimensional, se procedió a la explotación analítica de la información con el objetivo principal de estimar el **Customer Lifetime Value** (CLTV) y construir una visión integral de cada cliente, integrando datos transaccionales, demográficos y predictivos.

5.1. Estimación continua de la probabilidad de Churn.

Con el objetivo de enriquecer el análisis y disponer de una métrica de abandono más granular, se desarrolló un modelo de regresión lineal para estimar la probabilidad de churn como una variable continua. Esta estimación permite cuantificar el riesgo de abandono de cada cliente en un rango de valores entre 0 y 1, en lugar de utilizar una clasificación binaria, proporcionando mayor precisión para su integración posterior en el cálculo del **Customer Lifetime Value** (CLTV).

El punto de partida fue la construcción de un conjunto de datos específico que consolidara las principales variables explicativas del comportamiento de los clientes. Para ello, se diseñó una consulta sobre el entorno analítico local que agrupaba información clave a nivel de cliente y producto, generando un dataset con los siguientes atributos:

- El precio de venta al público (**PVP**) del producto adquirido.
- La **edad media del vehículo** asociado a cada cliente.
- El **kilometraje medio entre revisiones**.
- El **número medio de revisiones** realizadas.
- El **porcentaje medio de churn** observado por grupo de cliente y producto.

A partir de este conjunto de datos, se construyó un modelo de regresión lineal multivariable. El modelo fue entrenado sobre el conjunto completo de observaciones, validando su desempeño mediante métricas estándar: un **coeficiente de determinación (R^2) de 0,63** y un **error cuadrático medio (*MSE*) de 0,026**, lo que refleja un ajuste razonable para la finalidad analítica perseguida.

Como resultado del proceso de modelado, se obtuvieron **coeficientes asociados** a cada una de las variables explicativas, que cuantifican el efecto de cada atributo sobre la probabilidad estimada de churn. Estos coeficientes fueron almacenados en una nueva tabla denominada **Churn Coefficients**, situada en el **entorno local de SQL Server**, con el propósito de integrarlos posteriormente en la vista consolidada de cliente y permitir la estimación directa del riesgo de abandono desde el sistema analítico.

Documentación del modelado y conjunto de datos.

El procedimiento completo de extracción, construcción del conjunto de datos y evaluación del modelo de regresión lineal se encuentra documentado en un cuaderno técnico elaborado en *Jupyter Notebook*, ubicado en la carpeta `notebooks`. Adicionalmente, la consulta SQL utilizada para la obtención y modelado del conjunto de datos ha sido incluida en el código fuente del proyecto, ubicado en la carpeta `preprocessing`.

5.2. Integración del modelo predictivo y construcción de la vista consolidada de cliente.

Una vez obtenidos y almacenados los coeficientes del modelo de regresión lineal, se procedió a su integración en la infraestructura analítica local con el objetivo de consolidar la información predictiva y operativa en una única vista estratégica orientada al análisis de cliente.

Para ello, se diseñó y materializó una vista analítica, cuya finalidad es proporcionar una visión unificada y enriquecida de cada cliente, integrando tanto sus características demográficas y comerciales como la probabilidad estimada de abandono y las métricas clave para la posterior estimación del **Customer Lifetime Value** (CLTV). La construcción de esta vista se basó en la combinación de los siguientes elementos:

- **Datos demográficos y socioeconómicos:** Edad, código postal, provincia y clasificación Mosaic del cliente.
- **Variables operacionales:** Métricas de comportamiento como frecuencia de revisiones, antigüedad del vehículo, importe acumulado de ventas, margen medio y gasto medio por cliente.
- **Información predictiva:** Probabilidad estimada de churn, calculada a partir de la aplicación directa de los coeficientes obtenidos en el modelo de regresión lineal.
- **Variables financieras agregadas:** Ingresos históricos, margen medio por cliente y tasa de retención proyectada.

