Revista Científica Multidisciplinar G-ner@ndo ISSN: 2806-5905

Análisis de densidad de siembra óptima en el cultivo de camarón blanco del Pacífico (penaeus vannamei) mediante el modelado de algoritmo no supervisado en baja y alta salinidad

Analysis of optimal stocking density in Pacific white shrimp (penaeus vannamei) farming using unsupervised algorithm modeling in low and high salinity conditions

Andrés Alejandro Paredes Elaje, Joselyne Katiuska Ganchozo Vera & Ivan Leonel Acosta-Guzman

PUNTO CIENCIA.

julio - diciembre, V°6 - N°2; 2025

Recibido: 11-11-2025
Aceptado: 21-11-2025
Publicado: 30-12-2025

PAIS

- Ecuador, Guayaquil
- Ecuador, Guayaquil
- Ecuador, Guayaquil

INSTITUCION

- Universidad de Guayaquil
- Universidad de Guayaquil
- Universidad de Guayaquil

CORREO:

✉ andres.paredese@ug.edu.ec
✉ joselyne.ganchozov@ug.edu.ec
✉ ivan.acostag@ug.edu.ec

ORCID:

https://orcid.org/0009-0004-6161-4708
https://orcid.org/0009-0000-5756-0760
https://orcid.org/0000-0002-1589-1825

FORMATO DE CITA APA.

Paredes, A., Ganchozo, J. & Acosta-Guzman, I. (2025). Análisis de densidad de siembra óptima en el cultivo de camarón blanco del Pacífico (penaeus vannamei) mediante el modelado de algoritmo no supervisado en baja y alta salinidad. *Revista G-ner@ndo, V°6 (N°2). Pág. 3068 – 3089.*

Resumen

Este estudio propone un modelo innovador para optimizar la variable “Libras por Hectárea Final” que se mide en el momento de la cosecha de cultivos de camarón blanco (Penaeus vannamei), integrando técnicas de aprendizaje automático no supervisadas con análisis estadísticos de datos reales de producción. Mediante el uso de algoritmos como k-means y XGBoost, se identifican patrones y relaciones clave entre variables como densidad de siembra, salinidad, estación del año, supervivencia y crecimiento del camarón, permitiendo clasificar los sistemas productivos en clusters con diferencias significativas en rendimiento. Los resultados evidencian que la densidad óptima no es única, sino que varía según el perfil de manejo y condiciones ambientales, y que la supervivencia y el peso promedio son fundamentales para maximizar la producción. La metodología de análisis exploratorio y minería de datos CRISP-DM permitió identificar en que clusters existe eficiencia productiva y que clusters existe bajo rendimiento asociados a prácticas inadecuadas. Este enfoque basado en inteligencia artificial ofrece una herramienta objetiva y flexible para la toma de decisiones en la acuicultura, contribuyendo a la sostenibilidad y rentabilidad del sector. Así, el estudio sienta bases sólidas para futuros desarrollos y aplicaciones en el manejo inteligente del cultivo de camarón, integrando tecnologías avanzadas con conocimiento especializado para optimizar el desempeño y recursos productivos.

Palabras clave: Densidad de Siembra, Aprendizaje Automático, Algoritmo No Supervisado.

Abstract

This study proposes an innovative model to optimize the variable “Final Pounds per Acre,” which is measured at the time of harvest of white shrimp (Penaeus vannamei) crops, integrating unsupervised machine learning techniques with statistical analysis of real production data. Using algorithms such as k-means and XGBoost, key patterns and relationships between variables such as stocking density, salinity, season, survival, and shrimp growth are identified, allowing production systems to be classified into clusters with significant differences in yield. The results show that the optimal density is not unique, but varies according to management profile and environmental conditions, and that survival and average weight are fundamental to maximizing production. The CRISP-DM exploratory analysis and data mining methodology made it possible to identify which clusters are productive and which clusters are underperforming due to inadequate practices. This artificial intelligence-based approach offers an objective and flexible tool for decision-making in aquaculture, contributing to the sustainability and profitability of the sector. Thus, the study lays a solid foundation for future developments and applications in the intelligent management of shrimp farming, integrating advanced technologies with specialized knowledge to optimize performance and productive resources.

Keywords: Planting Density, Machine Learning, Unsupervised Algorithm.

Introducción

El cultivo del camarón blanco del Pacífico (Penaeus vannamei) es fundamental para la acuicultura mundial debido a su aporte económico. La densidad de siembra es un factor clave que influye en la productividad, supervivencia y salud del camarón. Aunque aumentar la densidad puede incrementar la producción, un exceso genera efectos negativos como menor crecimiento, mayor conversión alimentaria y mayor mortalidad, afectando la rentabilidad y sostenibilidad. los métodos tradicionales empleados para establecer esta variable presentan limitaciones al no considerar la complejidad y la variabilidad del entorno productivo (Zhao et al., 2021). En este contexto, la aplicación de técnicas de aprendizaje automático no supervisado permite identificar patrones complejos y agrupar datos en clústeres (Arévalo et al., 2022), estrategia que ha ganado relevancia y efectividad en diversos campos, en particular para la detección de anomalías financieras (Woźniak, 2024), clasificación espacial y el análisis empresarial (Jácome & Flores, 2022; Jang et al., 2018). el uso de modelos de clustering y reducción dimensional (Hussein et al., 2021), y la aplicación de inteligencia artificial en minería y mantenimiento predictivo (Souza & Silva, 2024; Torres Guerra et al., 2021).

