Typologies of Technological Adoption in Agriculture: A Machine Learning Approach Applied to Regional Productive Systems

Javier Fornari[1] 0009-0004-1701-3526 Mariel Lopez[1] Francisco Gatto[1] Milton Jullier[1] Lucas Ramos[1]

[1] Universidad Nacional de Rafaela, Argentina
javier.fornari@unraf.edu.ar marielopez@unraf.edu.ar
franciscogatto@unraf.edu.ar miljullier21@gmail.com
lucasramos012002@gmail.com

Abstract. This study analyzes the structural and technological heterogeneity of agricultural producers through a combined approach of unsupervised learning and statistical validation. A dataset of 62 firms from the Castellanos department (Santa Fe, Argentina) was used, integrating productive, organizational, technological, demographic, and institutional linkage variables. After numerical encoding and standardization, K-Means clustering was applied evaluating solutions from k=2 to k=14. The selection of four clusters was based on statistical metrics (inertia, silhouette score) and analytical interpretability criteria. Silhouette values were low (0.02–0.04), indicating geometrically weak separations. A Random Forest classifier reproduced the cluster assignments with a mean cross-validation accuracy of 0.79 and a macro F1 of 0.78, albeit with wide uncertainty due to the small sample size. Non-parametric Kruskal-Wallis tests revealed statistically significant differences between clusters in key dimensions such as management type, advisory networks, technology adoption, and sociodemographic characteristics. A K-Medoids stability check showed only moderate agreement with K-Means (ARI = 0.165), leading to an interpretation of the groups as regions within a continuum of profiles rather than discrete types. The resulting exploratory taxonomy identifies four differentiated profiles based on farm scale, institutional linkage, digital adoption, advisory networks, and educational levels. Findings position institutional advisory and digital adoption as differentiation axes, providing preliminary evidence for extension policy design and innovation strategies in heterogeneous agricultural systems.

Keywords: agricultural production, technology adoption, digital transformation, machine learning, clustering.

Tipologías de adopción tecnológica en el agro: un enfoque de aprendizaje automático aplicado a sistemas productivos regionales

Resumen. Este estudio analiza la heterogeneidad estructural y tecnológica de productores agropecuarios mediante un enfoque combinado de aprendizaje no supervisado y validación estadística. Se utilizó un dataset de 62 empresas del departamento Castellanos (Santa Fe, Argentina) que integra variables productivas, organizacionales, tecnológicas, demográficas y de vinculación institucional. Tras la codificación numérica y estandarización de las variables, se aplicó clustering mediante K-Means evaluando soluciones entre k=2 y k=14. La selección de cuatro clusters se fundamentó en métricas estadísticas (inercia, silhouette score) y criterios de interpretabilidad analítica. Los valores de silhouette fueron bajos (0.02–0.04), indicando separaciones geométricamente débiles. Un clasificador Random Forest reprodujo las asignaciones con un accuracy promedio en validación cruzada de 0.79 y un macro F1 de 0.78, si bien con amplia incertidumbre asociada al tamaño muestral. Pruebas no paramétricas de Kruskal-Wallis revelaron diferencias estadísticamente significativas entre clusters en dimensiones clave como tipo de gestión, redes de asesoramiento, adopción tecnológica y características sociodemográficas. Una prueba de estabilidad con K-Medoids arrojó una concordancia sólo moderada con K-Means (ARI = 0.165), lo que llevó a interpretar los grupos como regiones de un continuo de perfiles más que como tipos discretos. La taxonomía resultante, de carácter exploratorio, identifica cuatro perfiles diferenciados según escala productiva, vinculación institucional, adopción digital, redes de asesoramiento y nivel educativo. Los hallazgos posicionan al asesoramiento institucional y la adopción digital como ejes de diferenciación, aportando evidencia preliminar para el diseño de políticas de extensión y estrategias de innovación en sistemas agroproductivos heterogéneos.

Palabras clave: producción agropecuaria, nivel de adopción tecnológica, transformación digital, aprendizaje automático, clustering.

1 Introducción

La agricultura contemporánea enfrenta el desafío de incrementar su productividad y sostenibilidad en un contexto de creciente complejidad tecnológica, climática y de mercado. La transformación digital del sector agropecuario, frecuentemente denominada AgTech o agricultura 4.0, comprende la incorporación de tecnologías como sensores, sistemas de posicionamiento global (GPS), plataformas digitales de gestión, herramientas de análisis de datos y maquinaria con manejo autónomo (Wolfert, Ge, Verdouw & Bogaardt, 2017). Sin embargo, la adopción de estas tecnologías no es homogénea: los productores agropecuarios difieren significativamente en sus capacidades, motivaciones y restricciones para incorporar innovaciones tecnológicas (Pierpaoli, Carli, Pignatti & Canavari, 2013).

Comprender esta heterogeneidad resulta fundamental para el diseño de políticas públicas de extensión agropecuaria, estrategias de transferencia tecnológica y programas de desarrollo rural que sean efectivos y focalizados. Los enfoques tradicionales de caracterización de productores, basados en categorías predefinidas como tamaño de explotación o rubro productivo, suelen resultar insuficientes para capturar la complejidad multidimensional de los factores que inciden en la adopción tecnológica (Lowder, Skoet & Raney, 2016). En este contexto, las técnicas de aprendizaje automático (machine learning), particularmente los métodos de clustering, ofrecen un enfoque basado en datos que permite identificar patrones latentes y construir tipologías empíricas sin recurrir a categorías a priori (Jain, 2010).

El departamento Castellanos, en la provincia de Santa Fe, Argentina, constituye un territorio productivo de alta relevancia, caracterizado por la coexistencia de actividades agrícolas, ganaderas y lecheras, con un tejido institucional consolidado que incluye entidades como el Instituto Nacional de Tecnología Agropecuaria (INTA), la Sociedad Rural de Rafaela y diversas cooperativas y empresas de servicios. No obstante, la información disponible sobre los niveles de adopción tecnológica y los factores que los condicionan en este territorio es aún limitada, lo cual dificulta la focalización de las intervenciones de asistencia técnica y política pública.

