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 de 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, distancia a grandes ciudades, dimensiones socio-económicas de las regiones y distancia a puertos marítimos en los casos que se quiera potenciar las exportaciones.

Al lograr esta agrupación, se espera tener zonas de interés agrícolas para diferentes fines como disminuir las importaciones nacionales de productos que ya se producen en nuestro país, así como aumentar exportaciones de productos demandados por otros países. Por lo anterior, se tendrían unos grupos potenciales que maximizarían los efectos de subsidios e inversiones agrícolas, mejorando la calidad de vida de las personas y de los agricultores en todo el país.

Con el proyecto se espera aportar en el desarrollo de focalización con un enfoque innovador que aprovecha la inteligencia artificial y la analítica para abordar este problema crítico. Al finalizar este proyecto, esperamos 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 agricultura.

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 Revisión preliminar de antecedentes en la literatura

La inversión en agricultura es una forma de apoyar el desarrollo rural, la seguridad alimentaria, la sostenibilidad ambiental y la reducción de la pobreza. Sin embargo, no todas las inversiones son responsables ni beneficiosas para las comunidades locales, los pequeños productores y el medio ambiente. Por eso, es importante focalizar la inversión hacia la agricultura de manera que respete los derechos humanos, los recursos naturales y los principios de buena gobernanza.

Para lograr este objetivo, se pueden seguir algunas recomendaciones, tales como:

- Aplicar los Principios para la inversión responsable en la agricultura y los sistemas alimentarios (Principios CSA-IRA[1]), que son un conjunto de directrices voluntarias acordadas por los Estados miembros de la FAO, el FIDA[2] y el Banco Mundial para orientar las políticas, los marcos legales y las prácticas de inversión [FAO (2014)].
- Utilizar incentivos para fomentar las inversiones que contribuyan a los objetivos de desarrollo sostenible, como exenciones fiscales, subsidios, créditos, garantías, infraestructura, servicios públicos, asistencia técnica, etc. Estos incentivos deben ser transparentes, equitativos y eficientes [Bulman, Cordes, Mehranvar, Merrill y Fiedler (2021)].
- Promover modelos empresariales inclusivos que involucren a los pequeños productores y a las comunidades locales como socios o accionistas de las inversiones, respetando sus derechos sobre la tierra y los recursos naturales, y asegurando su participación en la toma de decisiones y el reparto de beneficios [Sahan y Mikhail (2012)].
- Diversificar las fuentes de financiación para la agricultura, aprovechando el potencial de los fondos públicos, privados, nacionales e internacionales. Se puede recurrir a instrumentos financieros innovadores, como los bonos verdes o sociales, los fondos de impacto o las alianzas público-privadas [FAO (2013)].
- Fortalecer las capacidades de los actores involucrados en la inversión agrícola, como los gobiernos, los pequeños productores, las organizaciones de la sociedad civil, el sector privado y el mundo académico. Esto implica mejorar el acceso a la información, la educación, la asistencia técnica y jurídica, el diálogo y la cooperación [FAO (2017)].

Para focalizar la inversión hacia la agricultura con aprendizaje no supervisado, se requiere contar con datos de calidad, cantidad y diversidad suficientes, así como con algoritmos adecuados y herramientas informáticas potentes. Además, se debe tener en cuenta el contexto, los objetivos y los principios de la inversión responsable en agricultura, así como las limitaciones éticas y legales del uso de los datos.

El aprendizaje no supervisado puede tener varias aplicaciones en el ámbito de la agricultura, las importaciones y las exportaciones como, por ejemplo:

La agrupación en clústeres, que consiste en agrupar datos en función de sus similitudes o diferencias. Esta técnica puede servir para segmentar los mercados agrícolas, identificar los patrones de consumo y demanda, agrupar a los inversores potenciales según sus características, preferencias, necesidades o comportamientos, optimizar las cadenas de suministro y distribución, o detectar anomalías o fraudes. El uso del algoritmo de agrupación en clústeres k-medias para clasificar los cultivos según sus características físicas, químicas o biológicas, y así mejorar la gestión de los recursos, la calidad de los productos o la prevención de plagas. Esto también puede servir para diseñar incentivos personalizados y eficaces para atraer y retener a los inversores que contribuyan a la inversión responsable en agricultura.
Autores como Dankevych, Dankevych, y Pyvovar (2018), mediante la aplicación de K-means generaron tres clusters para determinar los factores que incentivan y desincentivan el comercio agrícola entre Ucrania y la Unión Europea para el periodo comprendido entre los años 2014, 2015 y 2016.