El avance tecnológico en plataformas inteligentes, como la desarrollada para la optimización de procesos industriales mediante inteligencia artificial (Wang et al., 2024), ejemplifica la importancia de integrar soluciones computacionales avanzadas en diversos sectores productivos. Inspirado en estos desarrollos y en el potencial de los algoritmos de aprendizaje automático no supervisado para abordar problemas complejos en diferentes áreas (Rameshbabu et al., 2023), este estudio propone maximizar la variable “Libras por Hectárea al Final” del cultivo de camarón blanco. Para ello, se desarrolla un Modelo No Supervisado de Clusterización que permite segmentar los datos productivos y determinar los rangos óptimos de variables asociadas a la siembra y manejo del cultivo bajo distintos escenarios productivos.

Métodos y Materiales

El presente estudio adopta un enfoque longitudinal observacional orientado a evaluar la influencia de variables productivas en el rendimiento del cultivo de P. vannamei. Para la optimización de las variables influyentes en sistemas de cultivo se partió desde una base de datos con variables productivas consiste y estructurada. Posteriormente, se aplicó análisis estadísticos como ANOVA, regresión y componentes principales para identificar los factores que más inciden en el rendimiento; Finalmente, se formularon recomendaciones y simulaciones para mejorar estas variables en el manejo del sistema.

Las variables del sistema productivo camaronero disponibles para el estudio son las siguientes.

Las variables de entrada en el sistema productivo camaronero corresponden a los factores físicos, operativos y ambientales que describen cada ciclo, como la camaronera, piscina, salinidad, año, ciclo, fecha y mes de siembra, fecha y mes de cosecha, época del año, superficie de la piscina (has), densidad de siembra, días de cultivo, peso de pesca, crecimiento lineal, supervivencia, factor de conversión alimenticia (FCA); todas estas variables se registran antes o durante el proceso de cultivo. La variable de salida principal es "LIBRAS/HA FINAL", que representa el rendimiento final de la cosecha por hectárea y se utiliza como indicador clave de productividad y eficiencia en el manejo del sistema.

Tabla 1. *Variables Productivas*

No.	Variable	Descripción	Unidad De Medida
1	CAMARONERA	Nombre/identificador de la finca o empresa camaronera	Texto (categoría)
2	PISCINA	Piscina específica dentro de la camaronera	Texto (categoría)
3	SALINIDAD	Categoría de salinidad del agua	Texto (ordinal)
4	AÑO	Año calendario de la siembra/cosecha	Año
5	CICLO	Número de ciclo productivo en el año	Entero
6	FECHA SIEMBRA	Fecha de inicio de siembra	Fecha (dd-mmm)
7	MES SIEMBRA	Mes abreviado de la siembra	Texto (ordinal)
8	FECHA COSECHA	Fecha de cosecha	Fecha (dd-mmm)
9	MES COSECHA	Mes abreviado de la cosecha	Texto (ordinal)
10	ÉPOCA	Época del año según clima	Texto (ordinal)
11	HAS	Superficie de la piscina	Hectáreas
12	DENSIDAD SIEMBRA	Número de larvas sembradas por hectárea	Unidades/ha
13	DÍAS CULTIVO	Duración del ciclo de cultivo	Días
14	PESO PESCA	Peso promedio obtenido en la cosecha	Gramos
15	CRECIMIENTO LINEAL	Tasa de crecimiento diaria (g/día)	Gramos/día
16	SOBREVIVENCIA	Porcentaje de larvas que llegaron a cosecha	Porcentaje (%)
17	FCA	Factor de conversión alimenticia	Sin unidad
18	LIBRAS/HA FINAL	Rendimiento final de cosecha por hectárea	Libras/hectárea
19	RHD	Rentabilidad Hectárea Día	Valor calculado

Nota: La Tabla 1 detalla las variables disponibles para el estudio.

Procedimiento, análisis estadístico y modelado

El procedimiento aplicado inicia con la importación de librerías y carga de los datos productivos desde un archivo Excel, seguido por la limpieza, estandarización y codificación de variables categóricas y numéricas para su análisis; posteriormente, se realiza exploración descriptiva, análisis de evaluación y ANOVA para buscar relaciones y diferencias estadísticamente significativas, y finalmente se emplean modelos como XGBoost y métodos de importancia de variables (MDI y Permutación de Importancia) para identificar y jerarquizar los factores más influyentes sobre el rendimiento de LIBRAS/HA FINAL; Se procede a determinar el número de clusters mediante la prueba del codo y análisis Silhuette, generar los clusters con el

objetivo de orientar con propuestas de optimización técnica en el manejo de los sistemas de cultivo.

Análisis de Resultados

Carga y limpieza de datos

En el procedimiento de análisis de resultados en el código se inició con la carga y limpieza de la base de datos, asegurando que las variables tengan formatos y tipos adecuados (numéricos y categóricos) para el procesamiento estadístico. Para la limpieza y preparación de datos se emplearon funciones de pandas para transformación y codificación.

Se cuenta con 6928 filas en el DataFrame, cada una con información de cultivos de camarón y características asociadas. Luego del análisis de la data interna fue necesario realizar la trasformación de columnas a categórica nominal, categórica ordinal, numérica y fecha, lo cual facilita tratamientos y modelado adecuado para cada variable durante el análisis, obteniendo 17 columnas para el procesamiento posterior.

Imagen 1 . *Variables disponibles para la investigación*

#	Column	Non-Null Count	Dtype
0	CICLO	6928 non-null	int64
1	HAS	6928 non-null	float64
2	DIAS_CULTIVO	6928 non-null	int64
3	PESO_PESCA	6928 non-null	float64
4	CRECIMIENTO_LINEAL	6928 non-null	float64
5	SOBREVIVENCIA	6928 non-null	float64
6	FCA	6928 non-null	float64
7	LIBRAS/HA FINAL	6928 non-null	int64
8	RHD	6928 non-null	float64
9	ID_CAMARONERA_ENC	6928 non-null	category
10	PISCINA_ENC	6928 non-null	category
11	SALINIDAD_ENC	6928 non-null	category
12	MES_SIEMBRA_ENC	6928 non-null	category
13	MES_COSECHA_ENC	6928 non-null	category
14	EPOCA_CAT	6928 non-null	category
15	FECHA_SIEMBRA_DATE	6928 non-null	datetime64[ns]
16	FECHA_COSECHA_DATE	6928 non-null	datetime64[ns]
17	DENSIDAD_SIEMBRA	6928 non-null	int64

Nota: La imagen 1 muestra en detalle la lista de variables disponibles del sector camaronero disponibles para el estudio.

