Grupo 8: Valentina Farkas Sánchez - Código 202226209, Oscar Andrés Patiño Patarroyo - Código 202223234, Andrés Sierra Urrego - Código 202120960, John Edinson Rodríguez Fajardo - Código 202226240

USO DE CLUSTERS AGRÍCOLAS EN LA FOCALIZACIÓN DE SUBSIDIOS E INVERSIONES PARA LA PRODUCCIÓN

1 Resumen

El proyecto que se plantea desarrollar se enfoca en abordar un desafío actual en el sector agrícola de Colombia. Este desafío tiene como principal reto la falta de una estrategia eficiente para dirigir subsidios e inversiones hacia las unidades productoras agrícolas que más lo necesitan en el país, es así como esto conlleva a un uso ineficiente de los recursos disponibles y dificulta potenciar el campo colombiano.

Debido a lo anterior, se plantea, como solución para la focalización y priorización de inversiones y subsidios, el desarrollo de un algoritmo de generación de clústeres que agrupe las unidades productivas agrícolas (UPA), con el objetivo de identificar y segmentar de manera precisa estos lugares con vocación agrícola en grupos con características similares, teniendo en cuenta factores como la ubicación geográfica, el tipo de cultivo, el área de la finca, el rendimiento y la producción. Al lograr esta agrupación, se espera tener zonas de interés agrícolas para diferentes fines como disminuir las importaciones o aumentar exportaciones de productos.

Con el proyecto se espera proporcionar una herramienta poderosa para impulsar el desarrollo sostenible del sector agrícola, asegurando que los recursos y los esfuerzos se concentren maximizando el impacto. Esto no solo beneficiará a los agricultores, sino que también tendrá un impacto positivo en la seguridad alimentaria, la economía nacional y la prosperidad de las comunidades rurales en todo el país.

2 Introducción

En el contexto actual del país y de acuerdo con lo trazado en el Plan Nacional de Desarrollo 2022-2026 – “Colombia potencia mundial de la vida” se plantean 5 transformaciones para Colombia. Entre estas se encuentra el derecho humano a la alimentación y la transformación productiva. Uno de los ejes de estas transformaciones es el sector agro colombiano. Debido a lo anterior, es de vital importancia para el país la asignación eficiente de los recursos del presupuesto nacional que se esperan invertir en las zonas productivas del país que más lo requieran. Este problema hace que se piense en soluciones innovadoras para agrupar la gran cantidad de fincas productoras del país para potenciar sus rendimientos de cultivos, de acuerdo con los intereses que puede llegar a tener la maximización de la producción agrícola en el país.

Es así como surge el proyecto de generación de clústeres, el cual tiene como objetivo agrupar estas UPA para poder focalizar los subsidios y las inversiones de interés para diferentes entidades nacionales y locales, así como organismos encargados de fomentar el agro colombiano. Por lo anterior, nuestro cliente potencial es el estado colombiano a través de su Ministerio de Agricultura y Desarrollo Rural y sus diferentes entidades adscritas, como la Agencia de Desarrollo Rural (ADR) y la Unidad de Planeación de Tierras Rurales, Adecuación de Tierras y Usos Agropecuarios (UPRA), encargadas de la planificación y ejecución de políticas del sector agrícola.

Este problema radica en la necesidad de optimizar el uso del presupuesto disponible para generar este cambio en el campo colombiano, para asegurar que los subsidios y las inversiones se dirijan hacia aquellas UPA que puedan generar un mayor impacto en la economía agrícola del país. Es así como para lograr lo anterior, recurrimos al campo del aprendizaje no supervisado, específicamente en la generación de clústeres para agrupar las UPA.

Este enfoque tiene como objetivo descubrir patrones y relaciones entre los datos de las UPA y de la información socioeconómica de interés de las diferentes regiones a analizar, esto lo realizaremos con esta metodología teniendo en cuenta que no hay una forma de agrupar definida para los diferentes objetivos de la política nacional. Estos grupos se crearán considerando diferentes variables, como la ubicación geográfica, el tipo de cultivo, el área de la finca, la distancia a ciudades principales, dimensiones socioeconómicas de las regiones y la proximidad a puertos marítimos.