[1] Principios CSA-IRA: Principios del Comité de Seguridad Alimentaria Mundial para la inversión responsable en la agricultura y los sistemas alimentarios. Instrumento internacional voluntario, desarrollado a través de un proceso inclusivo de múltiples partes interesadas, consistente en diez principios para conseguir inversiones que sean responsables, contribuyan a la seguridad alimentaria y promuevan el desarrollo sostenible.
[2] El Fondo Internacional de Desarrollo Agrícola (FIDA, en inglés IFAD).

Por su parte, Gonzáles y Ticona (2019), con información de 264 países, realizaron la generación de clusters según su condición de mediterraneidad y sus condicionantes en el comercio exterior, para su elaboración se emplearon Algoritmos K-means y K-Medoids (Partitioning Around Medoids-PAM). El enfoque presentado por estos autores está relacionado con la metodología que se planea abordar en el presente trabajo.

A nivel nacional Ramírez (2020), propone un modelo de Machine Learning para predecir el estado de la cosecha a partir de información de consumo de pesticidas y otras variables del cultivo, para lo cual se sigue la metodología de machine Learning. Al comparar los métodos KNN y el árbol de decisión, este último presenta un mejor desempeño ya que logra obtener un mejor ajuste, adicionalmente, lo logra con un coste computacional significativamente menor por lo cual se puede decir que es el mejor modelo para la predicción durante esta primera ronda de comparación. Adicionalmente, se propusieron un modelo de ensamble tipo Boosting a partir del modelo árbol de decisión con el cual se logró hacer una mejor aproximación con la suma varios modelos más simples dando como resultado el modelo seleccionado para la predicción del estado de la cosecha.

La asociación, que consiste en encontrar reglas o patrones que relacionen diferentes variables o elementos. Esta técnica puede servir para descubrir las preferencias o hábitos de los consumidores, recomendar productos o servicios complementarios, o predecir el comportamiento o las tendencias del mercado. El uso del algoritmo de asociación Apriori para analizar las transacciones comerciales entre países y así identificar los productos más demandados, los socios comerciales más rentables o las oportunidades de negocio.

Pintado (2022) se propuso construir un modelo de minería de datos, con el cual se analizó el comportamiento de las exportaciones de productos del Ecuador entre los años 2008 y 2018. Para ello utilizó el enfoque metodológico CRISP-DM, con el objetivo de identificar, analizar y seleccionar las variables, parámetros y técnicas de minería de datos utilizadas en el modelo. Este estudio determinó que para la clasificación el algoritmo de clustering que mejor se adapta a los datos estudiados es K-means y para la asociación es Apriori. Para la elaboración del modelo se validaron diversos algoritmos de segmentación (K-means, K-Medoids y "Clara"), en consecuencia y mediante la generación de reglas de asociación se identificó las relaciones existentes entre los productos exportados y sus países compradores. Se destaca la trascendencia de las asociaciones de productos, ya que al identificar los países que generan estas asociaciones y cuales no, se está en la condición de recomendar productos o partidas arancelarias, sobre las cuales se puedan definir estrategias comerciales, con el objetivo de incrementar las exportaciones. En cuanto al presente trabajo, éste coincide en utilizar la metodología de K-means y K-Medoids como lo menciona el autor anteriormente, en cuanto a las principales diferencias en los métodos empleados, planteamos utilizar DBSCAN el cual es un algoritmo de clustering basados en densidad, con la ventaja de eliminar el ruido de manera más eficiente. Adicionalmente, también emplearemos el Algoritmo Jerárquico Aglomerativo, el cual permite identificar la estructura jerárquica de los clústeres sin necesidad de especificar al número de clústeres previamente.