Análisis descriptivo

Se exploran los datos mediante análisis descriptivos para entender rangos, frecuencias y valores atípicos, y se agruparon las variables relevantes. Se utilizaron técnicas estadísticas como análisis descriptivo y visualización con seaborn y matplotlib.

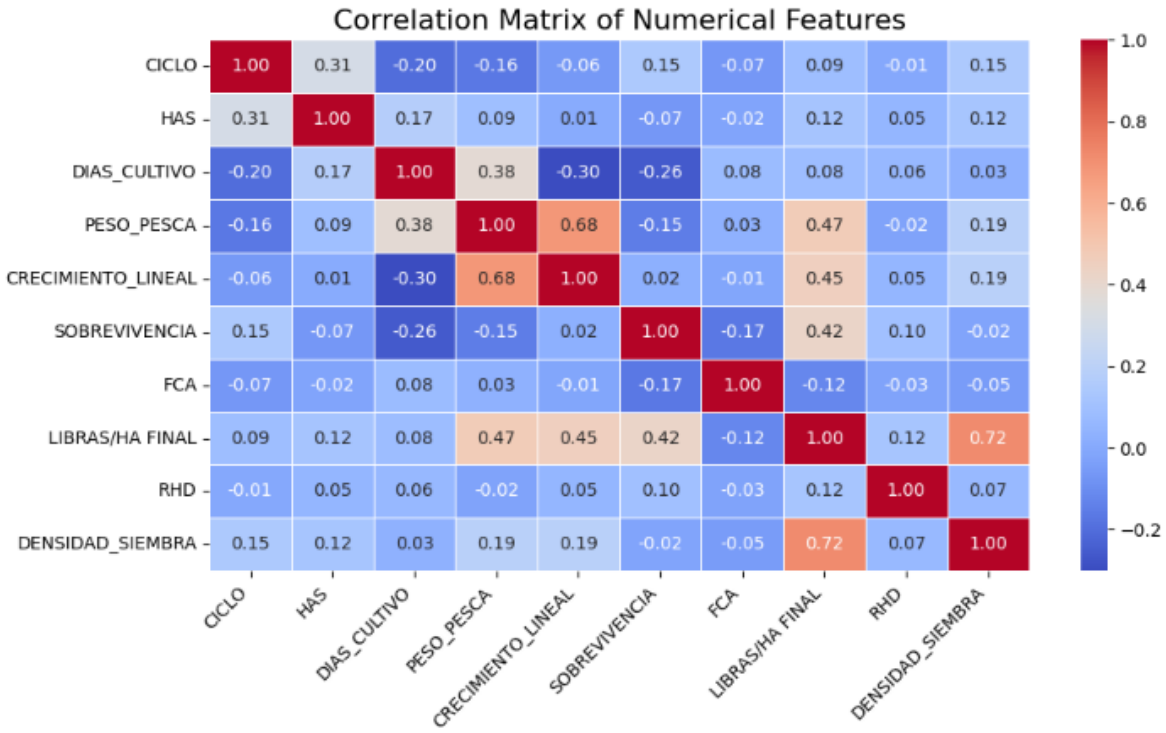
Correlación de variables

A continuación, se realizó un análisis de correlación de Person para identificar relaciones lineales entre variables numéricas, adicionalmente se realizó el análisis no lineal de factores influyentes mediante el método Disminución media de impurezas (Mean Decrease Impurity, MDI) y el método Permutación de Importancia (Permutation Importance, PI).

Correlación de Pearson

La estructura de la matriz sugiere que no hay multicolinealidad entre la mayoría de variables independientes (valor >0.70). Solo "DENSIDAD SIEMBRA" presenta una correlación muy fuerte y positiva con "LIBRAS/HA FINAL" (valor = 0.72), sugiriendo que prácticamente representan el mismo fenómeno y su inclusión simultánea en modelos puede causar redundancia (colinealidad). El resto de variables muestra correlaciones lineales nulas o muy débiles con la variable objetivo; esto implica que, por sí solas, no explican gran proporción de la variación de "DENSIDAD_SIEMBRA".

Imagen 2. *Matrix de correlación*.



Nota: La imagen 2 muestra la Matrix de coeficientes de correlación lineal de Pearson entre variables numéricas.

Cálculo importancia de variable MDI (Índice Gini)

Se entrenaron modelos predictivos como XGBoost, calculando la importancia relativa de cada variable en la predicción del rendimiento, lo que permite jerarquizar los factores más influyentes (por ejemplo, densidad de siembra, peso de pesca y supervivencia).

DENSIDAD_SIEMBRA es, por amplio margen, la variable de mayor influencia (0.612), demostrando que la cantidad con la que se siembra determina fundamentalmente la productividad en cosecha. Intervenciones y optimizaciones deben empezar por este factor.

SOBREVIVENCIA (0.315) y PESO_PESCA (0.271) ocupan el segundo y tercer lugar en importancia. Mantener elevadas tasas de supervivencia y buscar mayor peso de los animales cosechados son estrategias críticas para maximizar resultados.

El siguiente grupo, con menos peso, pero aun relevante, lo constituyen CRECIMIENTO_LINEAL (0.047), RHD (0.033), y variables como FCA, DIAS_CULTIVO, PISCINA_ENC, CICLO, y HAS (0.006 - 0.009), que muestran una contribución secundaria, pero pueden ser optimizados para ajustes finos del sistema productivo.

