GRUPO 23

PROYECTO APRENDIZAJE NO SUPERVISADO

Análisis Tipológico Enfocado para la Caracterización de los Agricultores de la Cadena de la papa en Colombia

Integrantes:

- JAIME ANDRES UNRIZA VARGAS
- JHON FARLEY ADARVE DIAZ
- JHON NICOLAS GARZON RODRIGUEZ
- JINNETH CAROLINA PARADA AMAYA
- VICTOR ERNESTO VELANDIA SUAREZ

1. Resumen

El problema que busca resolver este proyecto es la falta de una tipología clara que clasifique a los agricultores de papa en Colombia según sus rendimientos y características socioeconómicas. Esto es crucial para mejorar la planificación y toma de decisiones en el sector agrícola. Utilizando datos del Censo Nacional Agropecuario (CNA) y de la Unidad de Planificación Rural Agropecuaria (UPRA), aplicaremos técnicas de aprendizaje no supervisado, como K-medias y K-medoides, para generar una segmentación de los agricultores. Los resultados permitirán identificar patrones en la distribución de cultivos y características de los productores, proporcionando una herramienta valiosa para diseñar políticas públicas y estrategias agrícolas más eficientes y sostenibles.

2. Introducción

El sector agrícola en Colombia enfrenta desafíos relacionados con la variabilidad en los rendimientos agrícolas y las condiciones socioeconómicas de los productores. Particularmente, los agricultores de papa, exhiben gran diversidad en sus prácticas agrícolas, acceso a recursos y niveles de producción. Esta heterogeneidad dificulta la implementación de políticas y programas de apoyo eficaces.

El problema que abordamos en este proyecto es la falta de una caracterización clara y precisa de los agricultores de papa en Colombia, lo que limita la efectividad de las políticas públicas y los programas agrícolas. La pregunta de investigación que guía nuestro trabajo es: **¿Cómo podemos segmentar a los agricultores de papa en grupos homogéneos basados en sus rendimientos y características socioeconómicas, utilizando técnicas de clustering no supervisado?**

Estudios previos, como el de Sánchez-Toledano (2016), han demostrado que las segmentaciones en grupos con características similares pueden mejorar la efectividad de las intervenciones agrícolas al adaptarse a las necesidades de cada grupo. Sin embargo, la mayoría de estos enfoques se ha centrado en regiones específicas, y poco se ha explorado en el contexto colombiano. A nivel internacional, técnicas de minería de datos espacial y clustering han sido aplicadas en la agricultura para optimizar el uso de recursos y predecir rendimientos, como en los estudios de Gil-Torres (2019). Nuestro proyecto se sitúa en la intersección de estas investigaciones, aplicando técnicas de aprendizaje no supervisado para una mejor caracterización de los agricultores de papa en Colombia.

El cliente potencial de este proyecto son las entidades gubernamentales, como el Ministerio de Agricultura y la UPRA (Unidad de Planificación Rural Agropecuaria), además de cooperativas y organizaciones agrícolas. El contexto organizacional subraya la necesidad de una segmentación precisa para mejorar la planificación y ejecución de programas de apoyo al agricultor.

Utilizando técnicas de clustering, como K-medias y K-medoides, esperamos identificar patrones ocultos en los datos que permitan clasificar a los agricultores en grupos homogéneos basados en similitudes socioeconómicas y productivas. Si bien los métodos de clustering no supervisado son adecuados para abordar este tipo de problemas, su limitación radica en la dependencia de la calidad y cantidad de datos disponibles, lo que podría influir en la precisión de los clusters formados. No obstante, nuestra recomendación es continuar con este enfoque, ya que proporcionará una herramienta robusta para mejorar la planificación y ejecución de políticas agrícolas.