El presente trabajo tiene como objetivo analizar la diversidad de perfiles de productores agropecuarios del departamento Castellanos mediante técnicas de aprendizaje automático, con énfasis en la identificación de tipologías de adopción tecnológica. La contribución principal consiste en la aplicación de un enfoque metodológico que combina clustering no supervisado (K-Means), una evaluación de reproducibilidad mediante clasificación supervisada (Random Forest), inferencia estadística no paramétrica (Kruskal-Wallis con post-hoc de Dunn) y una prueba de estabilidad entre algoritmos (K-Medoids), proporcionando una segmentación exploratoria y empíricamente fundamentada de los sistemas productivos regionales, junto con una caracterización explícita de su incertidumbre. Los resultados aportan evidencia preliminar para orientar estrategias de extensión, transferencia tecnológica y políticas de desarrollo rural adaptadas a la heterogeneidad del territorio.

La investigación se enmarca en un proyecto más amplio desarrollado desde la Universidad Nacional de Rafaela (UNRaf), orientado a comprender los procesos de transformación digital en los sistemas agroproductivos de la región centro-oeste de

Santa Fe. El relevamiento de datos fue posible gracias a la articulación con instituciones clave del territorio como la Sociedad Rural de Rafaela y el INTA Rafaela, lo que aseguró el acceso a productores de distintos perfiles productivos de la zona, si bien mediante un mecanismo de contacto institucional cuyas implicancias sobre la representatividad se discuten en la Sección 5. El artículo se organiza de la siguiente manera: la Sección 2 presenta el marco teórico y los antecedentes; la Sección 3 describe los materiales y métodos; la Sección 4 presenta los resultados; la Sección 5 desarrolla la discusión; y la Sección 6 ofrece las conclusiones y líneas de trabajo futuro.

2 Marco teórico y antecedentes

2.1 Adopción tecnológica en el sector agropecuario

La adopción de tecnologías en el sector agropecuario ha sido objeto de extensa investigación desde las contribuciones seminales de Rogers (2003) sobre la difusión de innovaciones. En el contexto específico de la agricultura de precisión y las tecnologías digitales, diversos estudios han identificado factores que inciden en la adopción, incluyendo características del productor (edad, nivel educativo, experiencia), características de la explotación (tamaño, tipo de actividad, infraestructura), factores institucionales (asesoramiento técnico, redes de apoyo, políticas públicas) y características de la tecnología misma (complejidad, costo, compatibilidad) (Pierpaoli et al., 2013; Pivoto, Waquil, Talamini, Finocchio, Dalla Corte & de Vargas Mores, 2018).

Wolfert et al. (2017) analizan la transformación digital en la agricultura, destacando que la big data y las tecnologías inteligentes están reconfigurando las cadenas de valor agropecuarias, pero que la adopción sigue siendo desigual entre regiones y tipos de productores. Pivoto et al. (2018) realizan una revisión sistemática sobre los factores que influyen en la adopción de tecnologías de agricultura de precisión, concluyendo que los determinantes son multidimensionales y que las intervenciones de política deben considerar la heterogeneidad de los adoptantes.

En Argentina, la adopción de tecnologías agropecuarias presenta marcadas diferencias regionales y por tipo de productor. Estudios previos han documentado la coexistencia de establecimientos altamente tecnificados con unidades productivas que mantienen prácticas tradicionales, incluso dentro de una misma región (Lowder et al., 2016). Esta heterogeneidad plantea desafíos específicos para la transferencia tecnológica y la extensión agropecuaria, que requieren enfoques diferenciados y basados en evidencia empírica.

La región pampeana argentina, donde se ubica el departamento Castellanos, ha sido históricamente una de las áreas de mayor dinamismo en la adopción de innovaciones agropecuarias, desde la siembra directa hasta las tecnologías de agricultura de precisión. Sin embargo, la incorporación de tecnologías digitales de gestión, plataformas de comercialización electrónica y herramientas de análisis de datos ha sido más heterogénea y menos documentada. La coexistencia de tres actividades productivas principales (agricultura, ganadería y lechería) en el territorio genera perfiles de adopción diferenciados que requieren análisis específicos para cada configuración productiva.

2.2 Aprendizaje automático para la segmentación de productores

Conviene distinguir dos planos en la literatura relevante para este trabajo: por un lado, las contribuciones que desarrollan o describen las técnicas de aprendizaje automático en sí mismas y, por otro, los trabajos que aplican esas técnicas al sector agropecuario. En el primer plano, los métodos de clustering constituyen una familia de técnicas de aprendizaje no supervisado que permiten agrupar observaciones en función de su similitud, sin requerir etiquetas predefinidas (Jain, 2010); el algoritmo K-Means particiona los datos en k grupos mediante la minimización iterativa de la distancia intra-cluster, y su calidad geométrica suele evaluarse con métricas internas como el silhouette score (Rousseeuw, 1987). En el segundo plano, los trabajos aplicados muestran el uso de estas técnicas en agricultura: Liakos, Busato, Moshou, Pearson y Bochtis (2018) presentan una revisión de aplicaciones de machine learning en el sector, que abarcan la segmentación de productores, la predicción de rendimientos y la detección de patrones de uso de suelo. La selección de K-Means en el presente estudio se justifica por su eficiencia computacional, su interpretabilidad y su adopción en este tipo de contextos agropecuarios.