Las demás variables (EPOCA_CAT, MES_SIEMBRA_ENC, MES_COSECHA_ENC, ID_CAMARONERA_ENC, SALINIDAD_ENC) presentan importancias muy bajas (< 0.005), por lo que su influencia en la predicción del rendimiento es mínima y no justifican esfuerzos prioritarios para intervención directa.

Imagen 3. Ranking de variables. Elaborado por los autores.

Variable	MDI	PI	Ranking_Agregado
DENSIDAD_SIEMBRA	0.456271	0.775403	0.615837
SOBREVIVENCIA	0.174942	0.442302	0.308622
PESO_PESCA	0.205989	0.332423	0.269206
CRECIMIENTO_LINEAL	0.072368	0.025358	0.048863
RHD	0.039066	0.024711	0.031888
FCA	0.006067	0.010380	0.008224
PISCINA_ENC	0.007838	0.008246	0.008042
DIAS_CULTIVO	0.009727	0.006073	0.007900
HAS	0.006403	0.006831	0.006617
CICLO	0.003879	0.009124	0.006502
ID_CAMARONERA_ENC	0.006965	0.000421	0.003693
MES_SIEMBRA_ENC	0.004316	0.002667	0.003491
MES_COSECHA_ENC	0.003743	0.002319	0.003031
SALINIDAD_ENC	0.002426	0.000080	0.001253

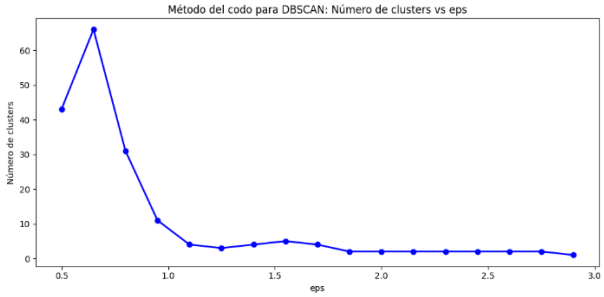
Nota: La imagen 3 muestra clara jerarquía de influencia y orienta la prioridad del manejo y decisión sobre el sistema productivo. Optimizar primero las variables top asegura el máximo retorno para el objetivo planteado.

Las demás variables son útiles para segmentar, ajustar o interpretar aspectos del proceso, pero no deben ser el foco principal de optimización cuando el objetivo es mejorar la cosecha total por hectárea.

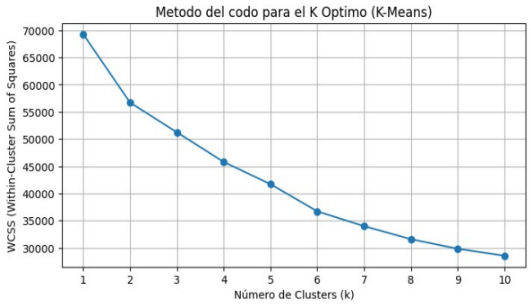
El enfoque basado en mediana entre MDI y PI otorga robustez y confianza al ranking, integrando tanto el efecto teórico del modelo (MDI) como el impacto real de la alteración de cada variable (PI).

Método del codo y análisis Silhouette

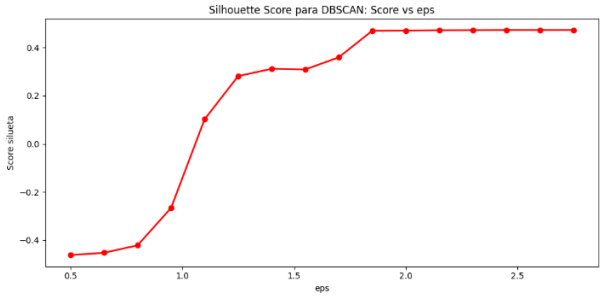
Imagen 4. Técnicas de evaluación de numero de cluster óptimo.



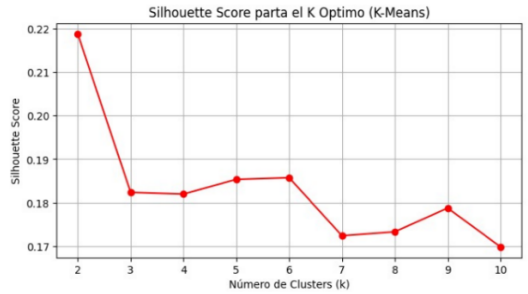
Técnica del codo aplicada a DBSCAN



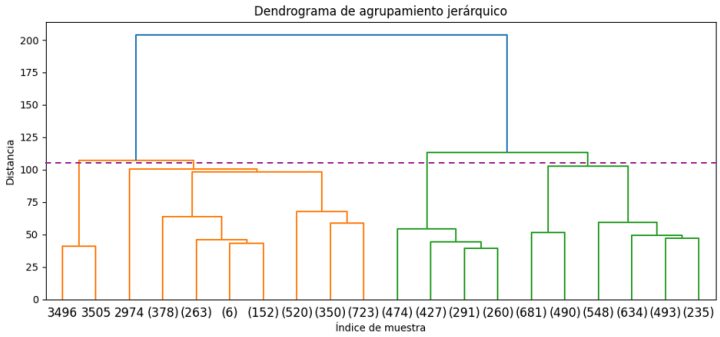
Técnica del codo para K-Means



Índice Silhouette aplicado a DBSCAN



Índice Silhouette aplicado a K-Means



Dendograma de Clusterización Aglomerativa.

Nota: La imagen 4 ilustra las técnicas de evaluación de numero de cluster óptimo para algoritmos No Supervisados de clusterizacion KMeans, DBSCAN, y Cluster Aglomerativo.