3 Materiales y métodos

Se cuenta con una fuente principal obtenida del censo agropecuario del 2014 realizado por el DANE y 10 fuentes de apoyo o complementarias. Se puede acceder al resumen de cada una de las fuentes en el siguiente link: https://github.com/grupovajo/proyectoANS/blob/main/documentos/Resumen_fuentes.pdf, allí se presenta para todas las

fuentes el origen de los datos, la cantidad de registros, los nombres y tipos de las variables para la fuente principal y complementarias.

La fuente principal cuenta con 12 variables, 7 categóricas y 5 numéricas, con 1'048.576 registros. De las 5 variables numéricas 2 se relacionan con la geolocalización y los 3 restantes son variables relacionadas al área y cantidad de producción de los cultivos. En la siguiente tabla se presenta cada una de las variables con su definición y tipo de dato.

Tabla 1, Definición de variables

No	Campos	Definición	Tipo de Dato
1.	**OBJECTID**	Identificador único de la UPA	Categórico
2.	**x_geo**	Información de coordenadas geográficas	Numérico
3.	**y_geo**	Información de coordenadas geográficas	Numérico
4.	**p_s6p46**	Variable codificada que representa la información del tipo de cultivo o plantación forestal que está en el lote	Categórico
5.	**MPIO_CDPMP**	Código del municipio DANE	Categórico
6.	**cultivo**	Tipo de cultivo	Categórico
7.	**MPIO_cultivo**	código del municipio concatenado con el tipo de cultivo	Categórico
8.	**rend**	Rendimiento (t/ha)	Numérico
9.	**prod**	Producción obtenida (t)	Numérico
10.	**area_cos**	Área cosechada de los cultivos (ha)	Numérico
11.	**Periodicid**	Frecuencia con la que se realiza siembra del cultivo en el lote	Categórico
12.	**Cultivo_1**	Cultivo anterior	Categórico

*t: toneladas, *ha: hectáreas

Valores nulos y outliers en cada columna

En la revisión de los datos se pudo identificar valores nulos en la variable de "Periodicid" y "Cultivo_1", que corresponde al 0.1% del total de los datos y también se identificó outliers en las variables de interés de "rend", "prod" y "area_cos". La cantidad de outliers por cada variable ordenadas de mayor a menor es presentada en la figura 1a.

En relación con las estadísticas descriptivas (Tabla 2 a) se observa que:

1. Se puede observar que las 3 variables de interés no contienen valores nulos.
2. Las variables "rend" y "prod" tienen desviaciones estándar cercanas y mucho más alta que la desviación estándar de la variable "area_cos", indicando con esto una distribución de los datos más dispersa.
3. Para todos los casos se presenta una dispersión de la distribución de los datos evidenciándose una cantidad significativa outliers para cada una de las variables de interés.

En relación con la correlación entre variables de interés (Tabla 2 b), se evidencia una correlación positiva entre las variables "area_cos" y "prod". Y una correlación negativa entre las variables "area_cos" y "rend". Lo anterior muestra para la correlación positiva, la relación proporcional que, al tener una mayor área de cosecha, la producción total es mayor, sin embargo, la correlación negativa evidencia que para áreas grandes de cosecha el rendimiento es menor.

Tabla 2, Estadísticas descriptivas y correlación de las variables de interés

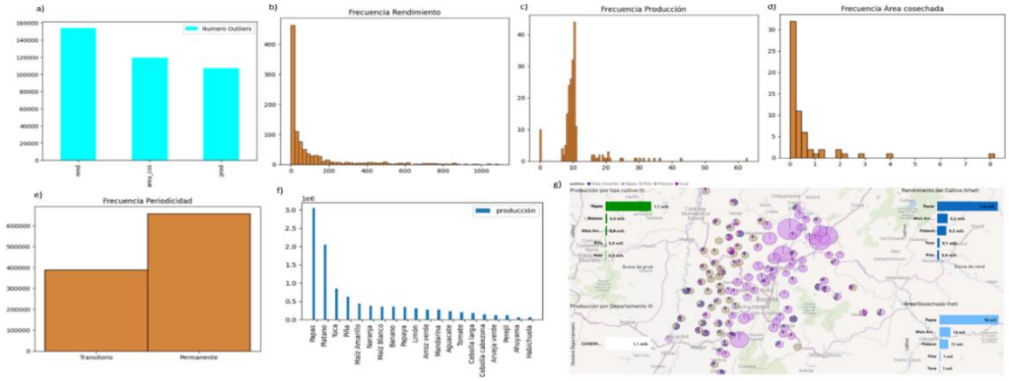
a)	count	mean	std	min	25%	50%	75%	max
rend	1048575.0	56.964931	154.937324	0.000003	2.188330	8.791615	37.789953	11420.76465
prod	1048575.0	10.829576	149.627018	0.000004	3.071679	5.060987	8.538944	86899.68240
area_cos	1048575.0	3.491945	48.083071	0.006200	0.143554	0.613336	2.112488	20000.00000