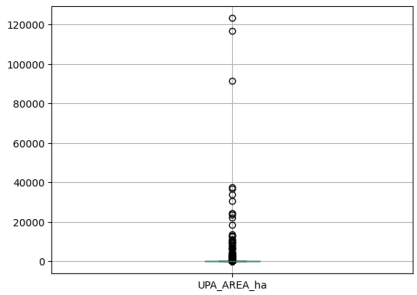
3. Materiales y métodos

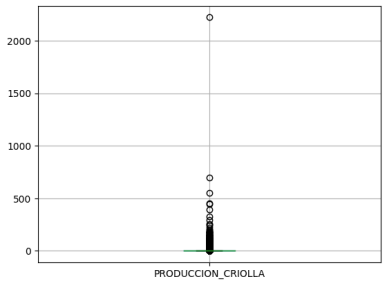
3.1 Datos

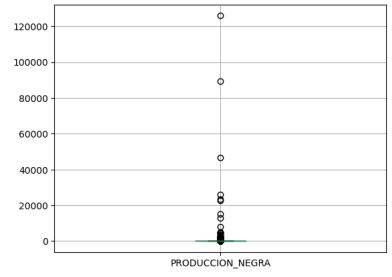
La base de datos utilizada para el análisis propuesto contiene información recolectada a través del Censo Nacional Agropecuario (CNA) y detalla diversas características asociadas a unidades productivas agrícolas (UPA) tales como (Variedad de papa cosechada, área sembrada por variedad, area total disponible, tipo de cultivo, entre otras.) lo que ofrece en primera instancia una amplia variedad de información para generar un contexto más claro acerca del estado actual de la producción de papa en Colombia.

En total se cuenta con 37,484 observaciones cada una asociada a una unidad de producción (UPA), y con 24 columnas o variables distribuidas de la siguiente forma: 13 variables numéricas y 11 categóricas; a continuación se presenta un pequeño resumen estadístico:

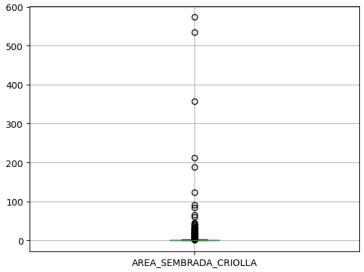
- Se encuentra una distribución de las áreas asociadas a cada unidad agrícola (UPA) extremadamente concentrada en valores relativamente bajos (menos de 1 hectárea), lo que refleja una predominancia de pequeños agricultores en la muestra. Información corroborada al observar la diferencia entre la mediana de la muestra (1.680011) y la media (30.816426).

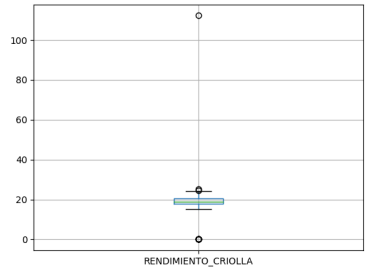


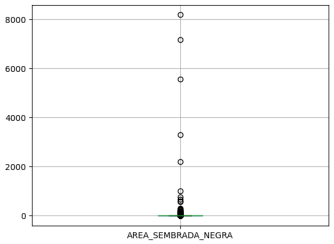


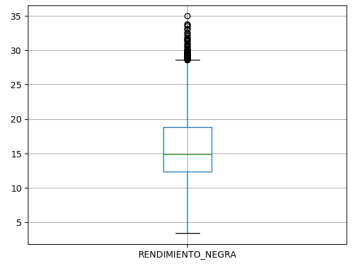


- Se cuenta con datos asociados a la producción de dos tipos de papa (Criolla - Negra) con una diferencia marcada entre unidades que se dedican a la producción de cada variedad (6221 - Criolla, 32869 - Negra), lo que genera diferencias interesantes en los números de producción por toneladas, cómo se observa en los gráficos anteriores (una producción media de Criolla de 4.947 toneladas, comparada con las 38.094 para la variedad Negra).
- Al analizar un poco más en profundidad estas diferencias de producción entre variedades, se observa que los rendimientos por cultivo (Rendimiento = Producción por Área sembrada) y las áreas sembradas por variedad muestran comportamientos opuestos; se encuentran mayores rendimientos promedio para la variedad criolla (19.086428 toneladas por hectárea, en comparación con las 15.589 de la variedad negra), y una mayor área sembrada promedio para la variedad negra (3.190960 hectáreas, en comparación con las 1.531621 de la variedad criolla).