El análisis combinado del método del codo y el puntaje de silueta permitió determinar los parámetros óptimos para DBSCAN y K-Means. Para DBSCAN, el número de clústeres se estabiliza alrededor de eps = 1.1-1.3, evitando la segmentación excesiva y el ruido, y el puntaje de silueta sugiere un rango óptimo entre 1.3 y 1.7. En K-Means, el método del codo indica un óptimo en k = 3 o 4, mientras que el puntaje de silueta es máximo en k = 2 y alto hasta k = 6,

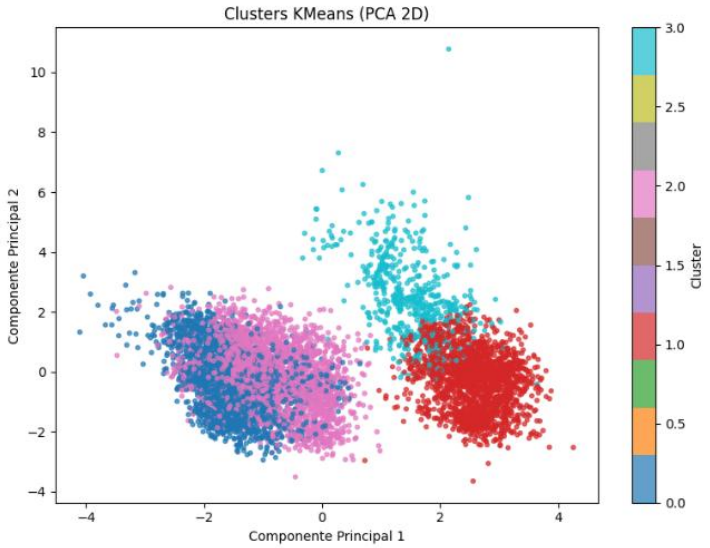
señalando que más de 3 a 5 clústeres no mejora significativamente la calidad. También se observó clusterización aglomerada identificándose el punto óptimo de corte en k = 4.

K-Means es más efectivo cuando los datos tienen agrupamientos de forma regular y centrados, mientras que DBSCAN destaca en identificar agrupaciones de forma irregular, eliminando ruido y detectando patrones conectados por densidad. En los resultados, DBSCAN logra mayor puntuación silueta, indicando clústeres mejor definidos. K-Means es útil cuando tienes hipótesis sobre el número de grupos, pero DBSCAN permite explorar sin esa restricción y es más robusto ante valores atípicos y estructuras complejas. De esta manera, se escogió k = 4 ya que es un valor intermedio generado con k-means; y es el valor que en la clusterización aglomerativa preserva la cohesión interna de los grupos y evita fusiones de subpoblaciones disímiles.

Clusterización

Para segmentar y agrupar los datos se aplicó el algoritmo de clustering K-Means por medio de PCA (Análisis de Componentes Principales), que permitió descubrir patrones y agrupamientos naturales en los datos; Este método facilitó interpretar qué factores impactan el rendimiento de manera robusta.

Imagen 5. *Clusters K-Means (PCA 2D). Elaborado por los autores*.



Nota: La imagen 5 muestra la visualización bidimensional (PCA 2D) de los clusters obtenidos por KMeans con k = 4.

Se observan cuatro grupos diferenciados de puntos, cada uno identificado por un color distinto en el esquema de clusters.

Los clusters principales muestran buena separación en el espacio proyectado, lo que indica que KMeans logró identificar regiones de alta densidad en el dataset. La gran mayoría de las observaciones están concentradas en torno a valores centrales de los componentes principales; sólo unos pocos puntos aparecen algo más dispersos o alejados del centro de los grupos. Hay fronteras suaves entre ciertos clusters, lo que puede indicar que existen algunas zonas de traslape o proximidad entre grupos, aunque la estructura general es claramente agrupada. No se observan valores extremos o ruido masivo que dificulte el agrupamiento.

Análisis y la caracterización de los 4 clusters obtenidos con KMeans

Cada cluster representa un perfil distinto de manejo y resultado: destaca el muy productivo (2), el de mayor duración (0), el de buen promedio y supervivencia (1), y el marginal o de baja producción (3). La segmentación permite entender la distribución y características de

los cultivos, facilitando la toma de decisiones para optimización, replicación o intervención en cada grupo.

Modelado y evaluación del modelo mediante XGBoost

Para el modelado con XGBoost se necesitó definir la variable objetivo (y) y la matriz de características (X), manejar las columnas no numéricas y categóricas en X, y luego dividir los datos en conjuntos de entrenamiento y prueba. Una vez que el modelo XGBoost fue entrenado, el siguiente paso lógico fue evaluar su rendimiento en el conjunto de prueba. Se usó el modelo entrenado para hacer predicciones sobre X_testy, luego se calculó el R cuadrado, el Error Cuadrático Medio (ECM)y el Error Cuadrático Medio (RMSE)respecto a los valores reales y_test.

Evaluación del rendimiento del modelo:

R-squared: 0.9603

Mean Squared Error (MSE): 733021.31

Root Mean Squared Error (RMSE): 856.17

El alto valor R cuadrado sugiere que las características elegidas explican una parte significativa de la varianza en «LIBRAS/HA FINAL», lo que convierte al modelo en un potente predictor para esta variable objetivo.

