



# **Segmentación de Clientes en el Sector Acuícola mediante Algoritmos de K- Medias y Árboles de Decisión: Un Enfoque Aplicado a la Industria del Balanceado para Camarón**

**Sofía Norelia Jiménez Onofre & Edwin Adrián Calderón Zambrano**

**Facultad de Posgrado**

**Maestría en Inteligencia de Negocios y Ciencia de Datos**

**14/junio/2025**

**Palabras clave: Componentes Principales, K-means, Árboles de Decisión,  
Inteligencia Artificial, Segmentación, Marketing, Pricing, Clústers.**

## A. Resumen Ejecutivo

En la industria del balanceado para camarón, la segmentación efectiva de clientes es esencial para optimizar estrategias comerciales como la fijación de precios, campañas de marketing y gestión de cartera. Tradicionalmente, muchas empresas del sector acuícola han basado esta segmentación únicamente en el volumen de compra mensual, lo que limita la comprensión integral del valor y comportamiento de los clientes (Wedel & Kamakura, 2012).

Este proyecto propone una segmentación más robusta y precisa mediante la aplicación de algoritmos de K-means (Dofadar, Khan, & Alam, 2024) y árboles de decisión (Breiman, Freidman, Olshen, & Stone, 1984), a una base de datos de clientes. Estas metodologías permiten identificar grupos homogéneos considerando múltiples variables como el margen de rentabilidad, puntualidad en pagos, tipo de producto adquirido y frecuencia de compra.

A través de este enfoque, se busca proporcionar a la empresa una herramienta analítica que facilite la clasificación estratégica de sus clientes, maximizando la eficiencia comercial y la rentabilidad. (John, Shobayo, & Ogunleye, 2023).

## B. Resumen Ejecutivo Gráfico

La *Ilustración 1*, muestra de forma visual el resultado del Análisis de Componentes Principales (PCA). En el diagrama de dispersión se tiene dos ejes que corresponden a las dimensiones del modelo Dim1 y Dim2, en este diagrama se visualiza cómo se agrupan y relacionan las variables del estudio. Finalmente se ve cómo las variables de Producto\_LARVAS, Margen, Riesgo y Tecnificado\_SI presentan una alta contribución a la varianza explicada.

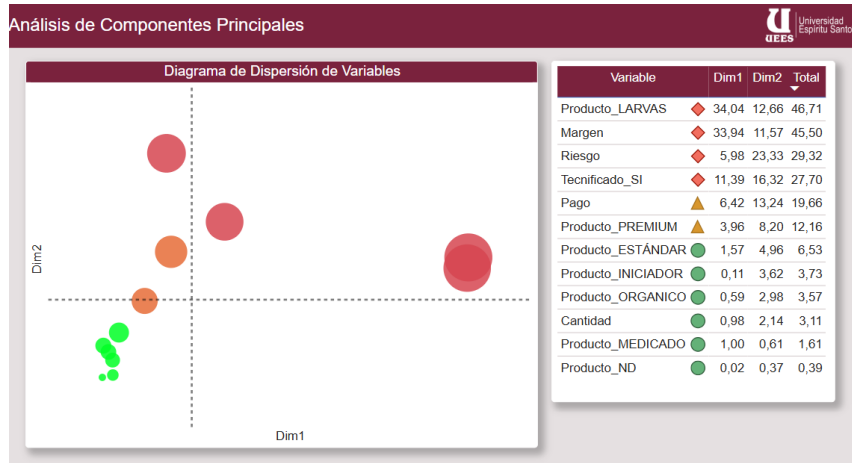


Ilustración 1. Análisis de Componentes Principales. Fuente: Power BI Desktop (Elaboración: Autores).