Sobre esta estructura unificada, se procedió al cálculo del **Customer Lifetime Value** (CLTV), indicador estratégico que permite estimar el valor económico potencial de cada cliente a medio plazo. El cálculo se realizó aplicando la siguiente fórmula:

$$CLTV = \sum_{t=1}^{5} \frac{\text{Margen} \times r^t}{(1+i)^t} \tag{1}$$

Dónde:

- **Margen**: Corresponde al margen económico medio asociado al cliente.
- r: Tasa de retención estimada para cada período.
- i: Tasa de descuento aplicada, fijada en un valor de 0,07.
- t: Período considerado en la estimación, con un horizonte temporal de cinco años.

Consulta para la estimación del CLTV.

La consulta SQL utilizada para la construcción de la vista **Visión Cliente** y la estimación del CLTV ha sido implementada en el **entorno local de SQL Server** y se encuentra disponible en el código fuente del proyecto, ubicado en la carpeta `databases/dwh`.

5.3. Exploración y análisis de la distribución del CLTV.

Una vez estimado el **Customer Lifetime Value** (CLTV) para la totalidad de la base de clientes, se procedió a su análisis exploratorio mediante visualizaciones interactivas desarrolladas en **Power BI**. El objetivo de este análisis fue identificar patrones de valor, segmentaciones naturales dentro de la cartera y posibles correlaciones temporales o geográficas asociadas al comportamiento económico del cliente. El dashboard desarrollado incluye, entre otros elementos clave:

- **Indicadores agregados** del total de clientes analizados, el CLTV promedio estimado a cinco años y el margen económico total proyectado.
- **Histogramas de distribución del CLTV**, que permiten visualizar la concentración de clientes por rango de valor.
- **Análisis temporal**, que muestra cómo varía el CLTV según el año de primera compra, el año de captación y el año de la última transacción, permitiendo detectar tendencias por cohortes.
- **Distribución geográfica** del valor de cliente, a través de mapas que cruzan el CLTV promedio con el volumen de clientes por provincia, permitiendo identificar zonas de mayor concentración de valor.

El análisis realizado refleja una distribución heterogénea del CLTV en la base de clientes, evidenciando la existencia de un grupo reducido de clientes cuyo valor potencial es significativamente superior al promedio, así como la presencia de segmentos cuya rentabilidad proyectada es limitada o negativa. Esta información resulta esencial para la posterior definición de estrategias comerciales y de fidelización orientadas a maximizar la rentabilidad de la cartera.

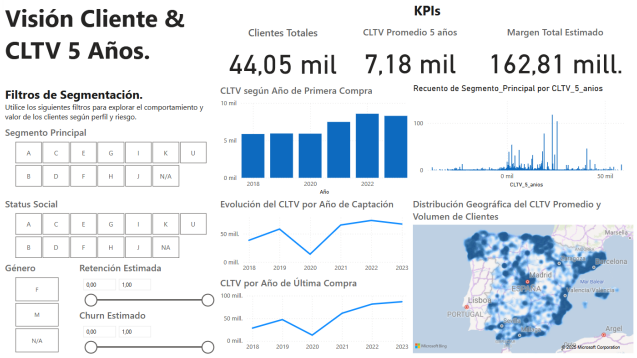


Figura 4: Panel CLTV interactivo.