El siguiente paso lógico es realizar un análisis SHAP para interpretar el modelo XGBoost entrenado, comprender la importancia y la dirección del impacto de cada característica

Tabla 2. *Análisis y caracterización de Cluster*

Cluster	Densidad Siembra	Has	Días Cultivo	Peso Pesca	Crecimien-to Lineal	Sobrevivencia	Libras /Ha Final	Características Distintivas
0	Muy alta (285,740)	Intermedio (10.6).	Muy alto (120 días)	Alto (24.3)	Intermedio (1.39).	Intermedia (50.4%).	Alta (5,108).	Este cluster agrupa cultivos de alto tiempo de cultivo y alta densidad final, con peso y rendimiento por hectárea elevados. Lo diferencia su larga duración y productividad.
1	Intermedia-alta (238,156).	Bajo (9.08).	Intermedio-bajo (77.9).	Bajo (16.7).	Intermedio (1.45).	Alta (76.6%).	Intermedia (3,586).	Este cluster se caracteriza por cultivos de menor superficie, con buena supervivencia y rendimiento intermedio, pero menos tiempo de cultivo y menor peso respecto a cluster 0.
2	Máxima del conjunto (339,582).	Alta (19.0).	Intermedio (100.2).	Intermedio (19.9).	Alto (2.23).	Alta (66.9%).	Máxima (11,974).	Predomina la mayor densidad y rendimiento por hectárea, con buen crecimiento y supervivencia. Este cluster es el más productivo, con grandes superficies y buen desempeño general.

3	Muy baja (8,593).	Mínimo (1.0).	Mínimo (24.0).	Muy bajo (0.72).	Muy bajo (1.0).	Máxima (55.4%), aunque con pocos individuos.	Mínimo (161).	Este cluster agrupa lotes marginales o de baja productividad, con cultivos cortos, superficies pequeñas, rendimiento y peso bajos, y escasa densidad final.

Nota: La Tabla 2 detalla las Características Distintivas de los cuatro clusters identificados en el estudio.

Análisis SHAP

DENSIDAD_SIEMBRA es la variable más influyente. Los valores altos (en rosa) y bajos (en azul) están dispersos a lo largo de todo el eje de impacto, lo que significa que tanto valores altos como bajos pueden ejercer un impacto notable en la predicción. Los valores positivos de SHAP indican que mayor densidad aumenta la predicción de

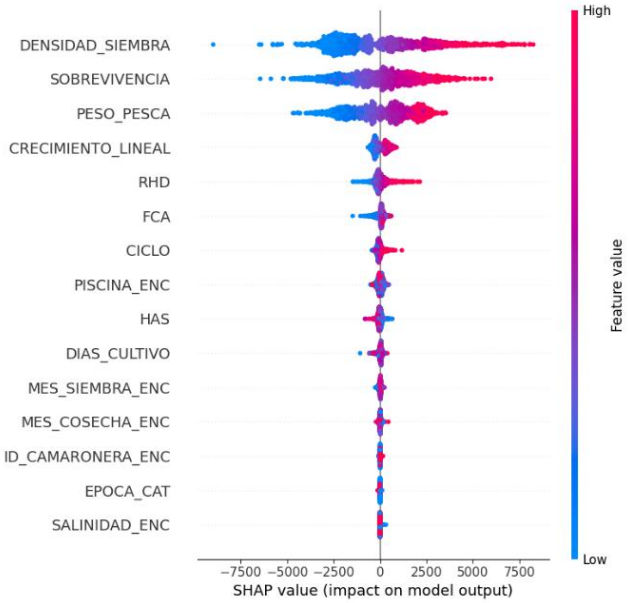
LIBRAS/HA FINAL, pero también es posible observar outliers: algunas densidades excesivas pueden tener efectos negativos (puntos más a la izquierda).

SOBREVIVENCIA y PESO_PESCA tienen también alto impacto. Valores altos tienden a incrementar la predicción, lo que implica que mejorar supervivencia y peso de captura eleva el rendimiento esperado.

CRECIMIENTO_LINEAL y RHD muestran un efecto moderado y concentrado, con la mayoría de sus SHAP values cerca de cero, lo que indica menos variabilidad, pero algún potencial efecto puntual.

El resto de variables (FCA, CICLO, PISCINA_ENC, etc.) tienen burbujas muy compactas centradas en cero, lo que refleja que modificar sus valores rara vez cambia de forma substancial la predicción del modelo.

Imagen 6. *Análisis Shap. Elaborado por los autores*.

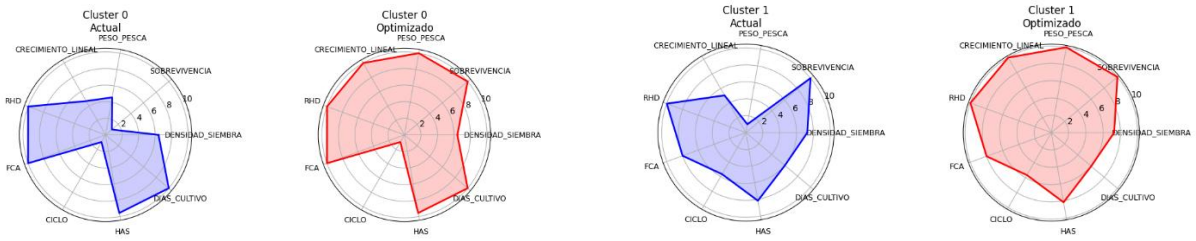


Nota: La imagen 6 muestra un gráfico resumen SHAP aplicado a un modelo predictivo en la industria del camarón blanco. En este gráfico se visualiza la importancia relativa y el efecto de cada variable sobre el resultado del modelo:

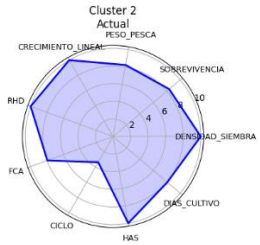
Optimización de variables

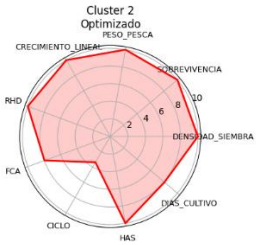
La optimización para cada cluster que permite maximizar el resultado de la variable LIBRAS/HA FINAL considerando que la meta que se busca es maximizar esa cifra de cosecha con la mejor DENSIDAD_SIEMBRA en combinación con las otras variables que influyen en ese objetivo.

Imagen 7. *Optimización de cluster 0, 1, 2, 3. Elaborado por los autores*.

