DIOTIMA
REVISTA CIENTÍFICA DE ESTUDIOS TRANSDISCIPLINARIA
Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

Modelado de usos del suelo con k-medias para la conservación territorial en Almoloya de Juárez

land use modelling with k-means for territorial conservation in Almoloya de Juárez

Raúl Salazar Godínez[1]

rsalazarg003@alumno.uaemex.mx

ORCID: 0009-0004-5242-1197

Carolina Herrera Mendoza

cherreram@uaemex.mx

ORCID: 0009-0005-2039-8532

Rosa María Valdovinos Rosas

rvaldovinosr@uaemex.mx

ORCID: 0000-0001-9954-0653

[1] Autores de la Ingeniería en Computación, Ciudad Universitaria UAEM COATEPEC, Facultad de ingeniería Universidad Autónoma del Estado de México.

Resumen

El presente estudio explora las aplicaciones de la Minería de Datos en el análisis del uso suelo para comprender las formas de reconfiguración y expansión urbana, para tal efecto, este articulo emplea el algoritmo de clasificación *K*-medias para la identificación de patrones de los usos del suelo de las áreas no urbanizables en el contexto del territorio mexicano, tomando como referencia el contenido previsto en la Tabla de Uso del Suelo del Plan Municipal de Desarrollo Urbano (PMDU) de Almoloya de Juárez vigente publicado en 2022 por la Secretaría de Desarrollo Urbano e Infraestructura del Estado de México. Los resultados sugieren que el modelo prioriza la clasificación de las áreas boscosas, maximizando la detección de áreas clasificadas como Natural Bosque No Protegido (N-BOS-N), incluso a costa de una posible restricción excesiva de tierras agrícolas.

***Palabras clave*:** Zonificación del suelo, Planeación Urbana, Áreas no urbanizables, *k*-medias

Abstract

This study explores the applications of Data Mining in land use analysis to understand the forms of urban reconfiguration and expansion. To this end, this article uses the K-means classification algorithm to identify land use patterns in non-urbanizable areas in the Mexican territory context. The study used the content provided in the Land Use Table of the current Municipal Urban Development Plan of Almoloya de Juárez published in 2022 by the Secretariat of Urban

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

Development and Infrastructure of the State of Mexico. The results suggest that the model prioritizes the classification of forest areas, maximizing the detection of areas classified as Natural Unprotected Forest (N-BOS-N), even at the cost of a possible excessive restriction of agricultural land.

***Keywords*:** Zoning Land, Urban Planning, Non-Urban areas, *k*-means

Fecha de envío: 21/05/2025

Fecha de aprobación: 18/07/2025

Fecha de publicación: 01/09/2025

Introducción

En la era digital, la generación de datos aumenta a un ritmo sin precedentes donde dada su complejidad y cantidad resulta necesaria la utilización de herramientas computacionales que permitan el procesamiento y análisis de datos estructurados y no estructurados. Ante este tenor en materia de Planeación Urbana (PLU) a consecuencia de la rápida urbanización y crecimiento poblacional en las ciudades, un elemento crucial para la optimización de recursos humanos, financieros y computacionales es el empleo de técnicas de juicio asistido como el aprendizaje automático para atender las problemáticas apremiantes de las ciudades (Peña, 2021).

La PLU y la gestión del uso del suelo son dos elementos de gran relevancia para contribuir los Objetivos de Desarrollo Sostenible (ODS), especialmente en el No. 11 cuyo propósito se centra en las Ciudades y comunidades sostenibles; en este sentido, además de contribuir con la agenda internacional propuesta de la Organización de las Naciones Unidas (ONU), evidencia científica

pone de manifiesto que la adecuada planificación de los usos del suelo permiten promover la cohesión social, la identidad histórica y la conservación del patrimonio natural y ambiental (Ramírez, Cárdenas y Alegría, 2024).

