

Aplicación de aprendizaje no supervisado en agricultura de cultivos sustitutos en diferentes regiones de Colombia

- **Resumen:**

Colombia posee una diversa variedad de zonas agrícolas, y ecológicas. Este proyecto propone la aplicación de múltiples modelos de aprendizaje no supervisado para generar recomendaciones sobre la viabilidad de reemplazar ciertos cultivos sin desmejorar su rendimiento. Concretamente, se identificaron los municipios idóneos para el cultivo, donde se pueda mantener o mejorar la producción de toneladas por hectárea cultivada. El enfoque metodológico que hemos implementado consta de los siguientes pasos: adquisición de datos [1], selección de algoritmos, entrenamiento del modelo, evaluación y análisis, interpretación de los resultados, aplicaciones prácticas y conclusiones.

Para reducir la dimensionalidad de nuestros datos se aplicó el Análisis de Componentes Principales (PCA). Se han implementado dos algoritmos de clustering, a saber, K-Means X-Means y DBSCAN, para identificar agrupamientos con alta similitud. Adicionalmente, se llevaron a cabo técnicas de visualización de datos geográficos.

La culminación de estos procesos nos permitirá generar recomendaciones fundamentadas para la toma de decisiones en cuanto a que cultivos podrían reemplazar el cultivo actual, y para nuevos proyectos basados en productos específicos nos ayuda a determinar ubicaciones más adecuadas, basados en el resultado siembras de especies con características similares.

- **Introducción:**

La sustitución de cultivos en Colombia representa un proceso estratégico y fundamental para la optimización de la agricultura en el país. Este enfoque se centra en la transición de una variedad de cultivos hacia alternativas más productivas y sostenibles. La amplitud de esta iniciativa abarca diversos productos agrícolas, tales como la acelga, el mango, el maíz, el tomate, la naranja, y numerosos otros. La importancia de este proceso radica en varios aspectos cruciales:

Diversificación Agrícola: La sustitución de cultivos impulsa la diversificación de la producción agrícola, enriqueciendo la disponibilidad de alimentos y productos en el ámbito local y regional [6]. Mejora de la Rentabilidad: Al reemplazar cultivos menos productivos o de baja demanda con alternativas más lucrativas, los agricultores tienen la capacidad de aumentar sus ingresos y mejorar su calidad de vida. Promoción de la Sostenibilidad: Este enfoque fomenta prácticas agrícolas sostenibles, lo que conlleva a una disminución en la presión ejercida sobre los recursos naturales y favorece la conservación del medio ambiente [5]. Seguridad Alimentaria: La diversificación agrícola contribuye a la seguridad alimentaria, al acceder a muchos alimentos frescos y nutritivos [5]. En síntesis, la sustitución de cultivos en Colombia desempeña un rol esencial en la mejora de la agricultura al promover la diversificación, la rentabilidad y la sostenibilidad. La amplia variedad de productos agrícolas incluidos en este proyecto brinda numerosas alternativas de sustitución.

Una de las principales limitantes para mejorar la seguridad alimentaria y mejorar el desempeño de los pequeños productores agrícolas es la falta de decisiones informadas debido a factores como falta de información a nivel local, estructura institucional y a políticas de estado [7]. Además, el uso de algoritmos de aprendizaje no supervisado permite identificar patrones en los datos sin requerir etiquetas previamente establecidas, lo que permite generar recomendaciones óptimas de sustitución sin afectar negativamente el rendimiento de los cultivos. Es importante mencionar que existen estudios agroclimáticos para departamentos en

Colombia que identifican las zonas aptas dentro los municipios para ser dedicadas a la agricultura como el trabajo de Araujo-Carrillo enfocado en la Alta Guajira [10], y referencias a la agricultura de precisión las podemos encontrar en trabajos como el de Arrubla-Hoyos en la aplicación de sistemas de sensores en la agricultura [11]. Es importante que el aprendizaje sobre producción agricultura en Colombia llegue a todos, para que los agricultores puedan tomar decisiones informadas.

- **Materiales y Métodos:**

Área de Estudio:

El ejercicio se lleva a cabo en Colombia, la cual está localizada en la parte noroeste de Sudamérica, este país colinda al norte con el mar Caribe y con el océano Pacífico al este, con un área de 1.1 millones de km². Colombia tiene una población de alrededor de 48.3 millones de habitantes (ver <http://www.dane.gov.co>), la altitud en este país oscila desde el nivel del mar hasta más de 5,000 m de altura sobre el nivel del mar en sus cumbres andinas o de la sierra nevada de Santa Marta, las diferencias en elevación y latitud son la causa de los diferentes climas, un ejemplo de ello es la comparación de precipitación anual de 350 mm en península de la Guajira a los 12,000 mm en las zonas bajas de la costa del pacífico, [8]. En el Informe Nacional sobre el estado de la Biodiversidad en Colombia 1997 del Instituto de Investigación de Recursos Biológicos Alexander von Humboldt se identifican 26 ecoregiones y 63 ecosistemas [9]. Por otro lado, en la investigación conducida por Sánchez-Cuervo [8] se redujeron las ecoregiones a 13 debido a que algunas ecoregiones era representadas por solo un poco de municipalidades, un ejemplo que nos muestran es que el municipio de Tumaco comprendía toda una ecorregión, la cual fue agregada a una región más colindante más extensa.