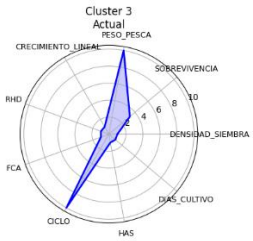


Optimización de cluster 0 ***Optimización de cluster 1***





Optimización de cluster 2



Optimización de cluster 3

Nota: La imagen 7 muestra las propuestas de optimización específicas para maximizar LIBRAS/HA FINAL en cada cluster, considerando la mejor combinación posible de DENSIDAD_SIEMBRA y demás variables influyentes en el resultado:

Cluster 0 - Productividad alta con densidad y tiempo de cultivo elevados

Optimización recomendada:

Mantener densidad relativamente alta, pero optimizar el tiempo de cultivo para encontrar el punto donde el crecimiento y el peso compensen la densidad sin saturar el sistema.

Mejorar el manejo nutricional y la calidad de agua para potenciar el crecimiento y peso de pesca incluso con altas densidades.

Implementar sistemas de monitoreo para identificar el momento óptimo de cosecha, buscando maximizar peso sin incrementar mortalidad ni disminuir rendimiento.

Cluster 1 - Supervivencia alta

Optimización recomendada:

Potenciar la supervivencia usando estrategias de bioseguridad y control ambiental, lo que permite elevar la densidad de siembra ligeramente y elevar LIBRAS/HA FINAL.

Ajustar la duración del ciclo y alimentación para incrementar peso de pesca y crecimiento lineal, logrando que una densidad intermedia produzca el mayor rendimiento individual y por superficie.

Analizar el impacto de incrementar tiempo de cultivo, siempre y cuando no afecte negativamente la supervivencia.

Cluster 2 - Máxima densidad y rendimiento

Optimización recomendada:

Evaluar si una leve reducción de densidad puede ser compensada con mayor peso, supervivencia o crecimiento, mejorando la eficiencia de conversión y rendimiento por hectárea.

Mantener prácticas avanzadas de manejo integral (aireación, suplementación nutricional, monitoreo frecuente) para sostener alto rendimiento con la mejor densidad alcanzable.

Ajustar variables ambientales (calidad de agua) y tecnológicas para soportar altas densidades sin pérdida de productividad ni aumento de mortalidad.

Cluster 3 - Muy baja densidad y productividad

Optimización recomendada:

Aumentar progresivamente la densidad, bajo estricto monitoreo, para detectar el punto donde la mayor población incrementa LIBRAS/HA FINAL sin comprometer supervivencia ni calidad.

Fortalecer prácticas de manejo (mejor alimentación, control sanitario, calidad de agua) para elevar crecimiento y rendimiento.

Introducir mejoras tecnológicas básicas y capacitación para incrementar peso individual y maximizar cosecha.

Estrategias transversales para todos los clusters Integrar análisis predictivo que combine DENSIDAD_SIEMBRA, supervivencia, crecimiento y tiempo de cultivo para identificar el óptimo adaptado a cada grupo.

Monitorear sistemáticamente los indicadores clave y hacer ajustes dinámicos de densidad, alimentación y ambiente para cada cluster.

Priorizar la gestión de calidad y salud en los sistemas para que un aumento (o reducción) de densidad no deteriore el rendimiento final.

Estas propuestas apoyan la búsqueda del máximo rendimiento con la mejor combinación de densidad y variables ambientales/operativas en función del perfil de cada grupo, permitiendo maximizar el objetivo de LIBRAS/HA FINAL

Discusión

El presente estudio aporta una contribución significativa en la optimización de la densidad de siembra en el cultivo de Penaeus Vannamei mediante el uso combinado de técnicas de Aprendizaje Automático No Supervisado y Análisis Estadístico. El modelo K-Means con cuatro clústeres permitió segmentar el sistema productivo en grupos con características diferenciadas, facilitando la identificación y el diseño de estrategias específicas para maximizar el rendimiento.

Esta aproximación basada en inteligencia artificial se alinea con los planteamientos de Arévalo et al. (2022), Bakry et al. (2023) y Pandey (2022), quienes destacaron la eficacia y versatilidad de los métodos no supervisados, como clustering y reducción de dimensionalidad, para identificar patrones complejos en grandes volúmenes de datos, particularmente en ámbitos

financieros y de análisis de delitos. De manera análoga, Hussein et al. (2021) demostraron que la combinación de estos algoritmos permite evaluar sistemas complejos en otras industrias, respaldo metodológico que coincide con la robustez del enfoque utilizado en este estudio para segmentar y optimizar variables ambientales y operativas en acuicultura.

Además, la integración del modelo predictivo XGBoost con el análisis interpretativo SHAP para jerarquizar variables críticas, como densidad, peso y supervivencia, refuerza la relevancia de conectar técnicas avanzadas de modelado con el entendimiento biológico-productivo. Esta metodología, coherente con los trabajos de Li et. al (2020), quienes optimizaron sistemas mineros integrando Modelos No Supervisados, y Saputra et al. (2025), quienes aplicaron aprendizaje profundo multimodal para segmentar áreas con precisión, pone de manifiesto el potencial de la Inteligencia Artificial para mejorar la gestión sostenible y rentable en sectores naturales. En la línea de Shiraj et al. (2024), quienes aplicaron DBSCAN para detectar anomalías en series temporales financieras, el presente estudio también evidencia la utilidad de métodos robustos frente a datos heterogéneos y ruido. Así, el modelo propuesto sienta bases para ampliar el estudio de variables productivas y ambientales en acuicultura, contribuyendo con una base sólida para futuras investigaciones que integren análisis predictivos y segmentación inteligente.