b)	rend	prod	area_cos
rend	1.000000	0.009471	-0.025247
prod	0.009471	1.000000	0.237365
area_cos	-0.025247	0.237365	1.000000

Histogramas

- Rend: Variable continua que representa las toneladas producidas por hectárea sembrada. Esta presenta una distribución con un fuerte sesgo hacia la derecha, lo cual indica que la mayoría de los cultivos produce una cantidad relativamente baja por hectárea (figura 1b).
- Prod: Variable continua que representa las toneladas producidas por los diferentes tipos de cultivo. Se puede evidenciar la dispersión de los datos y un pico de 10 que evidencia el valor más común de toneladas producidas por cultivo (figura 1c).
- Area_cos: Variable continua que representa el área cosechada en hectáreas de los diferentes tipos de cultivo. Esta presenta una distribución con un fuerte sesgo hacia la derecha, evidenciando que la mayoría de las cosechas de los cultivos se realizan en un área menor a una hectárea (figura 1d).
- Periodicid: Variable categórica que representa la frecuencia o periodicidad del cultivo en el área sembrada. Cómo se puede evidenciar en la gráfica de los 1'048.576 registros que representa todos los tipos de cultivos en las diferentes zonas del país, más de 600.000 corresponden a cultivos permanentes (figura 1e).

Figura 1, Gráfico de barras, diagrama de frecuencias y mapa de burbujas



En la figura 1f se presentan los 20 tipos de cultivo con mayor producción en millones de toneladas en dónde los cultivos de papa, plátano, yuca, piña y maíz amarillo representan el 62% de la producción total del país. Por último, para el departamento de Cundinamarca (el cual cuenta con la mayor producción), en la figura 1g se muestra un mapa de burbujas de las zonas con el comportamiento de producción para los 5 tipos de cultivo mencionados anteriormente en donde la producción de papa corresponde al 93% de la producción total del departamento.

Después de realizar el proceso de limpieza y análisis descriptivo de los datos, damos paso a la implementación del algoritmo que nos va a permitir determinar las zonas de interés agrícolas en el país. En el presente trabajo nos vamos a concentrar en el cultivo de la soya, puesto que es el producto agrícola que más importan los países, llegando a duplicar al siguiente producto que es el maíz, siendo este producto de potencial para las exportaciones futuras del país.por otro lado, se creó una variable que se puede encontrar en el notebook del proyecto y que es un cálculo del potencial de producción que tiene cada UPA.

Producto	Valor (miles USD)
Soya beans; other than seed, whether or not broken	$ 519,028,427.87
Cereals; maize (corn), other than seed	$ 199,739,407.98
Cereals; wheat and meslin, other than durum wheat, other than seed	$ 160,675,307.05
Vegetable oils; palm oil and its fractions, other than crude	$ 130,950,670.30

Como primer paso, escalamos la base e implementamos los modelos de Aprendizaje No Supervisado, en este proyecto utilizamos los siguientes modelos: K-means, K-medoids, DBSCAN y Algoritmo Jerárquico Aglomerativo. A continuación, se explica el procedimiento realizado.