Datos y Fuentes de Datos:

La integración de las variables fue crucial para la zonificación de áreas agrícolas similares (clústeres). Este ejercicio usa un conjunto de variables relevantes (Tabla 1) para desarrollar las áreas de producción agrícolas similares.

En los datos se identificó que en Colombia se cultivan 223 especies diferentes, que se clasifican en 13 Grupos de cultivos que son: Hortalizas, Plantas Aromáticas, Condimentarias Y Medicinales, Tubérculos Y Plátanos, Frutales, Oleaginosas, Leguminosas, Fibras, Flores Y Follajes, Cereales, Otros Permanentes, Forestales, Hongos, Otros Transitorio.

En los Datos se cuenta con 32 departamentos, 1105 Municipios, 120 Subgrupos de Cultivos, 13 años de información sobre los cultivos para 36 periodos de siembra y la producción de toneladas registrada en cada siembra.

Tabla 1. Variables utilizadas para crear clústeres de zonas de producción agrícola

Variable / Tipo	Resolución	Fuente
Evaluaciones Agropecuarias (cultivo, ciclos de cultivo, municipalidad, código de municipio)	Municipio	<u>Evaluaciones Agropecuarias Municipales (EVA) - Datos.gov.co</u>
Datos geográficos (polígonos, código de municipio)	Municipio	Kaggle: <u>Colombia Municipios en Kaggle</u>

El número de ciclos para cada cultivo fue evaluado para cada municipalidad, los datos fueron evaluados con y sin estandarización. Estas son las dos variables utilizadas para la creación de los clústeres.

El rendimiento de cultivo fue utilizado para crear recomendaciones basadas en el clúster al que pertenece el territorio y los productos con un mayor crecimiento en rendimiento.

Preprocesamiento de los datos:

Para que hacer comparable el impacto de las variables, fue necesario estandarizar los valores de ciclos que cada cultivo se sembró en cada municipio. En este ejercicio, el análisis de componentes principales se utilizará para reducir la dimensión y mantener las variables importantes.

Creación de clústeres para definir zonas de producción agrícola:

El ejercicio de creación de clústeres en unidades similares donde se siembran cultivos parecidos, por lo que al relacionarse los municipios en regiones podrían introducirse nuevos cultivos al municipio al identificar su pertenencia a una zona agrícola en la que existen estos cultivos. La creación de clústeres es la clasificación de datos la cual se basa en la idea que estos datos comparten grupos si son similares. El objetivo es agrupar municipalidades donde se maximice la similitud en cultivos, al mismo tiempo maximizando la diferencia entre grupos diferentes.

Uno de los algoritmos más comunes utilizados para crear clústeres espaciales es K-means, el cual es simple de implementar y se puede trabajar con grandes volúmenes de información. Sin embargo, tiene la desventaja de que el usuario debe predefinir el número de clústeres a formar. Para ello se analizó el máximo número de clústeres utilizando el método del codo donde se graficó la varianza intra-clúster, donde el objetivo es encontrar el punto en el cual la mejora deja de ser notable; de la misma manera se evaluó el coeficiente de Silhouette, el cual identifica la calidad de los clústeres, y dicha calidad se espera maximizar. Adicionalmente se utilizó el método de creación de clústeres X-means, el cual se ha desarrollado como una extensión de K-means para brindar una clasificación más independiente con una mejor eficiencia computacional. Utiliza información como (BIC) Criterio de Información Bayesiana para seleccionar un modelo sobre el otro y si una partición del centroide se debe mantener o no.

- **Resultados y Discusión:**

Los resultados muestran que 179 dimensiones (componentes) explican más del 95% de la varianza acumulada. Como resultado, 179 componentes fueron seleccionados para continuar con el análisis de clústeres.

Número óptimo de clústeres:

Se utilizó el coeficiente de Silhouette para determinar el número óptimo de clústeres, donde los valores cercanos a +1 indican que la muestra se encuentra lejos del clúster vecino, un valor de 0 indica que la muestra se encuentra muy cercana a la frontera de decisión, y un valor negativo nos indicaría que las muestras han sido asignadas al clúster incorrecto. En base a los picos observados en la gráfica de Silhouette se muestran 2 ejemplos de clústeres (con 6 y 10 clases) los cuales pueden ser distinguidos en este ejercicio (Figura 1).