De manera específica, el Municipio de Almoloya de Juárez ubicado en el Estado de México, es una región de gran interés debido a su cercanía con la Capital de la Entidad (Toluca de Lerdo), que en los últimos años derivado de las políticas de vivienda (Política Estatal de Ciudades del Bicentenario) ha presentado un fenómeno de creciente urbanización lo que ha generado una fuerte influencia para el desarrollo de actividades industriales, provocando una tensión constante en la disminución de actividades agrícolas y de protección de las áreas naturales, así como también, experimentando una notable transformación del uso del suelo (SEDUI, 2022).

El Plan Municipal de Desarrollo Urbano (PMDU) de Almoloya de Juárez establece las bases para la PLU del Municipio y la gestión de los usos del suelo, en el cual también se establecen las políticas y estrategias de conservación del territorio principalmente de las áreas no urbanizables (SEDUI, 2022). Cabe señalar que el eje central de análisis del presente estudio comprende la aplicación de técnicas de aprendizaje no supervisado, siendo una de ellas el algoritmo K-medias para la identificación de patrones de uso del suelo, centrando su atención en las áreas catalogadas como Áreas No Urbanizables para fortalecer la literatura en la preservación de estas zonas, evitar daños al ecosistema, biodiversidad y la degradación del suelo.

Trabajos relacionados

La aplicación de algoritmos de aprendizaje automático en el campo del urbanismo ha adquirido popularidad en el análisis de la ocupación del suelo utilizando métodos de clasificación, cuyo

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

objetivo se ha centrado en comprender bajo un enfoque sistémico y multidimensional la manera en cómo se articulan las ciudades (Peña, 2021). Adicional a lo anterior, la literatura ha permitido evidenciar la importancia de aplicar técnicas de aprendizaje no supervisado para la clasificación en la zonificación del suelo, ya que estas técnicas además de auxiliar en una adecuada toma de decisiones, permite la integración de datos de diversas fuentes de información para un análisis integral de la realidad (Zhou, Gu, Shen, Ma, Miao, Zhang y Gong, 2017). Retomando las ideas de Steurer y Bayr (2020) la expansión urbana de los asentamientos humanos se constituye como un problema multidimensional y complejo en las ciudades, cuyas implicaciones impactan significativamente en el cambio de patrones de uso del suelo. Así mismo, existe evidencia de que la expansión urbana ha puesto en peligro el sistema de sostenibilidad de las ciudades generando consecuencias negativas en el contexto ambiental, social, de salud, entre otros (Liu, Peng, Wu, Jiao, Yu y Zhao, 2018).

En el contexto mexicano se puede observar que los efectos negativos sobre el cambio del uso del suelo en las ciudades, se asocian en contextos donde, por decir algunos ejemplos, el desarrollo de actividad humanas como la ganadería han provocado que los bosques desaparezcan en una tasa de 79% con una extensión de hasta 2,672 km^2 cuadrados cada año (SEMARNAT, s/f), escenario que de no cambiarse o solucionarse provocará afectaciones en el medio ambiente como degradación del suelo o perdida de la cobertura vegetal. Para minimizar los efectos negativos de una mala planeación urbana, la literatura muestra que se han utilizado diversos algoritmos de clasificación para el análisis de la zonificación del suelo, siendo uno de los más populares el algoritmo *k*-medias, cuyas aportaciones a la ciencia en el ámbito de la zonificación del suelo se han centrado en lo siguiente:

Tema central de análisis	Datos utilizados	Líneas abiertas de investigación
Medición de la expansión urbana utilizando información sobre usos del suelo	Baja densidad, dispersión y baja compacidad de la forma de la ciudad utilizando el conjunto de datos Corine Land Cover	Gracias a la aplicación de los índices de densidad, entropía y autocorrelación espacial se mejoran los procesos de medición tradicionales utilizando SIGs
Medición e identificación de patrones típicos de expansión urbana considerando las actividades humanas que son distintas de las físicas y/o medio ambientales aplicando métodos distintos a los de teledetección	Patrones de distribución de la densidad de población: área, área de borde, área rural, área de expansión y área ecológica interna. Se contemplan datos de uso del suelo como: edificios urbanos y rurales, áreas forestales, praderas	Los resultados sugieren que para identificar la expansión urbana a partir de la agrupación de K-medias, se realice un análisis de población cuadriculada aplicando la entropía espacial local (FCA)

Tabla 1: Estudios sobre zonificación del suelo aplicando k-medias (Steurer y Bayr, 2020) (Liu, Peng, Wu, Jiao, Yu y Zhao, 2018).