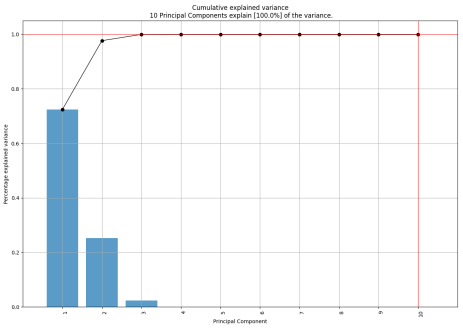


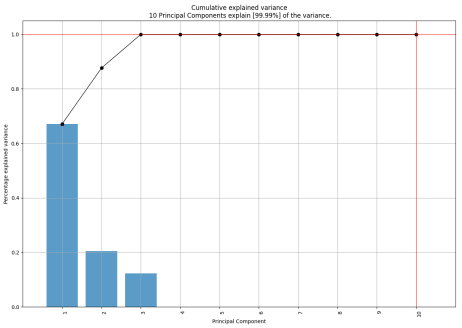


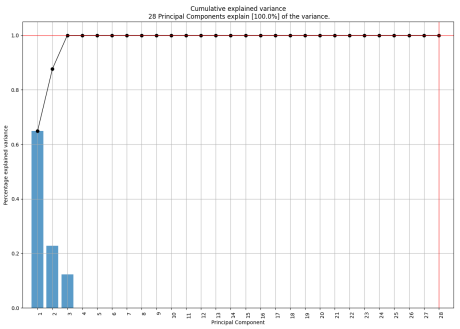
3.2 Métodos

3.2.1 Preparación de datos: Para profundizar en el análisis de las diferencias de producción entre variedades de papa, se decidió hacer un estudio centrado en la clasificación de UPAs de acuerdo a su variedad preferida para la producción, por lo que se dividió el conjunto de datos en tres grupos: uno con una preferencia por la producción de la variedad Criolla (6,221 observaciones); uno con una preferencia por la producción de la variedad negra (32,869 observaciones) y un grupo con el conjunto completo de datos. Para asegurar un análisis adecuado, se realizó una estandarización de los datos numéricos, y se codificaron las variables categóricas de interés a través de métodos como: reemplazo de categorías por valores numéricos (variables binarias); y generación de variables dummy (variables con un número de categorías elevado).

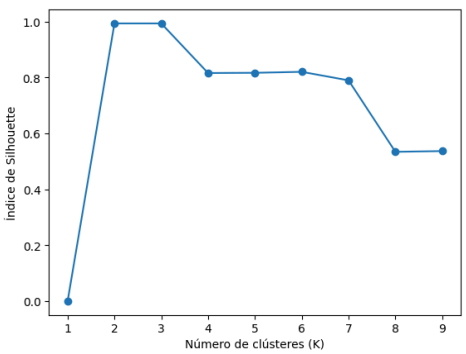
3.2.2 Reducción de dimensionalidad: En busca de optimizar el análisis y los recursos computacionales disponibles, se decidió que era necesario aplicar los conceptos de reducción de dimensionalidad aprendidos previamente de forma que no se utilizara información redundante durante los procesos de clasificación, por lo que se estimó un modelo de componentes principales (PCA) para cada grupo de datos, encontrando que los primeros 3 componentes lograban explicar más del 99% de la varianza de los datos, cómo se muestra a continuación:







K-means: Una vez establecidos los componentes principales a utilizar, se procedió a verificar el número óptimo de clusters para los datos disponibles. A través del coeficiente de Silhouette (que valiéndose de la varianza intra-cluster, logra medir qué tan coherente es la asignación de elementos a una agrupación), se estableció que el número óptimo de clusters sería entre 2-3 para cada agrupación de datos:



K-medoides: Como primera alternativa de comparación, se realizó una clusterización con el algoritmo de k-medoides, introduciendo variaciones en la estimación del modelo: se hizo un análisis sobre los datos estandarizados, sin reducción de dimensionalidad (para evaluar posibles efectos secundarios de la eliminación de información redundante sobre la clasificación de los datos).