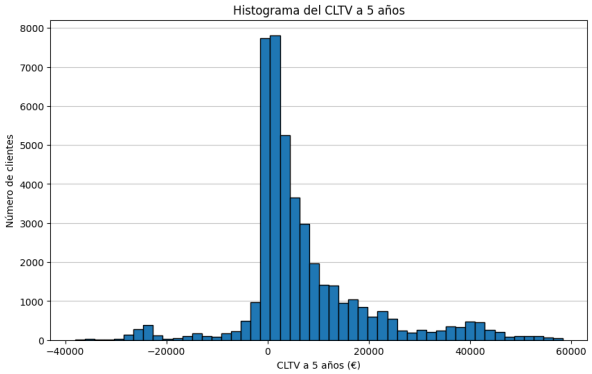


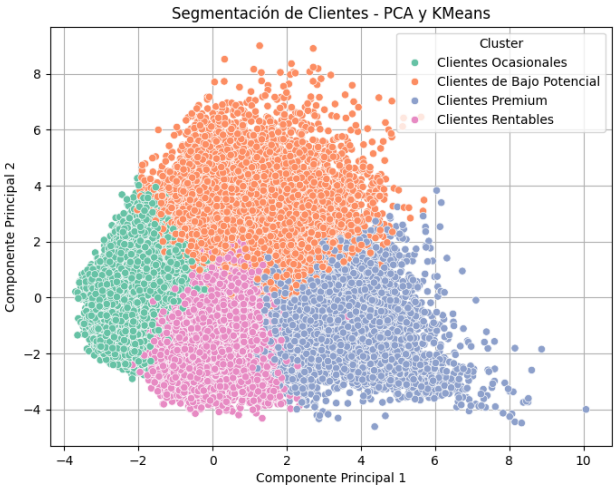
Figura 5: Histograma CLTV 5 años.

Acceso al cuadro de mando y datos utilizados.

El cuadro de mando interactivo desarrollado en **Power BI**, utilizado para la exploración y análisis del CLTV, se encuentra disponible en la carpeta `dashboard` del proyecto. Para su correcta visualización y funcionamiento en un entorno local, se proporciona un archivo `CSV` ubicado en la carpeta `data`, que contiene el conjunto de datos empleado en el análisis. Alternativamente, en caso de haberse ejecutado previamente todo el flujo de extracción y modelado descrito en los apartados anteriores, el cuadro de mando puede conectarse directamente al entorno local de **SQL Server**, utilizando la base de datos consolidada.

6. Segmentación Avanzada mediante PCA y Clustering.

Con el objetivo de identificar perfiles diferenciados dentro de la cartera de clientes, se llevó a cabo un proceso de segmentación basado en técnicas de aprendizaje automático no supervisado. La metodología se estructuró en dos etapas: en primer lugar, se aplicó un **Análisis de Componentes Principales** (PCA) para reducir la dimensionalidad de las variables relevantes extraídas de la vista **Visión Cliente**, concentrando la mayor parte de la información en dos componentes principales.
A partir de estos componentes, se empleó el algoritmo de **Clustering K-Means**, determinando como óptimo un total de cuatro segmentos. Cada grupo presenta características diferenciadas en términos de comportamiento, valor económico y probabilidad de abandono, facilitando la toma de decisiones estratégicas basadas en estos perfiles.



Disponibilidad del análisis de segmentación.

Los segmentos identificados han sido integrados en la vista consolidada y están disponibles para consulta en la aplicación interactiva desarrollada en **Streamlit** (`https://dashboard-clients.streamlit.app/`). La documentación técnica completa se encuentra en la carpeta `notebooks`.

7. Conclusiones y Plan de Mejora.

El proyecto ha permitido construir un entorno analítico robusto para estimar el **Customer Lifetime Value** (CLTV), analizar el comportamiento de los clientes y segmentar la cartera en función de su valor potencial y probabilidad de abandono. La combinación de técnicas de regresión, análisis de componentes principales (PCA) y clustering no supervisado ha aportado profundidad al análisis, permitiendo una toma de decisiones basada en datos y con visión estratégica.

7.1. Plan de mejora y continuidad.

De cara a su evolución, se propone un plan de mejora enfocado en escalar y profesionalizar la solución, con especial énfasis en su despliegue en la nube. Este plan incluye las siguientes líneas de actuación:

- **Contenerización del entorno analítico:** encapsular la base de datos, los scripts ETL y los notebooks analíticos mediante **Docker**, facilitando su portabilidad, mantenimiento y despliegue controlado.
- **Migración a la nube:** alojar el modelo dimensional y la vista de cliente en servicios gestionados como **Azure SQL Database** o **Amazon RDS**, garantizando disponibilidad, rendimiento y seguridad.
- **Automatización del pipeline ETL:** orquestar las tareas de extracción, transformación y carga mediante herramientas como **Azure Data Factory**, reduciendo la intervención manual y mejorando la eficiencia operativa.