Atendiendo a las líneas abiertas de estudio identificadas en la literatura, resulta esencial la aplicación de técnicas de Aprendizaje Automático con técnicas de modelado espacial con el propósito de predecir y generar información empírica más robusta que favorezca la Planeación

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

Urbana en diversos contextos territoriales y que permitan comprender el fenómeno de la expansión urbana y los cambios de uso del suelo.

Materiales y métodos

Para la elaboración del presente estudio se utilizó información del Repositorio Institucional de la Secretaría de Desarrollo Urbano e Infraestructura (SEDUIS) del Estado de México, centrando el análisis en el contenido de la Tabla de Uso del Suelo del Plan Municipal de Desarrollo Urbano (PMDU) de Almoloya de Juárez (SEDUI, 2022). Considerando el universo de información publicado por dicha institución, para el procesamiento de los datos se utilizó como herramienta *General Refine Expression Languaje* y Visual Studio Code bajo un lenguaje de programación en Python, considerando la metodología Knowledge Discovery in Databases (KDD) pues permite contar con un enfoque estructurado para asegurar una adecuada limpieza, análisis de datos e interpretación de los resultados obtenidos de la aplicación del Algoritmo k-Medias.

Los usos del suelo en áreas no urbanizables previstos en la Tabla de Usos del Suelo vigente en Almoloya de Juárez (SEDUI, 2022), considera los siguientes destinos (Observe Tabla 2):

Uso específico	Clave	Superficie Municipal (Ha)
Residencial Campestre con uso Habitacional Densidad Tres Mil Trescientos Treinta y Tres	RC-H-333-A	535.25
Natural Bosque No Protegido	N-BOS-N	4,660.64
Natural Bosque Protegido	N-BOS-P	2,676.11
Parque Natural No Protegido	N-PAR-N	223.48
Parque Natural Protegido	N-PAR-P	2,780.95
Ecoturístico	ECO-T	905.05
Agrícola de Mediana Productividad Temporal	AG-MP-T	25,231.02
Agroindustria	AGR-I	530.79
Mina a Cielo Abierto	M-C-A	12.59
Cuerpos de Agua	C-A	1,378.67

Tabla 2: Destinos y características de los del suelo no urbanizables en Almoloya de Juárez (elaboración propia con base en SEDUI, 2022.

Bajo otra orden de ideas, considerando la metodología planteada anteriormente, el proceso de recolección, análisis y transformación de la información realizado en este estudio puede observarse en la Ilustración 1.

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

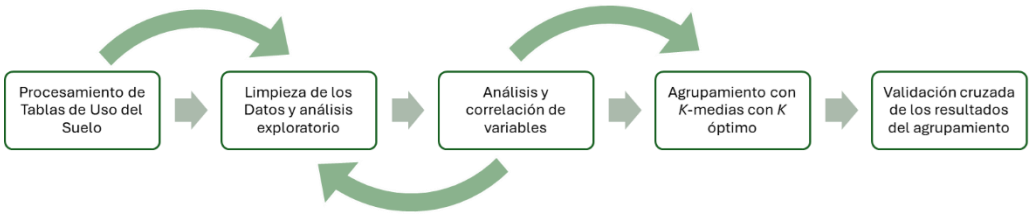


Ilustración 1. *Proceso metodológico para el desarrollo de la investigación (elaboración propia, 2025)*

Tratamiento de los datos

Debido a la naturaleza y formato de la información obtenida a través de la Tabla de Uso del Suelo de Almoloya de Juárez, para facilitar el análisis y manipulación de la información, en un proceso inicial fue necesaria la transformación del archivo original (PDF) a un formato CSV para la reorganización y depuración de datos inconsistentes dentro de nuestro conjunto de datos. Dado que los datos originales contaban con diversos problemas que dificultaban la ejecución del Algoritmo *K*-medias (presencia de datos categóricos y numéricos), fue necesario realizar un preprocesado de datos, en el que se incluye la conversión del formato de los datos, así como tratamiento de complejidades que se señala a continuación.