La *Ilustración 2*, muestra el resultado de la aplicación del algoritmo de K-medias para la clusterización de los clientes basándonos en los datos de sus componentes principales. El dashboard muestra de forma visual los detalles de cada segmento como por ejemplo, el Clúster 1 muestra una mayor proporción con 295 clientes, por otro lado muestra también el diagrama de dispersión usando las variables de Margen y Riesgo para identificar cómo se agrupan los clientes de cada clúster basados en esta información. Por último, junto con la *Ilustración 3*, se muestran gráficos de barras que muestran el promedio de cada variable para entender a mayor profundidad el detalle de cada clúster, por ejemplo, el clúster 2 tiene un mayor margen promedio respecto al resto de clústers. Mientras que su riesgo es similar al clúster 1.

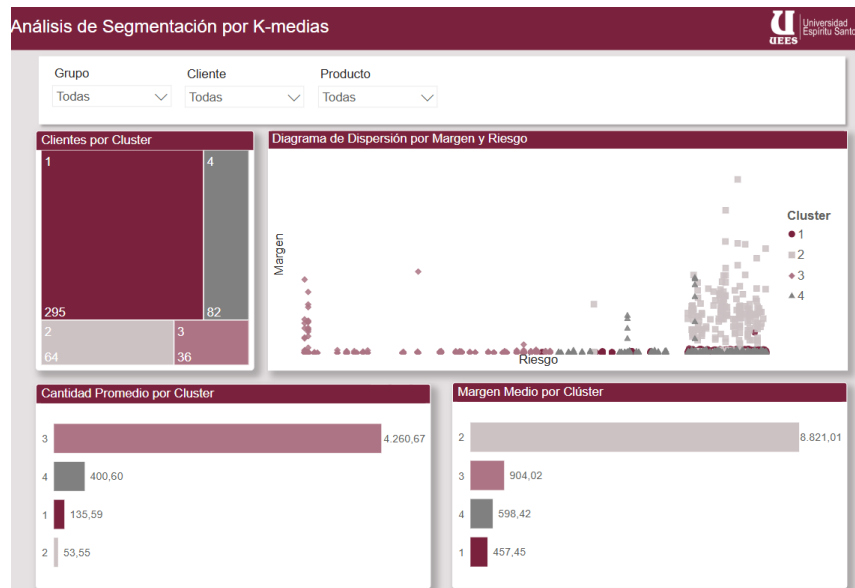


Ilustración 2. Análisis de Segmentación Final por K-medias. Fuente: Power BI Desktop (Elaboración: Autores)

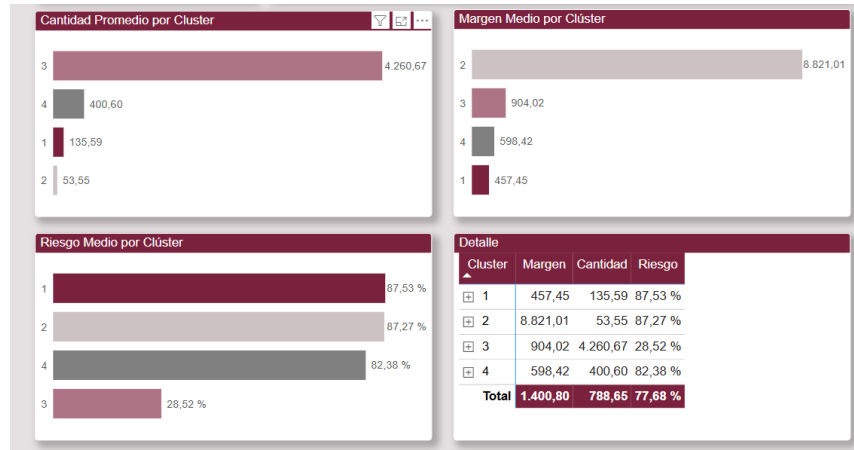


Ilustración 3. Detalle de los segmentos resultantes. Fuente: Power BI Desktop (Elaboración: Autores)

Por último, tenemos en la *Ilustración 4*, un árbol de decisión que nos muestra visualmente las reglas de decisión en las que se basó el Algoritmo de K-Medias para realizar la clusterización, este árbol nos ayuda para entender cómo se clasificaría un nuevo cliente basados en variables como *Tecnificado\_SI*, *Producto\_LARVAS* y *Riesgo*.