Adicionalmente, se empleó X-means para clasificar las regiones de producción agrícola en Colombia. El algoritmo identificó 13 clústeres para el cual utilizamos los 179 componentes de PCA. 5 de los clústeres solo contienen un municipio, los cuales podríamos considerar valores atípicos, y tendríamos solo 8 clústeres que se encuentran agrupando regiones. Este método de clasificación no requiere predefinir el número de clústeres, ya que minimiza el WSS sin la necesidad de producir agrupaciones redundantes. La varianza intra-clúster resultó menor con los 8 clústeres efectivos identificados por X-means. Estos clústeres o unidades se pueden considerar dominios a los que se les puede hacer recomendaciones específicas de cultivos.

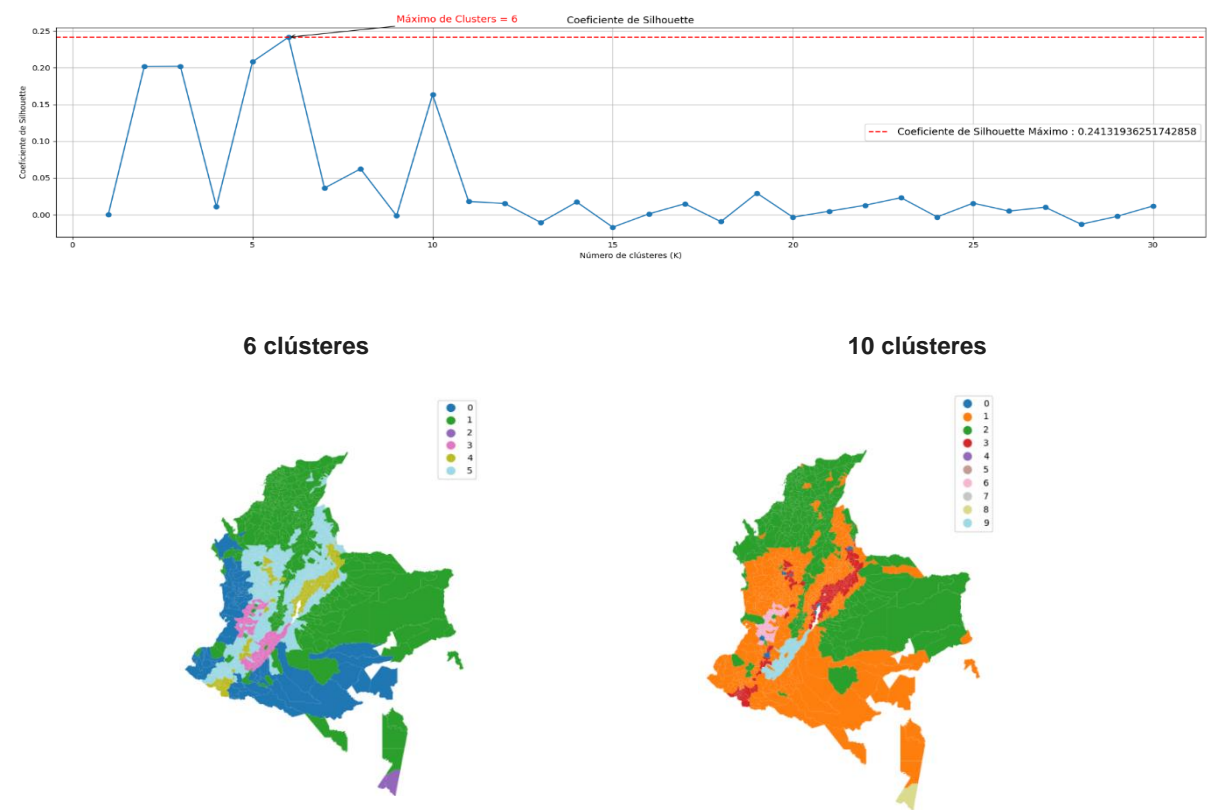


Figura 1. Número de clústeres utilizando el método de coeficiente de Silhouette en la creación de clústeres y el algoritmo K-means.

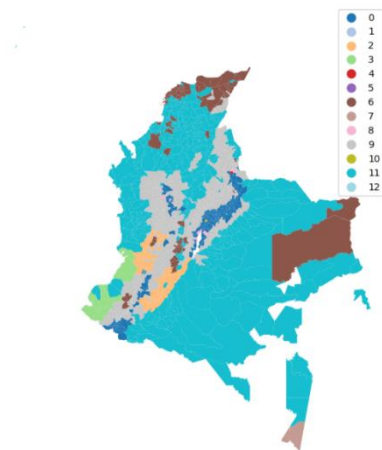


Figura 2. Representación clústeres basado en la distribución espacial óptima basado en el algoritmo de creación de clústeres X-means.