Conversión de archivos PDF:

Como bien se señaló anteriormente, la naturaleza original de los datos obtenidos para la conformación de nuestro conjunto de datos se encontraba inicialmente en un formato PDF, por lo que, para el análisis y procesamiento de estos, Open Refine no era una herramienta viable, por lo

tanto, se emplearon diversas bibliotecas de Python para hacer posible la lectura de la Tabla de Usos del Suelo (Camelot, PDF Plumber). Gracias a eso, se logró el proceso de transformación, generando así, archivos en formato CSV para su posterior lectura en la herramienta Open Refine. La generación de los archivos CSV se elaboró de manera separada, es decir, para mantener la integridad absoluta de los datos se hizo un archivo por tabla, de modo que se obtuvieron 7 tablas diferentes con extensión de renglones variada. Dado que no era óptimo evaluar y analizar la información por separado, se creó un script en Python que permitió la fusión de cada tabla, por supuesto sin la pérdida de información, paso esencial para trasladar la siguiente fase del proyecto a Open Refine.

Limpieza de los datos y análisis exploratorio

En virtud de que los datos estaban conformados por atributos tanto categóricos como numéricos, se emplearon técnicas de codificación de etiquetas y normalización; así mismo, se identificaron patrones atípicos como atributos duplicados o instancias mal etiquetadas, que generaban inconsistencias en la lectura de cada dato correspondiente a la Tabla de Usos.

Para los valores nulos y/o faltantes, se reemplazaron los valores vacíos por el término “ZX” por criterio de selectividad, es decir, la simbología mostrada en los archivos PDF, las formas de datos original del conjunto de datos, indicaba un uso de instancias diversas en formato textual, de modo que resultó adecuado ese nombramiento para evitar confusiones en la lectura de datos. Esto se llevó a cabo en las mismas funciones que ofrece Open Refine, a través de la selección de una columna deseada, se transformó la forma de las instancias de cada renglón por medio del uso de

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

expresiones GREL usando una lógica de programación similar a Python para su correcta aplicación.

Para los patrones atípicos, aquellos elementos vacíos o nulos, se desarrolló el mismo procedimiento de reetiquetado en instancias específicas, para eso, se utilizaron pequeños scripts en GREL para identificar adecuadamente aquellas instancias a renombrar. El procedimiento fue similar con el inconveniente de la duplicidad, pues con base a scripts específicos de selección se eliminaron y renombraron instancias determinadas. Todo este proceso con la utilización de Open Refine, su análisis selectivo a través de la línea de comandos de la aplicación abarcó la sección del estudio y solución de la complejidad de los datos por medio de expresiones GREL.

_K_-medias:

La literatura señala que en Aprendizaje No Supervisado, existen diversos algoritmos de clasificación dentro, en el cual *k*-medias es uno de los más populares para el análisis, transformación y procesamiento de los datos. A lo largo de los años este algoritmo ha evolucionado y se ha adaptado a diversas aplicaciones, desde la segmentación de clientes hasta ámbitos más especializados como el análisis de imágenes pues su eficacia ha convertido al algoritmo en una herramienta esencial para aplicarse en diversos campos multidisciplinarios. El funcionamiento del algoritmo *K*-medias opera de la siguiente manera:

Algoritmo 1 K- Medias

Inputs: Conjunto de datos $X = \{x_1 \dots x_n\}$, No. de clústeres k, No. máximo de iteraciones
Outputs: Los k centros de clústeres
Inicializar aleatoriamente los centros de los clústeres $\{c_1 \dots c_k\}$
for iteración $= 1$ hasta max_iter **do**
  for cada punto $x_i \in X$ **do**
    Asignar x_i al clúster más cercano a su centro c_j, donde $j = \arg\min_{1 \leq j \leq k} \|x_i - c_j\|$
  end for
  for cada clúster $j = 1 \dots k$ **do**
    Recalcular el centro de $c_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i$, donde S_j es el conjunto de puntos asignados al clúster j
  end for
  if los centros de los clústeres no cambian **then**
    Salir
  end if
end for
Retornar los centros C y las asignaciones de clústeres