La determinación del número óptimo de clusters representa un desafío metodológico central. Entre las métricas más utilizadas se encuentran la inercia (suma de distancias al cuadrado dentro de cada cluster), cuya curva permite identificar puntos de inflexión (método del codo), y el silhouette score, que mide la cohesión y separación de los clusters (Rousseeuw, 1987). Valores de silhouette cercanos a 1 indican clusters bien definidos, mientras que valores bajos sugieren solapamiento entre grupos. No obstante, silhouette scores bajos no invalidan necesariamente una segmentación si esta se sustenta en validaciones adicionales como clasificación supervisada y pruebas estadísticas.

En el ámbito agropecuario, Liakos et al. (2018) presentan una revisión comprehensiva de las aplicaciones de machine learning en agricultura, incluyendo la segmentación de productores, la predicción de rendimientos y la detección de patrones de uso de suelo. La combinación de clustering con métodos supervisados como Random Forest para validar las agrupaciones ha sido reportada como una práctica robusta que permite evaluar la capacidad predictiva de los clusters identificados y determinar las variables más discriminantes.

La validación estadística de los clusters mediante pruebas no paramétricas como Kruskal-Wallis complementa las métricas internas de calidad del clustering (silhouette, inercia) con una evaluación externa de la significancia de las diferencias entre grupos. Este enfoque multimétodo, que triangula evidencia de aprendizaje no supervisado, supervisado e inferencia estadística, ha sido recomendado para contextos donde la estructura de los datos no es claramente separable pero donde existen patrones latentes con relevancia práctica (Rousseeuw, 1987). En el contexto agropecuario, donde la heterogeneidad es inherente y las fronteras entre tipos de productores son difusas, esta aproximación resulta particularmente pertinente.

Un aspecto adicional relevante es la interpretabilidad de los modelos de machine learning en contextos de política pública. La importancia de variables derivada de modelos de ensamble como Random Forest permite identificar qué factores resultan

más discriminantes entre grupos, facilitando la traducción de los resultados técnicos en recomendaciones accionables para extensionistas y decisores de política.

3 Materiales y métodos

3.1 Diseño del instrumento y recolección de datos

Para la recolección de datos se diseñó un cuestionario mediante la plataforma LimeSurvey, estructurado en cinco bloques temáticos: (a) descripción general de la empresa, (b) descripción detallada de la actividad productiva, (c) infraestructura y prácticas, (d) conectividad y asociación, y (e) caracterización tecnológica. El instrumento fue diseñado de forma colaborativa por el equipo de investigación, orientado a relevar información estratégica sobre empresas agropecuarias. La elaboración del cuestionario siguió las recomendaciones metodológicas para investigación por encuestas de Kitchenham y Pfleeger (2002), en cuanto a la definición de objetivos, la formulación y secuenciación de preguntas y la organización en bloques temáticos. Se reconoce, no obstante, que el instrumento no fue sometido a un proceso formal de validación psicométrica ni a una prueba piloto sistemática previa, lo que constituye una limitación metodológica a considerar en la interpretación de los resultados.

La población objetivo estuvo integrada por empresas y productores agropecuarios del departamento Castellanos, provincia de Santa Fe, con especial énfasis en los radicados en las proximidades de la ciudad de Rafaela. El contacto con los destinatarios fue facilitado por instituciones referentes del sector, tales como la Sociedad Rural de Rafaela y el INTA Rafaela. El cuestionario fue distribuido a aproximadamente 440 participantes, obteniéndose 62 respuestas completas que constituyeron el dataset final para el análisis. La tasa de respuesta resultante (14.1%) se encuentra dentro del rango habitual en encuestas voluntarias al sector agropecuario, aunque su magnitud, junto con el mecanismo de contacto institucional, introduce un sesgo de autoselección que se discute más adelante (Sección 5).

3.2 Variables y preprocesamiento

Del total de variables relevadas, se seleccionó un subconjunto de 22 variables priorizando aquellas con presunta vinculación con la adopción tecnológica. Cada variable fue sometida a un proceso de limpieza y transformación para asegurar su compatibilidad con los algoritmos de análisis. La Tabla 1 sintetiza las variables utilizadas, agrupadas por dimensión.

Tabla 1. Variables empleadas en el análisis, agrupadas por dimensión.

Dimensión	Variables	Procesamiento
Productiva y organizacional	Antigüedad, agricultura, ganadería, lechería, servicios a terceros, superficie, tipo de gestión	Escalas ordinales, binarias y temporales

Demográfica y educativa	Proporción propietarios, edad promedio, proporción masculinos, nivel educación, áreas formación	Ratios, promedios ponderados, proporciones
Infraestructura y conectividad	Tipo de acceso, conectividad internet, registros digitales, participación BPA	Índices ordinales y binarios
Tecnológica y de gestión digital	Plataformas digitales, maquinaria tecnológica, software de gestión, inversión tecnológica	Proporciones sobre opciones posibles
Asesoramiento	Índice de asesoramiento (profesionales, organismos públicos, cooperativas, empresas, ONG)	Índice ponderado compuesto

Todas las variables fueron estandarizadas mediante StandardScaler (media=0, desvío estándar=1) para evitar el sesgo debido a diferencias en las escalas de medición.

El procesamiento de cada variable siguió criterios específicos para maximizar la información capturada. La antigüedad de la empresa se calculó como la diferencia en años desde la fecha de inicio de actividades. Las actividades productivas (agricultura, ganadería, lechería) se transformaron de un ranking cualitativo a una escala numérica según la prioridad asignada por cada productor. La superficie productiva se recodificó en una escala ordinal a partir de rangos de hectáreas. El tipo de gestión se codificó como variable binaria (1=grupal, 0=unipersonal). Para las variables demográficas del grupo directivo, se calcularon proporciones (propietarios, masculinos) y promedios ponderados (edad, nivel educativo).