4 Descripción detallada de los datos

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

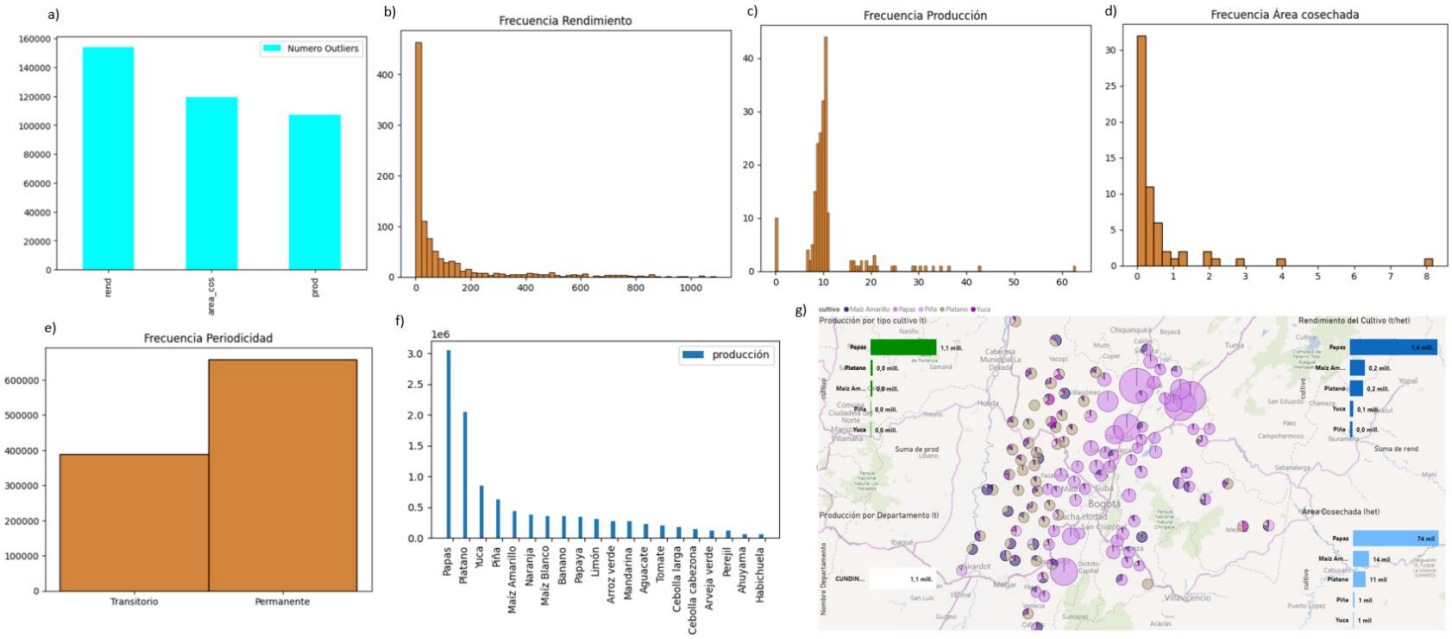
	rend	prod	area_cos
rend	1.000000	0.009471	-0.025247
prod	0.009471	1.000000	0.237365
area_cos	-0.025247	0.237365	1.000000

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, Gráfica de frecuencias



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.

5 Propuesta metodológica

La idea del proyecto es focalizar los subsidios y las inversiones que permitan fomentar el sector agropecuario colombiano. Para esto, debemos agrupar las UPA de acuerdo con las variables relacionadas a ellas, al igual que variables socioeconómicas. A partir de lo anterior, surgen distintos modelos de Aprendizaje No Supervisado los cuales nos permitirán potencializar el rendimiento en los cultivos de cada una de las UPA.

Dentro de los algoritmos de Aprendizaje No Supervisado, tenemos los modelos de clustering, los cuales nos permitirán segmentar las UPA de acuerdo con similitudes en las observaciones. Algunos modelos útiles para este proyecto son:

- **K-means:** este algoritmo nos permite agrupar las UPA en clústeres basados en similitudes de las variables seleccionadas, en donde utilizaremos métricas como el método del codo o el índice de Silhoutte para determinar el número óptimo de clústeres.
- **K-medoids:** 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.
- **DBSCAN:** el algoritmo DBSCAN nos permite identificar clústeres de una manera más flexible sin necesidad de especificar la cantidad de clústeres. Adicionalmente, podemos ajustar los hiperparámetros radio épsilon (eps) y el número mínimo de muestras (min_samples) para obtener mejores resultados en las agrupaciones.
- **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.

6 GITHUB

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

7 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.