- **Conclusiones:**

Los resultados del presente ejercicio deben ser analizados por expertos antes de introducir un cultivo nuevo en un municipio ya que se necesita un profundo entendimiento de los factores biofísicos del campo, pero pueden ser de gran aporte para agricultores que estén en busca de sustituir cultivos, dicha búsqueda puede responder a diversas condiciones como lo poco atractivo de los precios del cultivo actual, suficiencia de agua (la cual cambia año con año).

El objetivo del ejercicio fue utilizar la información abierta disponible, y se decidió agrupar los municipios de acuerdo con los cultivos producidos y contabilizar el número de ciclos de siembra registrados para cada cultivo. La premisa es que, si en los municipios se cultivan productos similares, se pueden introducir nuevos cultivos ya cultivados en otros municipios del mismo clúster.

Los siguientes pasos deben enfocarse en la validación de las zonas de producción agrícola identificadas y ajustar su aplicabilidad. Adicionalmente incluir información de geomorfología, temperaturas medias, radiación, precipitación, tipo de riego, disponibilidad de agua, tecnologías de siembra. La data de todos los factores anteriores resultará en analíticas mejoradas y producirá mejor zonificación, ya que estas son determinantes para el éxito de las siembras.

Anexo: [Repositorio Github - MIAD ANS AEZ](#)

- **Bibliografía:**

- [1] Datos oficiales del gobierno colombiano sobre Evaluaciones Agropecuarias Municipales (EVA): [Evaluaciones Agropecuarias Municipales \(EVA\) - Datos.gov.co](#)
- [2] Conjunto de datos y análisis relacionados con los municipios de Colombia disponible en Kaggle: [Colombia Municipios en Kaggle](#)
- [3] Datos sobre la relación entre el área cosechada por cultivo y el rendimiento en Colombia: [Área Cosechada por Cultivo Vs. Rendimiento - Datos.gov.co](#)
- [4] Información del gobierno colombiano sobre los cultivos con mayor potencial en Colombia: [Cultivos con Mayor Potencial en Colombia - agronet.gov.co](#)
- [5] Noticia sobre la importancia de la agricultura sostenible en el bienestar de las sociedades, proporcionada por la Alcaldía de Medellín: [La Agricultura Sostenible - Alcaldía de Medellín](#)
- [6] Recurso informativo sobre la diversificación de operaciones agrícolas: [Diversification of Your Operation, Why](#)
- [7] Tamene, L., Abera, W., Bendito, E., Erkossa, T., Tariku, M., Sewnet, H., Tesfaye, K. (2022). Data-driven similar response units for agricultural technology targeting: An example from Ethiopia. *Experimental Agriculture*, 58, E27. doi:10.1017/S0014479722000126
- [8] Sánchez-Cuervo, A. M., Aide, T. M., Clark, M. L., & Etter, A. (2012). Land Cover Change in Colombia: Surprising Forest Recovery Trends between 2001 and 2010. *PLoS ONE*, 7(8). <https://doi.org/10.1371/journal.pone.0043943>
- [9] Chaves ME, Arango N, editors (1998) Informe Nacional sobre el estado de la Biodiversidad en Colombia 1997. Bogotá, DC, Colombia: Instituto de Investigación de Recursos Biológicos Alexander von Humboldt.
- [10] ARAUJO-CARRILLO, GUSTAVO ALFONSO, MARTÍNEZ-MALDONADO, FABIO ERNESTO, DEANTONIO-FLORIDO, LEIDY YIBETH, & GÓMEZ-LATORRE, DOUGLAS ANDRÉS. (2019). Agroclimatic zoning: a planning strategy for agricultural and livestock systems in Alta Guajira, Colombia. *Revista Colombiana de Ciencias Hortícolas*, 13(3), 435-447. Epub August 08, 2020. <https://doi.org/10.17584/rcch.2019v13i3.9648>
- [11] Arrubla-Hoyos W, Ojeda-Beltrán A, Solano-Barliza A, Rambauth-Ibarra G, Barrios-Ulloa A, Cama-Pinto D, Arrabal-Campos FM, Martínez-Lao JA, Cama-Pinto A, Manzano-Agugliaro F. Precision Agriculture and Sensor Systems Applications in Colombia through 5G Networks. *Sensors*. 2022; 22(19):7295. <https://doi.org/10.3390/s22197295>