El índice de conectividad se construyó como un indicador compuesto que toma valor 0 (sin conexión), 0.5 (solo wifi o solo datos móviles) o 1 (ambas tecnologías disponibles). El índice de digitalización de registros capturó la forma de registrar la producción en tres niveles: 0 (papel), 1 (digital manual) y 2 (digital automático). El índice de asesoramiento se diseñó como un indicador ponderado que combina la frecuencia de consulta con la diversidad de fuentes utilizadas (profesionales externos, organismos públicos, cooperativas, empresas proveedoras y ONG), mediante la fórmula $I = (\sum fi)/(k \cdot fmax) \cdot (k/n)$, donde fi es la frecuencia ponderada, k el número de asesores utilizados y n el total de fuentes posibles. Las variables de adopción tecnológica (plataformas digitales, maquinaria con tecnología, software de gestión) se expresaron como proporciones sobre el total de opciones posibles en cada categoría.

3.3 Descripción de la muestra

La muestra de 62 empresas agropecuarias presenta características heterogéneas que reflejan la diversidad productiva del departamento Castellanos. En términos de actividad principal, se observa la coexistencia de establecimientos orientados a la agricultura extensiva, la ganadería bovina y la lechería, con un subconjunto de

productores que ofrecen servicios a terceros. La superficie productiva varía ampliamente, desde pequeñas explotaciones hasta establecimientos de gran escala, lo que genera diferencias estructurales en las capacidades de inversión y adopción tecnológica.

Respecto a los perfiles demográficos del grupo directivo, se observa una predominancia masculina con una distribución etaria concentrada en los rangos de 45 a 64 años, aunque con presencia de productores jóvenes y de productores mayores de 65 años. Los niveles educativos son heterogéneos, con representación desde educación secundaria incompleta hasta estudios universitarios de posgrado, y áreas de formación que abarcan ciencias agropecuarias, ciencias económicas y de gestión, y otras disciplinas.

En materia de infraestructura y conectividad, la muestra incluye establecimientos con distinto tipo de acceso (desde caminos de tierra hasta rutas pavimentadas), diferentes niveles de conectividad a internet y datos móviles, y diversos grados de digitalización de los registros productivos. La adopción tecnológica presenta un amplio rango de variación: desde productores sin ninguna tecnología digital hasta aquellos que utilizan maquinaria con GPS, sensores y manejo autónomo, complementada con software de gestión y plataformas digitales de comercialización. Esta variabilidad es consistente con la hipótesis de heterogeneidad estructural que motiva el análisis de clustering.

3.4 Análisis de clustering

Se aplicó el algoritmo K-Means evaluando soluciones de k=2 a k=14 clusters. La selección del número óptimo de clusters se basó en la combinación de tres criterios: (a) el análisis de la curva de inercia, buscando puntos de inflexión donde la reducción marginal de la distancia intra-cluster disminuye significativamente; (b) el silhouette score, que evalúa la cohesión y separación de los clusters; y (c) un criterio de representatividad práctica, entendido como el balance en el tamaño de los clusters y la posibilidad de interpretarlos en términos sustantivos.

La caracterización de cada cluster se realizó comparando el promedio de cada variable dentro del cluster con el promedio general de la muestra, lo que permitió identificar los rasgos distintivos de cada grupo en términos de desviaciones respecto al perfil típico.

Como análisis complementario de robustez, se aplicó también el algoritmo K-Medoids (PAM, Partitioning Around Medoids), que difiere de K-Means en que utiliza puntos reales del dataset como representantes de cada cluster (medoides) en lugar de centroides promedio. Esta característica lo hace más robusto frente a valores atípicos y proporciona representantes interpretables directamente como productores concretos. La evaluación del número óptimo de clusters se realizó con los mismos criterios empleados para K-Means. La comparación entre los resultados de ambos algoritmos permite evaluar la estabilidad de la segmentación frente a variaciones en el método de agrupamiento.

3.5 Validación supervisada

Para evaluar la calidad predictiva de los clusters identificados, se entrenó un clasificador supervisado Random Forest utilizando la asignación de clusters como variable dependiente y las variables originales como predictoras, mediante validación cruzada estratificada con 5 folds. Random Forest se seleccionó por su capacidad de capturar interacciones no lineales entre variables y de cuantificar la importancia relativa de cada predictor. Cabe señalar que este procedimiento evalúa la separabilidad de las etiquetas generadas por K-Means en el mismo espacio de variables que las produjo; por lo tanto, un desempeño elevado indica que los clusters son internamente consistentes y reproducibles, pero no constituye, por sí solo, evidencia independiente de que correspondan a grupos naturales preexistentes en la población.

3.6 Inferencia estadística

Dado que las pruebas de normalidad (Shapiro-Wilk) indicaron que prácticamente ninguna variable cumplía el supuesto de distribución normal, se optó por pruebas no paramétricas. Se aplicó el test de Kruskal-Wallis para detectar diferencias significativas entre clusters en cada variable, seguido de comparaciones post-hoc de Dunn con corrección de Bonferroni para identificar qué pares de clusters presentaban diferencias estadísticamente significativas. Se adoptó un nivel de significancia de $\alpha=0.05$.

4 Resultados

4.1 Selección del número de clusters

El análisis de la curva de inercia mostró una disminución significativa al pasar de 3 a 4 clusters, con una reducción marginal considerablemente menor al incrementar a 5 o más clusters. Los valores de silhouette score resultaron bajos en todos los casos evaluados (rango 0.02-0.04), indicando una estructura geométrica débil de los clusters. No obstante, la solución de k=4 representó el mejor compromiso entre tres criterios: (a) reducción significativa de la inercia, (b) silhouette score aceptable dentro del rango observado, y (c) tamaño suficiente de cada cluster para permitir análisis estadístico confiable e interpretación sustantiva. Con k superiores a 5, algunos clusters contenían muy pocos registros, comprometiendo la robustez del análisis.