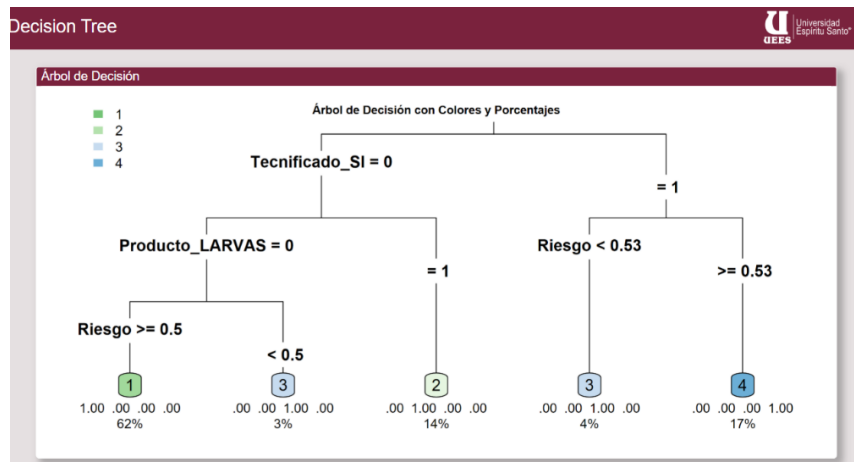


Ilustración 4. Árbol de decisión con los criterios de Clusterización. Fuente: Power BI Desktop (Elaboración: Autores)

## C. Detalle Técnico

### 1. Fuente de Datos

Para el desarrollo de este proyecto, se utilizó inicialmente una base de datos interna proporcionada por una empresa del sector acuícola. Esta información fue extraída del sistema ERP corporativo y exportada a archivos en formato Excel, abarcando el período comprendido entre 2023 y abril de 2025. La base de datos incluyó únicamente a los clientes activos registrados en el sistema cuya actividad comercial está vinculada a la compra de balanceado para camarón. Sin embargo,

debido a políticas de confidencialidad y a la no obtención de una autorización formal para su uso público, se procedió a realizar un proceso de anonimización y adaptación de los datos. Este procedimiento permitió preservar las características estructurales y patrones de comportamiento relevantes, garantizando al mismo tiempo la integridad del análisis y el cumplimiento de las normativas.

Dado que el objetivo principal del proyecto es segmentar a los clientes de manera más precisa para la toma de decisiones, se realizó una selección criteriosa de variables con base en dos principios: disponibilidad real en el sistema y relevancia estratégica para el negocio. Esta metodología se alinea con las recomendaciones de (Wedel & Kamakura, 2012), quienes señalan que una segmentación efectiva debe considerar variables que reflejen tanto el comportamiento observable del cliente como características internas que impacten la rentabilidad, por lo que, se incluyeron las siguientes variables:

- **Cliente:** Identificador único indispensable para el proceso de segmentación.
- **Tipo de alimento adquirido:** Clasificado en categorías premium, estándar y económico. Esta variable permite capturar el perfil de consumo del cliente y su disposición a pagar por valor agregado.
- **Toneladas compradas:** Se utilizó como una medida del volumen de compra, históricamente considerada como el criterio base en la segmentación actual de la empresa.
- **Margen por tonelada:** Indicador de rentabilidad por cliente, fundamental para distinguir clientes no solo por volumen, sino por su contribución real al negocio. Esto responde a la necesidad de incorporar indicadores de valor en la segmentación, tal como sugieren (Kotler & Keller, 2016).
- **Riesgo financiero:** Medido como el porcentaje de veces que el cliente ha cancelado sus facturas dentro del plazo establecido. Esta variable permite incorporar una dimensión financiera en la segmentación, considerando no solo cuánto compra un cliente, sino también cuán confiable es en términos de pago (Verhoef & Lemon, 2013).
- **Tecnificación:** Variable dicotómica que indica si el cliente cuenta con sistemas de alimentación automática

en sus piscinas de cultivo. Se considera un proxy del nivel de adopción tecnológica, lo cual puede relacionarse con una mayor sofisticación en las decisiones de compra y en la planificación de abastecimiento.