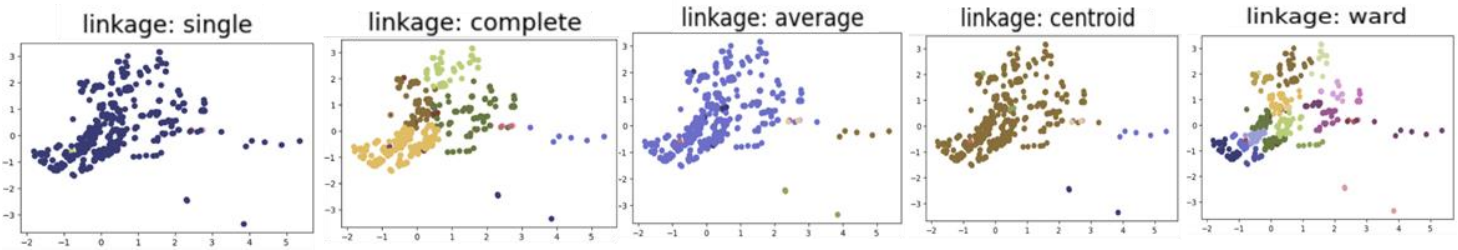
K-means: este algoritmo de clustering basado en centroides, nos permite agrupar las UPA en clústeres basados en simi litudes de las variables seleccionadas, en donde utilizamos la métrica del índice de Silhoutte para determinar el número óptimo de clústeres, en este caso se obtuvo que en un rango entre 2 y 10 el número óptimo de clústeres es de 5 con un c oeficiente de Silhouette de 0.6089.

K-medoides: esta es una alternativa al modelo K-means ya que es robusto a valores atípicos y utiliza medoides en lugar de centroides para representar los clústeres. En este algoritmo, los medoides son puntos reales del conjunto de datos, lo que significa que los centroides pueden ser puntos de datos reales, lo que lo hace más robusto en comparación con K-Means en presencia de valores atípicos o datos ruidosos. Implementamos el modelo con los 5 clusters mencionados anteriormente.

DBSCAN: es un algoritmo basado en la densidad de las observaciones y, a diferencia de K-medias, permite identificar datos atípicos. Además, determina clústeres de una manera más flexible sin necesidad de especificar la cantidad de clústeres. Primero, se calcula la distancia desde cada punto a su vecino más cercano implementando NearestNeighbors, se establece n_neighbors = 10, lo que significa que se buscarán los 10 vecinos más cercanos para cada punto del conjunto de datos, se utiliza KneeLocator para encontrar el codo de la curva, que es el punto en donde el gráfico cambia de pendiente e indica un valor apropiado para el parámetro k, obteniendo un resultado de k= 0.15897.

Después, implementamos el algoritmo de DBSCAN para diferentes valores de min_samples, con el fin de explorar cómo cambian los resultados a medida que se cambian los valores de min_samples. Por último, utilizando un valor de eps: 0.15897 y min_sample:10, graficamos los resultados, los cuales son presenta dos en la siguiente sección.

Algoritmo Jerárquico Aglomerativo: este algoritmo permite obtener una estructura de árbol de clústeres utilizando diferentes métodos de enlace, tales como single, complete, average, centroid y Ward. Generamos un bucle para comparar los diferentes métodos de enlace, se calcula la matriz de distancia mediante la métrica de distancia euclidiana, se utiliza la función fcluster para calcular los clústeres y visualizamos los resultados:



Los anteriores métodos de enlace producen diferentes estructuras de clústeres, se selecciona el método de Ward puesto que las agrupaciones son más homogéneas.

4 Resultados y discusión

A continuación, se presentan los resultados arrojados por los cuatro algoritmos de clusterización mencionados en el apartado anterior de manera gráfica para facilitar su interpretación:

Figura 5. Algoritmo K-medias.

Figura 6. Algoritmo K-medoides.

Figura 7. Algoritmo DBSCAN.

Figura 8. Algoritmo Jerárquico Aglomerativo.

Las visualizaciones anteriores permiten mostrar que el modelo de K-medoides, al parecer, es el que muestra una mejor homogenización en la agrupación de las ubicaciones de los municipios. Adicionalmente, vale la pena mencionar que los algoritmos de K-medias y Jerárquico Aglomerativo también muestran una buena distribución de los clústeres. Sin embargo, hay puntos de ciertos clústeres de dichos algoritmos los cuales parecieran estar agrupados en otros clústeres, por lo que se podría pensar que no están haciendo una buena asignación.