4.2 Caracterización de los clusters

La solución de cuatro clusters distribuyó los 62 productores en grupos de tamaño razonablemente equilibrado: Cluster 0 (19 productores, 30.6%), Cluster 1 (11 productores, 17.7%), Cluster 2 (17 productores, 27.4%) y Cluster 3 (15 productores, 24.2%). La Tabla 2 presenta la caracterización de cada cluster según sus variables distintivas.

Tabla 2. Caracterización de los cuatro clusters identificados mediante K-Means.

Cluster	n	Rasgos distintivos
0	19	Mayor superficie de explotación. Bajo asesoramiento externo (cooperativas, empresas, profesionales). Uso de software de gestión por encima del promedio. Baja vinculación institucional.

1	11	Muy baja adopción de maquinaria digital (GPS, sensores, manejo autónomo). Baja conectividad por datos móviles. Formación en áreas no vinculadas al sector agropecuario.
2	17	Mayor vinculación con asesoramiento de organismos públicos, ONG y empresas. Predominio de grupos etarios 45-54 y 65+. Fuerte inserción en redes de asesoramiento.
3	15	Gestión unipersonal predominante. Mayor proporción de varones. Bajo nivel de universitarios completos. Menor adopción de software de gestión y herramientas de análisis de datos.

El análisis detallado de cada cluster revela perfiles productivos y organizacionales claramente diferenciados. El Cluster 0 agrupa a los productores de mayor escala productiva (superficie por encima del promedio), que exhiben una gestión empresarial más formalizada y mayor adopción de software de gestión, pero paradojalmente mantienen baja vinculación con fuentes externas de asesoramiento (cooperativas, empresas consultoras, profesionales independientes). Este perfil sugiere una estrategia de autosuficiencia tecnológica basada en recursos propios, posiblemente asociada a la capacidad económica para contratar personal especializado o adquirir tecnologías directamente.

El Cluster 1 representa el perfil de mayor rezago tecnológico, con muy baja adopción de maquinaria digital (GPS, sensores, manejo autónomo) y conectividad limitada por datos móviles. Un hallazgo particular de este grupo es la sobrerrepresentación de formación en áreas no vinculadas al sector agropecuario, lo que podría indicar que estos productores provienen de trayectorias profesionales diversas y carecen del capital social y técnico específico del sector que facilita la adopción tecnológica.

El Cluster 2 se distingue por su fuerte inserción en redes de asesoramiento institucional, particularmente de organismos públicos (INTA), ONG y empresas de servicios agropecuarios. El predominio de productores de mayor edad (45-54 y 65+ años) en este grupo sugiere que la vinculación institucional se ha construido a lo largo de trayectorias prolongadas en el sector. Este perfil podría representar a productores que, si bien no son los más avanzados tecnológicamente, mantienen una disposición activa hacia la innovación mediada por la confianza en las instituciones de extensión.

Finalmente, el Cluster 3 presenta el perfil de mayor vulnerabilidad desde la perspectiva de la innovación tecnológica: gestión predominantemente unipersonal, mayor proporción de varones en la dirección, bajo nivel de educación universitaria completa y reducida adopción de software de gestión y herramientas de análisis de datos. La gestión unipersonal implica que las decisiones de adopción tecnológica recaen en una sola persona, sin el beneficio de la deliberación grupal y la diversidad de perspectivas que pueden facilitar la innovación. Estos productores representan un segmento crítico para las políticas de extensión, ya que su aislamiento decisional y menor nivel educativo pueden constituir barreras significativas para la adopción.

4.3 Validación supervisada mediante Random Forest

El clasificador Random Forest alcanzó un accuracy promedio en validación cruzada estratificada de 5 folds de 0.79 (±0.11) y un macro F1 de 0.78 (±0.11). Dado el reducido tamaño muestral (n=62, con clusters de entre 11 y 19 observaciones), la validación cruzada opera sobre conjuntos de prueba de aproximadamente 2 a 4 observaciones por cluster, lo que vuelve estas estimaciones poco estables. Para dimensionar esa inestabilidad se aplicó validación cruzada estratificada repetida (5 folds × 40 repeticiones): la distribución resultante del macro F1 tuvo una media de 0.78 pero un intervalo empírico del 95% que abarca de 0.52 a 1.00. En consecuencia, el desempeño puntual de 0.78 no debe interpretarse como una estimación precisa, sino como un valor central rodeado de una incertidumbre considerable propia del tamaño muestral. Con esa salvedad, el resultado indica que los clusters de K-Means son razonablemente reproducibles a partir de las variables originales. La Tabla 3 presenta las variables con mayor importancia para el modelo.

Tabla 3. Importancia de variables en el clasificador Random Forest.

Variable	Importancia
Tipo de gestión (unipersonal/grupal)	0.0932
Asesoría de cooperativas	0.0633
Asesoría de empresas	0.0501
Antigüedad de la empresa	0.0473
Asesoría de organismos públicos	0.0332
Maquinaria con sensores	0.0321
Tipo de acceso al establecimiento	0.0308
Uso de software de gestión	0.0295
Asesoría de ONG	0.0294
Nivel universitario completo	0.0293

Las variables de asesoramiento institucional (cooperativas, empresas, organismos públicos, ONG) concentran conjuntamente una importancia de 0.176, posicionándose como el factor más discriminante entre clusters. El tipo de gestión (unipersonal vs. grupal) emerge como la variable individual más importante (0.093), seguida por la antigüedad y las variables de adopción tecnológica (sensores, software de gestión). Esta convergencia entre las variables que caracterizan los clusters y las que el modelo utiliza para predecirlos refuerza la consistencia de la segmentación.

4.4 Inferencia estadística: prueba de Kruskal-Wallis

Las pruebas de Kruskal-Wallis revelaron diferencias estadísticamente significativas ($p<0.05$) entre clusters en 28 de las variables analizadas. La Tabla 4 presenta las variables con mayor significancia estadística.

Tabla 4. Variables con diferencias significativas entre clusters (Kruskal-Wallis, $p<0.05$).