4. Resultados

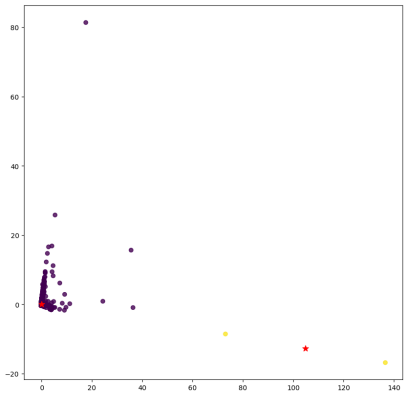
Luego de aplicar los algoritmos seleccionados sobre los conjuntos de datos se encontraron los siguientes resultados:

4.1. Subgrupo (Variedad Criolla)

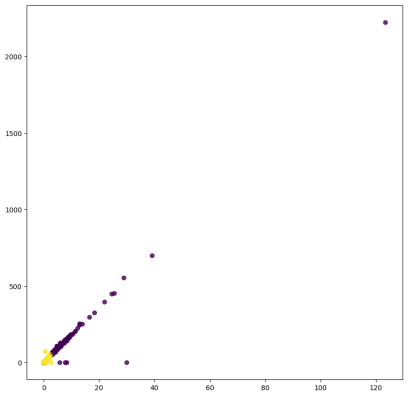
Se observa que los datos para la variedad criolla se logran representar en dos grupos:

- UPAs medianas-pequeñas con niveles de producción bajos (agricultores pequeños, grupos familiares) que no privilegian la producción de ninguna variedad y que extraen unos rendimientos mínimos para subsistir de la actividad agrícola sin demasiados excedentes.
- UPAs grandes: Tienen extensiones de tierra mayores que logran dedicar a la producción de ambas variedades; sin embargo la variedad criolla no representa su mayor flujo de rendimientos, la variedad negra representa una fuente de rendimientos e ingresos mayor.

K-medias



K-medoides



4.2 Subgrupo (Variedad Negra)

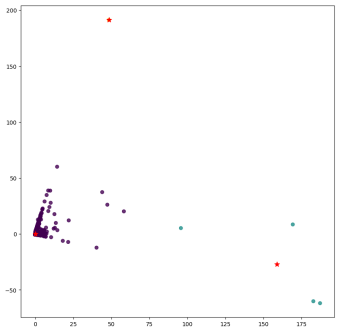
Se encuentran resultados similares con agrupaciones de datos divididos de acuerdo al tamaño de la UPA , y con una marcada diferenciación en el nivel de diversificación de la producción (Unidades pequeñas buscan maximizar el uso y el rendimiento del espacio reducido, mientras que unidades grandes buscan maximizar rendimientos con un mayor nivel de especialización de variedades).

cluster	UPA_AREA_ha	AREA_SEMBRADA_CRIOLLA	AREA_COSECHADA_CRIOLLA	PRODUCCION_CRIOLLA	RENDIMIENTO_CRIOLLA	AREA_SEMBRADA_NEGRA
0	23.138658	0.083695	0.023213	0.405241	0.254213	2.454286
1	36885.501239	276.921651	0.000000	0.000000	0.000000	6055.891530

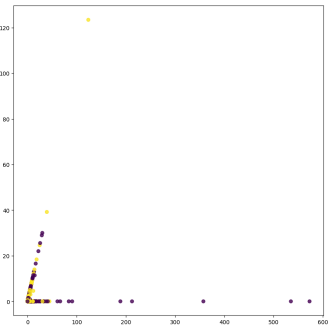
RENDIMIENTO_CRIOLLA	AREA_SEMBRADA_NEGRA	AREA_COSECHADA_NEGRA	PRODUCCION_NEGRA	RENDIMIENTO_NEGRA
0.254213	2.454286	1.769485	29.422105	15.589700
0.000000	6055.891530	4947.908883	71296.659745	14.841087