En la figura 9, para el algoritmo de K-medoides se presenta el recuento y la suma del potencial de producción para cada clúster. Se puede evidenciar que las agrupaciones con mayor potencial de producción son los clústeres 3 y 0, con valores de 65164 y 14539 respectivamente. Para el clúster 3 se evidencia el mayor número de unidades agrícolas con un valor de 504. También se observa en el diagrama de caja (figura 11) para el clúster 0 que la distribución entre el máximo esperado y la mediana tiene una dispersión mayor que los demás clústeres significando varias unidades productoras agrícolas con valores de potencial de producción relativamente altos, sin considerar los outliers.

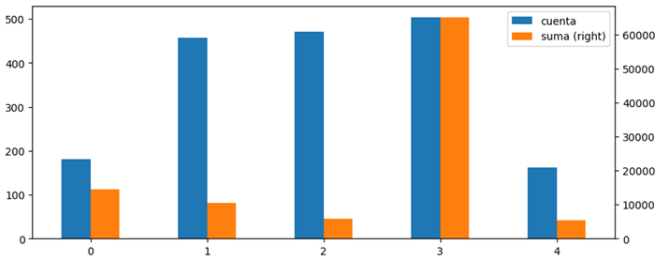


Figura 9. Recuento y suma del potencial de producción por clúster-K-medoides

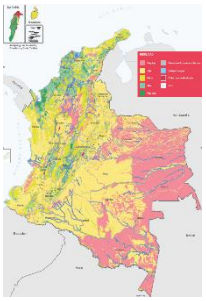


Figura 10. Mapa de fertilidad de los suelos en Colombia -fuente: IGAC 2022

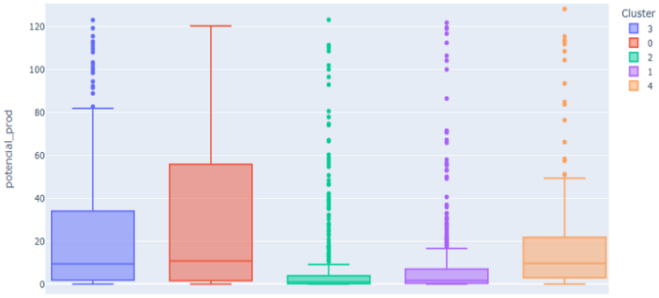


Figura 11. Diagrama de caja

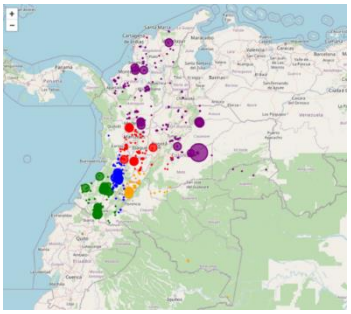


Figura 12. Mapa de burbujas

En la figura 9 también se puede observar para los clústeres 1 y 2 que la cantidad de unidades agrícolas es relativamente alta de 457 y 471 respectivamente. En la distribución de los clústeres 1 y 2 (figura 11) se evidencia rangos Inter cuartil pequeños mostrando que la mayoría de las unidades productoras agrícolas se encuentran con un potencial de producción menor a 10.

En la figura 12 se presenta un mapa de burbujas donde se observa el potencial de producción de cada una de las unidades productivas agrícolas de Soja. Se puede observar en color morado el clúster 3, asociado a las zonas del caribe, parte de la región andina y Orinoquia, donde las burbujas más grandes representan los outliers. En este clúster existen 57 outliers, de los cuales algunas unidades productivas agrícolas con alto potencial agrícola y bajo rendimiento se correlacionan con la baja fertilidad de los suelos en Colombia (figura10). [1]Los suelos con fertilidad baja se encuentran ampliamente distribuidos en el país y es en la Región Andina donde se presentan con mayor frecuencia y extensión. Los suelos con fertilidad muy baja se presentan diseminados en diferente proporción en el territorio nacional, pero su mayor extensión, concentración y frecuencia se encuentra en las regiones de la Orinoquia y Amazonia.