Variable	H	p-valor
Tipo de gestión	47.009	<0.00001
Maquinaria con GPS	25.844	0.00001
Maquinaria con sensores	23.857	0.00003
Asesoría de cooperativas	20.601	0.00013
Asesoría de empresas	19.703	0.00020
Software de gestión	17.426	0.00058
Asesoría de organismos públicos	16.354	0.00096
Maquinaria con manejo autónomo	15.781	0.00126
Proporción de masculinos	13.846	0.00312
Actividad agrícola	12.091	0.00708

Las variables con mayor significancia estadística corresponden a las dimensiones de gestión organizacional (tipo de gestión, $H=47.009$, $p<0.00001$), adopción de maquinaria tecnológica (GPS, sensores, manejo autónomo) y asesoramiento institucional. Las comparaciones post-hoc de Dunn con corrección de Bonferroni permitieron identificar qué pares de clusters difieren significativamente en cada variable, confirmando que las diferencias no se deben a un único par de clusters atípicos sino que reflejan patrones sistemáticos de diferenciación.

Las variables significativas abarcan múltiples dimensiones del análisis: productivas y organizacionales (tipo de gestión, antigüedad, agricultura, lechería, servicios a terceros), demográficas (grupos etarios 45-54, 55-64 y 65+, proporción de masculinos, nivel educativo), de formación específica (áreas de ciencias económicas), de infraestructura y tecnología (tipo de acceso, datos móviles, registros digitales manuales y automáticos, inversión tecnológica, software de gestión, herramientas de análisis), de asesoramiento (organizaciones públicas, cooperativas, empresas, ONG) y de plataformas y maquinaria (plataformas para compraventa de insumos, maquinaria con GPS, sensores y manejo autónomo). Esta amplitud en las dimensiones con diferencias significativas refuerza la idea de que los clusters capturan perfiles integrales y no simplemente variaciones en una única dimensión.

4.5 Validación con K-Medoids

Para evaluar la estabilidad de la segmentación frente a un cambio de algoritmo de agrupamiento, se realizó un análisis complementario con K-Medoids (PAM, Partitioning Around Medoids), que difiere de K-Means en que utiliza puntos reales del

dataset como representantes de cada cluster (medoides) en lugar de centroides promedio, lo que lo hace más robusto frente a valores atípicos. Al igual que en K-Means, los valores de silhouette de K-Medoids fueron bajos en todo el rango evaluado (máximo de 0.034 en k=2 y 0.030 en k=4), confirmando la ausencia de una estructura geométrica nítida en los datos. Se mantuvo la solución de k=4 para permitir la comparación directa con K-Means, priorizando el criterio de interpretabilidad sustantiva por sobre la métrica interna, dado que ésta no discrimina con claridad entre valores de k.

La solución de K-Medoids distribuyó los 62 productores en cuatro grupos de tamaños 18, 17, 20 y 7. La concordancia entre las particiones de K-Means y K-Medoids, medida mediante el índice de Rand ajustado (ARI), fue de 0.165, valor que indica una coincidencia moderada y superior al azar, pero lejos de una correspondencia uno a uno entre ambas soluciones. Es decir, los dos algoritmos no reproducen los mismos grupos: la asignación individual de productores difiere de manera sustancial según el método. No obstante, las dimensiones que discriminan a los grupos son parcialmente convergentes entre ambos: la antigüedad, la actividad lechera, el tipo de asesoramiento (organismos públicos, ONG, profesionales externos) y el nivel educativo universitario reaparecen como variables distintivas en la solución de K-Medoids, en línea con los ejes identificados por K-Means. Esta divergencia en la composición, junto con la coincidencia en los ejes de diferenciación, es coherente con la interpretación de un espacio continuo de perfiles (ver Sección 5) más que con la existencia de tipos discretos y bien separados.

La concordancia moderada entre algoritmos (ARI = 0.165) debe interpretarse con cautela: no constituye una validación fuerte de la segmentación, sino evidencia de que la partición en cuatro grupos es sensible al criterio de agrupamiento empleado. En conjunto con los bajos valores de silhouette, este resultado indica que los límites entre grupos son difusos y que ninguna solución de clustering individual debe leerse como una taxonomía definitiva. La evidencia más sólida sobre la existencia de diferencias entre productores no proviene de la coincidencia entre algoritmos, sino de las pruebas estadísticas no paramétricas (Sección 4.4), que confirman diferencias significativas en variables sustantivas con independencia de la geometría del agrupamiento.

5 Discusión

Los resultados del presente estudio permiten proponer una taxonomía exploratoria de productores agropecuarios que debe interpretarse con cautela, a la luz de tres limitaciones convergentes: la estructura geométrica débil de los clusters (silhouette próximo a 0.03), la inestabilidad de las métricas supervisadas asociada al tamaño muestral (intervalo del macro F1 entre 0.52 y 1.00) y la concordancia sólo moderada entre algoritmos de agrupamiento (ARI = 0.165 entre K-Means y K-Medoids). Ninguna de estas tres fuentes, por sí sola, constituye evidencia concluyente de la existencia de grupos naturales. En particular, la validación supervisada con Random Forest mide la reproducibilidad de las etiquetas en el mismo espacio de variables que las generó, por lo que no debe leerse como confirmación independiente de la segmentación. La evidencia más robusta proviene de las pruebas no paramétricas de Kruskal-Wallis, que detectan diferencias estadísticamente significativas entre los grupos en numerosas

variables sustantivas con independencia de la geometría del clustering. En conjunto, estos elementos respaldan no la imagen de tipos discretos y bien separados, sino la de un espacio continuo de perfiles productivos con regiones de mayor densidad, cuya utilidad es heurística y orientadora para la política pública más que clasificatoria en sentido estricto.