La elección de estas variables busca ir más allá de una clasificación empírica basada únicamente en el volumen, como se hace actualmente, y proponer una segmentación que integre comportamiento de compra, rentabilidad, riesgo y nivel de desarrollo del cliente, lo que resulta más representativo para fines estratégicos como fijación de precios, gestión de cartera, diseño de ofertas diferenciadas y planificación comercial (Dolničar, 2004).

A continuación, se presenta el diccionario de datos con las variables enlistadas para el desarrollo del siguiente proyecto.

Tabla 1. Diccionario de datos.

Variable	Tipo de Dato	Descripción
Grupo	Categorico	Grupo o Compañía
Cliente	Categorico	Subgrupo/Cliente final
Tecnificado	Categorico	Si el cliente cuenta con equipo tecnológico (SI/NO)
Pago	Nominal	Días en los que el cliente paga facturas (8, 15, 30, 45, 60, etc)
Año	Nominal	La base de datos cuenta con información desde el 2023
Producto	Categorico	Tipo de producto que consumió cada cliente por año (Estándar, Larvas, etc.)
Riesgo	Continua	% de facturas no pagas dentro del tiempo
Cantidad	Continua	Cantidad en toneladas promedio que consumió cada cliente por año
Margen	Continua	Margen de ganancia generada por cliente

Fuente: Power BI Desktop (Elaboración: Autores).

## 2. Técnica

Para el desarrollo del presente proyecto se implementaron tres técnicas multivariadas entre las cuales están: Análisis de Componentes Principales (PCA), Análisis de K-medias y Árboles de decisión. Cada una presenta un impacto para ayudar a reducir la dimensionalidad, agrupar de forma no supervisada y también para comprender patrones.

### **2.1. Análisis de Componentes Principales (PCA).**

El PCA principalmente fue implementado con la finalidad de reducir la dimensionalidad, esta técnica transforma las variables originales en un nuevo conjunto de variables no correlacionadas (componentes principales), las cuales se orden con base en la varianza explicada del modelo. Así se logró identificar las combinaciones lineales que le aportan mayor variabilidad al comportamiento del cliente, lo que redujo la redundancia y mejoró la interpretación de los datos.

En bases de datos donde se almacena información como comportamiento de clientes basándonos en variables como: margen, riesgo, tecnificación y uso de productos, PCA ayuda a detectar estructuras latentes y preparar los datos para una segmentación más estable. Además, permite visualizar los datos en espacios de menor dimensión, facilitando la comprensión de la distribución del mercado. (Jolliffe & Cadima, 2016)

### **2.2. Análisis de K-medias**

El uso del algoritmo de K-means se usó con la finalidad de realizar un clustering no supervisado basado en las variables que expresan patrones comunes entre clientes usando las variables derivadas del PCA. Esta técnica sirvió para identificar grupos internos en los datos que no están definidos de forma explícita. Enfocándonos en la problemática, K-means fue relevante para hacer una segmentación robusta reemplazando la segmentación empírica actual. Este algoritmo es usado frecuentemente en estrategias de marketing y pricing en empresas con volúmenes de información grandes. (Jan, Kamber, & Pei, 2012)

### **2.3. Árboles de decisión.**

El algoritmo de K-means fue relevante para realizar la segmentación, pero las reglas de clasificación no quedan claras. Por ello, fue necesario levantar un árbol de decisión que muestre de forma amigable cuáles fueron los criterios aplicados en la segmentación. Al usar como variable objetivo el clúster asignado, el árbol identifica y muestra las reglas de decisión que permiten entender cómo variables como el riesgo, días de pago, el margen de ganancias, el uso de productos diferenciados (larvas/premium) o el nivel de tecnificación definen la pertenencia a un grupo específico. Los árboles de decisión ayudan a la interpretación visual y destacan por no requerir supuestos estadísticos complejos, lo

que los convierte en una herramienta ideal para la comunicación de resultados con áreas no técnicas. (Breiman, Freidman, Olshen, & Stone, 1984)