Validación de resultados

Para el análisis de resultados se partió de una matriz de confusión, como la mostrada a continuación:

	Clase +	**Clase -**
Clase +	VP	FN
Clase -	FP	VN

Donde:

VP (Verdadero Positivo), y VN (Verdadero Negativo) son las instancias bien reconocidas por el modelo, en tanto que los errores se representan por FN (Falso Negativo) y FP (Falso Positivo).

De esta matriz se obtuvieron varias de las métricas ampliamente utilizadas en aprendizaje automático y minería de datos para análisis de resultados.

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

1	**Exactitud:** Cantidad de predicciones que fueron correctas	$Exactitud = \frac{VP + VN}{VP + VN + FP + FN}$
2	**Precisión:** Proporción de VP identificados por el modelo.	$Precisión = \frac{VP}{VP + FP}$
3	**Sensibilidad o Recall:** Proporción de VP correctamente identificados por el modelo	$Sensibilidad = \frac{VP}{VP + FN}$
4	**Especificidad:** Proporción de VN correctamente identificados por el modelo	$Especificidad = \frac{VN}{VN + FP}$
5	**F1 Score:** Medida armónica entre precisión y sensibilidad	$F1\ Score = 2\left(\frac{(Precisión)(Sensibilidad)}{Precisión + Sensibilidad}\right)$

Resultados

Una vez solucionadas las complejidades de datos, el proceso siguiente consistió en identificar la correlación existente entre una variable y otra a través del ploteo de los datos, procedimiento que consistió en la obtención de medidas estadísticas como el total de elementos por columna, media, mediana, desviación estándar, métricas necesarias para la evaluación de los pasos posteriores.

Matriz de correlación

El objetivo de esta sección fue determinar la distribución de los datos, la visualización de su comportamiento con demás variables, y en general, la correlación entre variables para deducir la matriz de correlación por cada columna del conjunto de la data set. Debido a que se contaba con datos numéricos y tipo cadena, estos últimos se etiquetaron con números negativos, ya que de no hacerlo producían errores en las operaciones de tipo NaN (Not a Number). La Ilustración 2 muestra la matriz resultante posterior al emplear un análisis de correlación por medio del "*coeficiente de correlación de Pearson*".

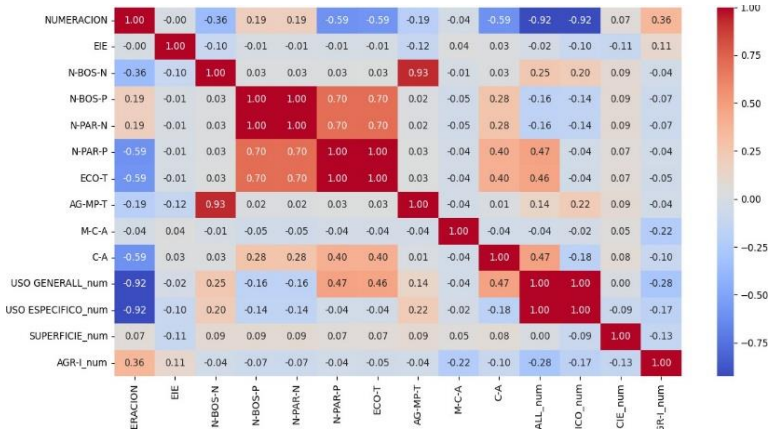


Ilustración 2. *Matriz de correlación de los datos numéricos (elaboración propia, 2025)*

Los resultados de la Matriz de correlación (Ilustración 2) muestra la relación entre cada variable numérica, donde aquellos más cercanos a 1 indican una fuerte correlación positiva. Por su contra parte, los valores cercanos a -1 muestran una fuerte correlación negativa. Igualmente existieron casos cercanos a cero entre variables. Se tomaron dos elementos numéricos con una correlación muy fuerte y significativa respecto a una variable de la otra, evitando

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

multicolinealidad, evitando una alta correlación entre variables para evitar redundancia, distorsión de resultados y la dificultad de interpretación.