Los bajos valores de silhouette observados (0.02-0.04) merecen una reflexión particular. En contextos agropecuarios, donde la heterogeneidad es inherente y las transiciones entre perfiles productivos son graduales más que discretas, es esperable que los clusters no presenten fronteras nítidas en el espacio de variables estandarizadas. Esto no invalida la utilidad práctica de la segmentación, siempre que los grupos identificados presenten diferencias estadísticamente significativas en variables relevantes para la toma de decisiones, como efectivamente se demostró mediante las pruebas de Kruskal-Wallis. La analogía apropiada no es la de clusters bien separados como en problemas de clasificación de imágenes, sino la de un espacio continuo de perfiles productivos donde es posible identificar regiones de mayor densidad que corresponden a configuraciones típicas.

El hallazgo de que las variables de asesoramiento institucional constituyen el factor más discriminante entre clusters es consistente con la literatura que posiciona a las redes de extensión y asesoramiento como determinantes críticos de la adopción tecnológica (Rogers, 2003; Pierpaoli et al., 2013). En particular, la diferenciación entre el Cluster 0 (productores grandes con bajo asesoramiento externo pero alto uso de software) y el Cluster 2 (productores vinculados a redes de asesoramiento público) sugiere que existen múltiples trayectorias de adopción tecnológica, no necesariamente lineales ni asociadas exclusivamente al tamaño de la explotación.

El tipo de gestión (unipersonal vs. grupal) emerge como la variable individual más discriminante, lo cual es relevante para el diseño de estrategias de intervención. El Cluster 3, caracterizado por gestión unipersonal, predominio masculino, bajo nivel educativo y reducida adopción de herramientas digitales, representa el perfil de mayor vulnerabilidad tecnológica y, probablemente, el que mayor beneficio podría obtener de programas focalizados de asistencia técnica y capacitación digital.

Los modelos de clasificación supervisada para predecir el nivel de adopción tecnológica alcanzaron precisiones moderadas (F1-macro de 0.67-0.71), lo cual sugiere que las variables consideradas capturan parcialmente la dinámica de adopción pero que existen factores adicionales no relevados en este estudio. Entre las variables más influyentes se encuentran el nivel educativo, la superficie productiva y el tipo de gestión, lo cual es consistente con los determinantes reportados en la literatura (Pivoto et al., 2018; Wolfert et al., 2017).

Un aspecto novedoso del análisis es la evidencia sobre la existencia de múltiples trayectorias de adopción tecnológica. Los modelos clásicos de difusión de innovaciones (Rogers, 2003) tienden a conceptualizar la adopción como un proceso lineal y unidimensional, desde adoptantes tempranos hasta rezagados. Los resultados del clustering sugieren, en cambio, que los productores del departamento Castellanos siguen diferentes caminos hacia la tecnificación: algunos adoptan software de gestión sin vincularse a redes institucionales (Cluster 0), otros se apoyan fuertemente en el

asesoramiento externo como motor de innovación (Cluster 2), mientras que un tercer grupo enfrenta barreras estructurales múltiples que limitan cualquier forma de adopción (Clusters 1 y 3). Esta diversidad de trayectorias tiene implicancias directas para el diseño de políticas de transferencia tecnológica, que deben contemplar múltiples puntos de entrada y estrategias diferenciadas según el perfil del productor.

Es importante señalar las limitaciones del estudio. La primera es el tamaño muestral. Con n=62 y cuatro clusters de entre 11 y 19 observaciones, los métodos de aprendizaje automático operan cerca de su límite inferior de aplicabilidad: la validación cruzada se calcula sobre subconjuntos de pocas observaciones por grupo, lo que se traduce en estimaciones de alta varianza, como evidencia el amplio intervalo del macro F1 (0.52–1.00). La literatura reciente sobre aprendizaje automático con datos limitados advierte explícitamente sobre los riesgos de sobreajuste y de baja generalización en estos regímenes muestrales, y propone enfoques específicos para mitigarlos (Naser, 2026; Hollmann et al., 2025). En consecuencia, los resultados de este trabajo deben entenderse como exploratorios y locales, no como estimaciones poblacionales precisas ni como una taxonomía generalizable. La segunda limitación es el mecanismo de muestreo. El contacto con los productores se canalizó a través de la Sociedad Rural de Rafaela y el INTA Rafaela, lo que introduce un sesgo de conveniencia no aleatorio: los productores vinculados a estas instituciones están probablemente sobrerrepresentados, en particular en el Cluster 2, definido justamente por su alta inserción en redes de asesoramiento institucional. Esto implica que la centralidad del asesoramiento como eje discriminante —uno de los hallazgos principales— podría estar parcialmente inflada por el propio diseño de captación de la muestra, y no reflejar exclusivamente un patrón de la población general de productores del departamento. La validación de este resultado con un muestreo probabilístico e independiente de las instituciones de extensión constituye una prioridad para futuros trabajos. La ampliación del muestreo y la incorporación de un diseño de encuesta con prueba piloto y validación formal del instrumento complementan estas líneas de mejora.

Desde una perspectiva metodológica, el diseño multimétodo empleado en este estudio constituye su aporte principal, aun reconociendo sus límites. La práctica habitual en estudios de clustering agropecuario se limita frecuentemente al análisis de las métricas internas de calidad (inercia, silhouette). En este trabajo, la combinación de clustering no supervisado con clasificación supervisada (Random Forest), inferencia estadística no paramétrica (Kruskal-Wallis con post-hoc de Dunn) y una prueba de estabilidad entre algoritmos (K-Medoids) ofrece un marco de evaluación más completo que el uso aislado de cualquiera de esos componentes. Es relevante destacar que este marco no sólo sirvió para respaldar la segmentación, sino también para acotar honestamente sus límites: la concordancia sólo moderada entre K-Means y K-Medoids fue, en sí misma, un resultado informativo que llevó a reinterpretar los grupos como regiones de un continuo y no como tipos nítidos. Este diseño metodológico, incluida la práctica de reportar las métricas de inestabilidad en lugar de ocultarlas, puede replicarse en otros contextos agropecuarios regionales para generar tipologías comparables y honestas respecto de su incertidumbre.