## **2.4. Herramientas Aplicadas.**

En el desarrollo del presente proyecto se usó Power BI como herramienta de visualización, esta herramienta cuenta con un motor que permite la ejecución de scripts de R dentro de Power Query tanto para establecer una fuente de datos como aplicarlo a los pasos dentro de la transformación de una tabla. Además, contamos con la opción de Obtener un Objeto visual a partir de la ejecución de un script de R usando librerías como `rpart.plot` que, en la aplicación práctica, nos muestra el árbol de decisión. Esta integración aporta un valor agregado a los análisis tradicionales, ya que evita ejecuciones manuales en R Studio aprovechando al máximo la integración de la información. (Microsoft Corporation, 2024)

## **D. Resultados**

### **1. Análisis de Componentes Principales (PCA)**

El gráfico de dispersión de variables muestra que las variables con mayor contribución a las dos primeras dimensiones principales son:

- Producto\_LARVAS (46.7%)
- Margen (45.5%)
- Riesgo (29.3%)
- Tecnificado\_SI (27.7%)

Esto indica que estas cuatro variables explican gran parte de la varianza en los datos y, por tanto, son importantes para diferenciar a los clientes. Las variables con bajo aporte como Cantidad, o Producto\_MEDICADO fueron descartadas en fases posteriores, reduciendo el ruido en el análisis. Esta etapa nos ayudó a seleccionar las variables más influyentes antes de aplicar algoritmos de agrupamiento. Sin embargo, también se decide considerar las variables Productos (Premium y Estándar).

### **2. Segmentación por K-means**

El clustering generó 4 grupos de clientes bien diferenciados:

- ✓ Cluster 1 (295 clientes): Bajo margen (457.45), baja cantidad, pero alto riesgo (87.53%). Representa clientes con comportamiento financiero riesgoso, pero bajo valor económico.

- ✓ Cluster 2 (64 clientes): Alto margen (8,821.01) y baja cantidad. Es un segmento estratégico: clientes de alto valor que deben ser priorizados.
- ✓ Cluster 3 (36 clientes): Alta cantidad (4,260.67) pero margen bajo (904.02), aunque también con riesgo alto (82.38%). Este grupo puede reflejar clientes grandes en volumen pero con rentabilidad limitada.
- ✓ Cluster 4 (82 clientes): Perfil intermedio en cantidad y margen, con riesgo moderado (28.52%). Potencial para desarrollo comercial.

La cantidad de clientes por clúster indica una distribución razonable sin necesidad de sobreajuste.

### **3. Árbol de Decisión como Interpretador**

El árbol de decisión proporciona una lectura clara de las reglas que definen la asignación a cada cluster. Las variables que definen las divisiones son:

- Tecnificado\_SI
- Producto\_LARVAS
- Riesgo

Por ejemplo:

Si el cliente no es tecnificado y no compra LARVAS, pero tiene riesgo alto, se clasifica como Cluster 1.

Si el cliente es tecnificado y tiene riesgo bajo ( $< 0.53$ ), cae en Cluster 3; mientras que si el riesgo es alto, pasa a Cluster 4.

Este modelo permite traducir los resultados de segmentación en lógica de negocio, facilitando la toma de decisiones para priorizar o mitigar riesgos según la información del cliente.