La inversión en unidades productivas agrícolas ubicadas en suelos con fertilidades bajas implica mayor inversión en insumos agrícolas y disminución de la relación costo beneficio, por esta razón se recomienda en estudios posteriores desarrollar metodologías que involucren el componente de beneficio económico en la focalización de recursos para las unidades productivas agrícolas. También para estudios posteriores se recomienda involucrar el mapa de tenencia de la tierra en Colombia el cual muestra que la distribución de la propiedad de la tierra es muy desigual. El 1% de las fincas de mayor tamaño acapara el 81% de la tierra, mientras que el 99% de las fincas restantes se reparten el 19% de la tierra (https://www.minagricultura.gov.co/).

[1] Fertilidad de los suelos de Colombia - INSTITUTO GEOGRÁFICO AGUSTÍN CODAZZI

5 Conclusiones

Con la aplicación de clusters en la metodología propuesta se encontró patrones y relaciones entre los datos de las unidades productivas agrícolas. En la metodología para el tipo de cultivo Soja (Soya) se definió una nueva variable dependiente de la producción y el área cosechada nombrada como el potencial de producción. Se encontró con la aplicación del algoritmo K-medoids cinco grupos susceptibles de los cuales se identificó el grupo con mayor potencial de producción, asociado a las zonas del caribe, parte de la región andina y Orinoquia.

Se observó que dentro de los cinco grupos encontrados con el algoritmo de Kmedoides, existe con el grupo asociado a las zonas del caribe, parte de la región andina y Orinoquia, una relación con el top de unidades productivas agrícolas de mayor potencial de producción del país.

Se seleccionó dentro de la evaluación de 4 algoritmos (K-medias, K-medoides, DBSCAN y Jerárquico Aglomerativo) el algoritmo de K-medoides, ya que se observó distribuciones consistentes en los cluster considerando las ubicaciones y el número de unidades productivas agrícolas en cada cluster.

6 GitHub

GitHub del proyecto: https://github.com/grupovajo/proyectoANS

7 Video

https://youtu.be/hUJHjngwdrs

8 Bibliografía

Bulman, A., Cordes, K.Y., Mehranvar, L., Merrill, E. y Fiedler, Y. 2021. Guía sobre incentivos para la inversión responsable en la agricultura y los sistemas alimentarios. Roma, FAO y Centro Columbia sobre Inversión Sostenible. Roma. https://doi.org/10.4060/cb3933es.

Dankevych, V., Dankevych, Y., & Pyvovar, P. (2018). Clustering of the International Agricultural Trade Between Ukraine and the Eu. Management Theory and Studies for Rural Business and Infrastructure Development, 40(3), 307–319. https://doi.org/10.15544/mts.2018.29

FAO. 2013. Tendencias e impactos de la inversión extranjera en la agricultura de los países en desarrollo – Datos de estudios de casos. Roma, Italia. FAO. 382 pp. (También disponible en: http://www.fao.org/fileadmin/templates/est/INTERNATIONAL-TRADE/FDIs/Trends_publication_12_November_2012.pdf).

FAO. Food and Agriculture Organization of the United Nations. 2014. Principios para la Inversión Responsable en la Agricultura y los Sistemas Alimentarios.

FAO. 2017. Resumen del Programa Marco de la FAO de apoyo a la inversión responsable en la agricultura y los sistemas alimentarios. Roma, Italia, FAO. 12 pp. (También disponible en: http://www.fao.org/publications/card/es/c/92303c68-3a89-4fe5-be2c-0ea6eff22501).

Gonzáles, H., y Ticona, U. (2019). Clustering, mediterraneidad y comercio internacional: aplicación empírica de los algoritmos Partitioning Around Medoids y K-means. Revista Latinoamericana de Desarrollo Económico, 32, 96–130. https://doi.org/10.35319/lajed.201932400

Gopinath, M., Batarseh, F., & Beckman, J. (2021). Machine Learning in Gravity Models: An Application to Agricultural Trade. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3603781

Pintado, P. X. (2022). Modelo de minería de datos para la recomendación de exportaciones de productos - estudio de caso Ecuador. Universidad del Azuay. Cuenca. Ecuador

Ramírez, C. A. (2020). Aplicación del machine learning en agricultura de precisión. Revista CINTEX, 25 (2), 14–27. https://doi.org/10.33131/24222208.356

Sahan, E., y Mikhail, M. 2012. Inversión privada en agricultura: por qué es fundamental y qué se necesita. Documento de trabajo de Oxfam.