Los resultados tienen implicancias directas para el diseño de políticas de extensión y transferencia tecnológica en el territorio. La identificación de perfiles diferenciados permite superar el enfoque genérico de las intervenciones y avanzar hacia estrategias

segmentadas. Para el Cluster 0 (productores de mayor escala con baja vinculación institucional pero alto uso de software), las estrategias podrían orientarse a fortalecer la conexión con redes de innovación y facilitar el acceso a tecnologías avanzadas de análisis de datos. Para el Cluster 1 (baja conectividad y digitalización limitada), resulta prioritario abordar las brechas de infraestructura básica (conectividad) como condición habilitante para la adopción tecnológica. El Cluster 2 (fuerte inserción en redes de asesoramiento) puede funcionar como vector de difusión de innovaciones hacia otros grupos, capitalizándose como agentes multiplicadores. Finalmente, el Cluster 3 (gestión individual con bajo nivel educativo) requiere intervenciones integrales que combinen formación básica digital, acompañamiento personalizado y facilitación del acceso a servicios de asesoramiento.

La relevancia del asesoramiento institucional como eje diferenciador entre clusters se alinea con las políticas de extensión agropecuaria promovidas por instituciones como el INTA y las universidades nacionales. Los resultados sugieren que fortalecer y diversificar los canales de asesoramiento (no solo públicos sino también cooperativos, empresariales y de organizaciones de la sociedad civil) puede tener un impacto significativo en la adopción tecnológica del sector. La Universidad Nacional de Rafaela, como institución del sistema educativo y científico del territorio, puede desempeñar un rol articulador entre los distintos actores del ecosistema de innovación agropecuaria.

6 Conclusiones y trabajo futuro

El presente estudio aporta un marco empírico replicable para analizar la diversidad productiva y tecnológica en contextos agropecuarios regionales. La aplicación combinada de clustering no supervisado, validación supervisada e inferencia estadística permitió identificar cuatro perfiles diferenciados de productores en el departamento Castellanos, caracterizados por distintas configuraciones de gestión organizacional, inserción en redes de asesoramiento, adopción de tecnologías digitales y perfil sociodemográfico.

Los hallazgos posicionan al asesoramiento institucional y la adopción digital como ejes centrales de diferenciación, aportando evidencia para el diseño de políticas de extensión y estrategias de innovación adaptadas a la heterogeneidad del territorio. En particular, los resultados sugieren que las intervenciones de transferencia tecnológica deberían: (a) fortalecer las redes de asesoramiento para productores del Cluster 0 (grandes pero desvinculados institucionalmente); (b) priorizar la conectividad y la capacitación digital básica para el Cluster 1; (c) capitalizar la inserción institucional del Cluster 2 como canal de difusión tecnológica; y (d) diseñar programas específicos de formación y acompañamiento para el Cluster 3, que presenta el perfil de mayor vulnerabilidad.

Como trabajo futuro, se proyecta ampliar la base de datos mediante nuevas rondas de relevamiento, incorporar variables adicionales (rentabilidad, acceso al crédito, percepciones sobre tecnología), explorar algoritmos de clustering alternativos (DBSCAN, clustering jerárquico) y desarrollar herramientas de visualización interactiva que faciliten el uso de los resultados por parte de extensionistas y tomadores de decisiones del sector agropecuario regional.

La metodología empleada resulta transferible a otros contextos agropecuarios regionales, permitiendo generar taxonomías comparables que alimenten el diseño de políticas basadas en evidencia. La combinación de aprendizaje no supervisado, validación supervisada e inferencia estadística constituye un marco robusto y replicable que supera las limitaciones de los enfoques puramente descriptivos o basados en categorías predefinidas. En un contexto donde la transformación digital del agro se acelera y la brecha tecnológica entre productores puede profundizarse, contar con herramientas empíricas para segmentar y caracterizar la diversidad productiva resulta indispensable para la eficiencia de los programas públicos orientados al desarrollo rural sostenible.

En síntesis, el estudio aporta tanto al conocimiento científico sobre la adopción tecnológica agropecuaria como a la práctica de la extensión y la política pública, demostrando que las técnicas de aprendizaje automático pueden complementar y enriquecer los enfoques tradicionales de diagnóstico y planificación territorial en el sector agropecuario.

Referencias

Hollmann, N., Müller, S., Purucker, L. et al. (2025). Accurate predictions on small data with a tabular foundation model. Nature, 637, 319–326.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666.

Kitchenham, B. A. y Pfleeger, S. L. (2002). Principles of survey research (parts 2, 4 y 5). ACM SIGSOFT Software Engineering Notes, 27(1), 18–20; 27(3), 20–23; 27(5), 17–20.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S. y Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674.

Naser, M. Z. (2026). A review of machine learning with small and limited data. Journal of Big Data, 13, 18.

Lowder, S. K., Skoet, J. y Raney, T. (2016). The number, size, and distribution of farms, smallholder farms, and family farms worldwide. World Development, 87, 16–29.

Pierpaoli, E., Carli, G., Pignatti, E. y Canavari, M. (2013). Drivers of precision agriculture technologies adoption: A literature review. Procedia Technology, 8, 61–69.

Pivoto, D., Waquil, P. D., Talamini, E., Finocchio, C. P. S., Dalla Corte, V. F. y de Vargas Mores, G. (2018). Scientific development of smart farming technologies and their application in Brazil. Information Processing in Agriculture, 5(1), 21–32.

Rogers, E. M. (2003). Diffusion of Innovations (5th ed.). Free Press.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

Wolfert, S., Ge, L., Verdouw, C. y Bogaardt, M. J. (2017). Big data in smart farming – A review. Agricultural Systems, 153, 69–80.