Conclusiones

El modelo KMeans con 4 clusters logró segmentar los datos en cuatro grupos relativamente homogéneos y compactos. La separación visual entre clusters sugiere que la estructura interna de los datos permite distinguir subpoblaciones o patrones diferenciados. Puede considerarse que la elección de k = 4 fue adecuada para este conjunto, dado que no hay grupos muy pequeños ni predominancia absoluta de un solo cluster. Este resultado facilita la posterior interpretación y análisis de las características y diferencias entre los clústeres formados, siendo útil para tareas de segmentación, análisis descriptivo o modelización basada en grupos. En

síntesis, el gráfico respalda la decisión de segmentar los datos en 4 clusters y muestra que KMeans consigue una partición coherente y utilizable para fines analíticos.

El conjunto de datos se preparó correctamente para el modelado, con 5542 muestras y 15 características utilizadas para el entrenamiento, y 1386 muestras y 15 características para la prueba. Las características categóricas se convirtieron en códigos numéricos. Se entrenó un modelo regresor XGBoost utilizando los datos preparados. El modelo mostró un buen rendimiento con un valor R cuadrado de 0,9603. El error cuadrático medio (MSE) del modelo se calculó en 733 021,31, y el error cuadrático medio (RMSE) fue de 856,17, lo que indica un error de predicción bajo en los valores predichos para la variable objetivo “Libras por Hectárea al Final” del cultivo.

Referencias bibliográficas

Arévalo, Barucca, Téllez-León, Rodríguez, Gage, & Morales. (2022). Identifying clusters of anomalous payments in the salvadorian payment system. Latin American Journal of Central Banking, 3(1), 2-12. doi: https://doi.org/10.1016/j.latcb.2022.100050

Bakry, Alsharkawy, Farag, & Raslan. (2023). Combating Financial Crimes with Unsupervised Learning Techniques: Clustering and Dimensionality Reduction for Anti-Money Laundering. Al-Azhar Bulletin of Science., 35(1), 2-12. https://www.researchgate.net/publication/378084632

Hussein, Stewart, Sacrey, Wu, & Athale. (2021). Unsupervised machine learning using 3D seismic data applied to reservoir evaluation and rock type identification. Interpretation, 9(2), 2-12. doi: https://doi.org/10.1190/INT-2020-0108.1

Jácome, & Flores. (2022). Identificación de Clusters Espaciales de Empresas y la Influencia de Factores Externos en su Constitución. Revista Politécnica, 3(1), 2-12. doi: https://doi.org/10.33333/rp.vol50n3.0

Jang, H.-J., Kim, B., Kim, J., & Jung, S.-Y. (2018). An Efficient Grid-Based K-Prototypes Algorithm for Sustainable Decision-Making on Spatial Objects. Sustainability, 10(8), 2614. https://doi.org/10.3390/su10082614

Pandey. A. (2022). Machine Learning. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 11(8), 864-869. doi: https://doi.org/10.22214/ijraset.2022.44376

Li, S., Sari, Y.A. & Kumral, M. (2020) Optimization of Mining–Mineral Processing Integration Using Unsupervised Machine Learning Algorithms. Natural Resources Research 29, 3035–3046. doi: https://doi.org/10.1007/s11053-020-09628-0

Rameshbabu, V., Vijayakumaran, C., & Edwin Prabhakar, P.B. (2023). Machine Learning. Character Lab Tips. doi: https://doi.org/10.53776/tips-gratitude-machine-learning

Saputra, M. R. U., Bhaswara, I. D., Nasution, B. I., Ern, M. A. L., Husna, N. L. R., Witra, T., Feliren, V., Owen, J. R., Kemp, D., & Lechner, A. M. (2025). Multi-modal deep learning approaches to semantic segmentation of mining footprints with multispectral satellite imagery. Remote Sensing of Environment, 318, Article 114584. https://doi.org/10.1016/j.rse.2024.114584.

Shiraj, M. M. B., Rahman, M. M., Al-Imran, M., Liza, M. Z. A., Murshed, M. M., & Akhter, N. (2024). Anomaly detection in financial time series data via mapper algorithm and DBSCAN clustering. World Journal of Advanced Engineering Technology and Sciences, 13(1), 70-84. doi: https://doi.org/10.30574/wjaets.2024.13.1.0396

Souza, A. C. A., & Silva, P. R. N. da. (2024). APLICAÇÃO DA INTELIGÊNCIA ARTIFICIAL NA ENGENHARIA DE CONFIABILIDADE E MANUTENÇÃO PREDITIVA: UM ESTUDO DE CASO NA INDÚSTRIA DE MINERAÇÃO. Revista Ibero-Americana de Humanidades, Ciências E Educação, 10(10), 3646–3659. https://doi.org/10.51891/rease.v10i10.16252

Torres Guerra, J. A., Mejía Cáceres, D., Moreyra Ramos, P., Oré Grados, J., & Oscco Barrientos, S. (2021). Geometalurgia y el futuro de la minería digital en el Perú. Revista Del Instituto De investigación De La Facultad De Minas, Metalurgia Y Ciencias geográficas, 24(47), 163-179. https://doi.org/10.15381/iigeo.v24i47.20661

Wang Y, Feng Y, Xi C, Wang B, Tang B, Geng Y. (2024) Development of an Intelligent Coal Production and Operation Platform Based on a Real-Time Data Warehouse and AI Model. Energies.; 17(20):5205. https://doi.org/10.3390/en17205205

Woźniak, R. J. (2024). Unsupervised machine learning in financial anomaly detection: clustering algorithms vs. dedicated methods. Przegląd Teleinformatyczny, 11(1), 29–46. doi: https://doi.org/10.5604/01.3001.0054.8748

Zhao, S., Zhang, S., Liu, J., Wang, H., Zhu, J., Li, D., & Zhao, R. (2021). Application of machine learning in intelligent fish aquaculture: A review. ELSEVIER, 1. https://doi.org/10.1016/j.aquaculture.2021.736724