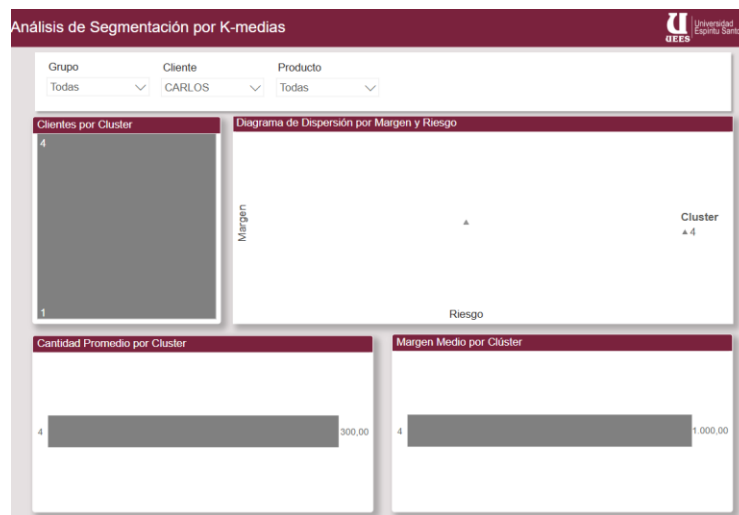
La combinación de PCA, K-means y Árboles de Decisión permitió segmentar de forma objetiva a los clientes, identificando no solo grupos de alto valor económico, sino también aquellos con riesgos que podrían impactar la rentabilidad.

El uso del PCA aseguró que se utilizaran únicamente las variables más relevantes, reduciendo la complejidad del modelo.

K-means generó clusters con diferencias marcadas y consistentes en las variables estratégicas.

El Árbol de Decisión tradujo estos grupos en reglas claras, útiles para personalizar estrategias comerciales y de cobranza.

Por ejemplo, se añadió al Cliente 'CARLOS' el cual cuenta con tecnificación y un riesgo superior a 0.53 por lo que según las reglas de clusterización mostradas en el árbol de decisión, el algoritmo de K-Medias lo asignó al clúster 4.



*Ilustración 5. Resultado de prueba de segmentación con nuevo cliente. Fuente: Power BI Desktop (Elaboración: Autores)*

A partir de los resultados obtenidos, podemos concluir que la combinación de técnicas utilizadas (Análisis de Componentes Principales, K-means y Árboles de Decisión) nos permitió construir una segmentación mucho más completa y útil para el negocio. Como se muestra en la *Ilustración 1*, el PCA nos ayudó a identificar cuáles eran las variables que realmente aportaban valor al modelo, simplificando el análisis. Con esto, al aplicar K-means (ver *Ilustración 2*), logramos visualizar cómo se agrupan los clientes en cuatro clústers diferenciados, cada uno con características claras en términos de margen, riesgo y volumen. La *Ilustración 3* profundiza aún más, mostrando los promedios de estas variables para cada grupo, lo que facilita su interpretación comercial. Finalmente, la *Ilustración 4* nos entrega un árbol de decisión que permite entender, de forma muy práctica, cómo

se asignaría un nuevo cliente a un clúster específico en función de sus características. En conjunto, este análisis no solo nos permite clasificar mejor a nuestros clientes, sino que también aporta detalles muy valiosos que pueden ser utilizados en estrategias futuras de pricing, segmentación y gestión de cartera.

## E. Bibliografía

- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. New York: Chapman & Hall/CRC.
- Dofadar, D., Khan, R., & Alam, G. (2024). LRFS: Online Shoppers' Behavior-Based Efficient Customer Segmentation Model. *IEEE Access*, 96462 - 96480. doi:DOI:10.1109/ACCESS.2024.3420221
- Dolničar, S. (2004). Beyond "Commonsense Segmentation": A Systematics of Segmentation Approaches in Tourism. 52-58. doi:<https://doi.org/10.1177/0047287503258830>
- Jan, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann.
- John, J. M., Shobayo, O., & Ogunleye, B. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics*, 809-823. doi:<https://doi.org/10.3390/analytics2040042>
- Jolliffe, I., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 20150202.
- Kotler, P., & Keller, K. (2016). *Marketing Management*. Pearson Education.
- Microsoft Corporation. (2024). *Power BI Desktop*. Retrieved from Power BI Desktop: <https://powerbi.microsoft.com>
- Verhoef, P., & Lemon, K. (2013). Successful customer value management: Key lessons and emerging trends. *European Management Journal*, 1-15. doi:<https://doi.org/10.1016/j.emj.2012.08.001>
- Wedel, M., & Kamakura, W. (2012). *Market Segmentation: Conceptual and Methodological Foundations*. Springer Science & Business Media.