4.3 Subgrupo (Todas las variedades)

K-means



K-medoids



cluster	UPA_AREA_ha	TIPO_CULTIVO	AREA_SEMBRADA_CRIOLLA	AREA_COSECHADA_CRIOLLA	PRODUCCION_CRIOLLA	RENDIMIENTO_CRIOLLA	AREA_SEMBRADA_NEGRA
0	26.067704	0.850903	0.221380	0.042608	0.761988	0.640164	2.152142
1	36885.501239	0.250000	276.921651	0.000000	0.000000	0.000000	6055.891530
2	30589.459897	0.000000	123.417969	123.417969	2222.049240	18.003776	0.000000

AREA_COSECHADA_CRIOLLA	PRODUCCION_CRIOLLA	RENDIMIENTO_CRIOLLA	AREA_SEMBRADA_NEGRA	AREA_COSECHADA_NEGRA	PRODUCCION_NEGRA	RENDIMIENTO_NEGRA
0.042608	0.761988	0.640164	2.152142	1.551645	25.799981	13.670469
0.000000	0.000000	0.000000	6055.891530	4947.908883	71296.659745	14.841087
123.417969	2222.049240	18.003776	0.000000	0.000000	0.000000	0.000000

Se generó una división en 3 clusters que a diferencia de las clasificaciones anteriores no parecen centrar su clasificación en el tamaño de las unidades productivas, sino en su especialidad: un grupo dedicado a la extracción de rendimientos de la variedad criolla, un grupo que basa su producción en la variedad negra y un grupo con una distribución de cosechas diversificada.

5. Conclusión

El análisis tipológico de los agricultores de papa en Colombia utilizando técnicas de clustering no supervisado ha permitido identificar patrones clave en su producción y características socioeconómicas. A través de K-means y K-medoides, se lograron segmentar los agricultores en tres grupos principales: pequeños y medianos productores que se enfocan en la subsistencia con bajos rendimientos, grandes productores especializados en una variedad (principalmente papa Negra), y un tercer grupo de agricultores diversificados que cultivan tanto papa Criolla como Negra, maximizando sus ingresos a través de la diversificación. Las técnicas de reducción de dimensionalidad (PCA) facilitaron el análisis al explicar más del 99% de la varianza de los datos con solo tres componentes principales. Este enfoque permite obtener una visión clara y precisa de las diferencias entre los agricultores, particularmente en relación con el tamaño de las Unidades Productivas Agrícolas (UPAs) y los rendimientos por hectárea. La segmentación obtenida es crucial para la toma de decisiones en el sector agrícola, ya que ofrece información valiosa para diseñar políticas públicas más efectivas y personalizadas, mejorando así la planificación y la ejecución de programas de apoyo para los agricultores. En conclusión, este análisis proporciona una herramienta robusta para optimizar las intervenciones en el sector de la papa en Colombia.

6. Bibliografía

- Salazar, A. & Espinosa, R. (2016). *Análisis y segmentación de productores agrarios para la transferencia del riesgo*. Revista de Estudios Agrarios, 24(2), 34-48.
- Gil-Torres, A. F., Monroy-García, A. L., & González-Sanabria, J. S. (2019). *Minería de datos espacial en la agricultura en Latinoamérica: Una aproximación conceptual*. Revista de Tecnología Agrícola, 35(3), 123-137.
- Sánchez-Toledano, B. I., Kallas, Z., & Gil, J. M. (2017). *Importancia de los objetivos sociales, ambientales y económicos de los agricultores en la adopción de maíz mejorado en Chiapas, México*. Revista Mexicana de Ciencias Agrícolas, 8(3), 269-287.