Al observar la matriz de correlación, se identificaron 14 variables con una correlación igual o superior al 70%, más específicamente, se crearon grupos para relacionar las variables, de modo que, se categorizaron de la siguiente manera:

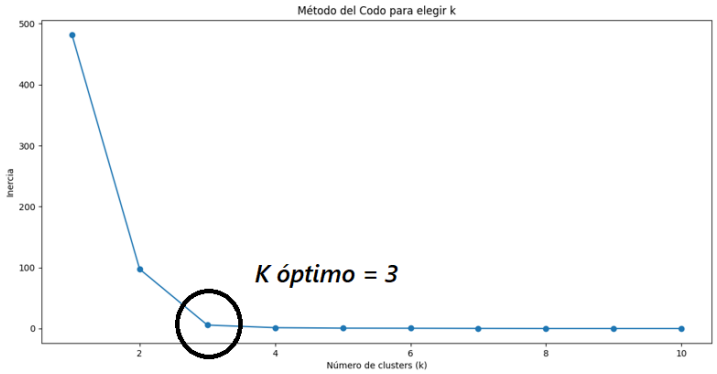
Grupo	Variable 1	Variable 2	Nivel de correlación
1	N-BOS-N	AG-MP-T	0.93
2	N-BOS-P	N-PAR-N	1.00
3	N-PAR-P	ECO-T	1.00
4	N-BOS-P	N-PAR-P	0.70
5	N-PAR-P	N-PAR-N	0.70
6	N-BOS-P	ECO-T	0.70
7	N-PAR-N	ECO-T	0.70

Tabla 2: Tabla de clasificación de las variables candidatas

Considerando lo anterior, tras analizar el nivel de correlación de cada grupo se determinó que el Grupo 1, contaba con las variables más idóneas, pues los demás grupos, pese que en algunos casos mostraban niveles de correlación con una aparente relación prometedora (Grupo 2 y Grupo 3), sus variables presentaban valores numéricos negativos, por tal motivo, su uso no fue posible a pesar de haber sido transformadas pues su verdadera naturaleza impidió trabajarse con ellas.

Agrupamiento de datos

Una vez comprendida la naturaleza de los datos, y determinadas las instancias candidatas "N-BOS-N" y "AG-MP-T" se entrenó el algoritmo *k*-medias. Para determinar el valor de *k,* se utilizó el "Método del Codo" o "Elbow Method", variando el valor de *k* de 2-10 grupos (Ilustración 3).



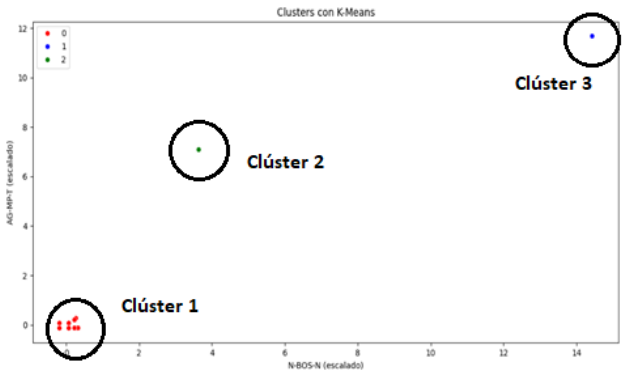
lustración 3. Diagrama del Método del Codo (Elbow Method) en k=3 *(elaboración propia, 2025).*

El objetivo fue determinar el valor de k donde la disminución de la inercia fuera más lenta y cercana al eje X, formando un "codo" en la gráfica. En la ilustración 3 se observa una disminución muy rápida de la inercia desde k=1 hasta k=3. A partir de k=3, la disminución de la inercia se vuelve mucho más gradual, casi plana, de modo que el valor óptimo de k hallado fue 3. Con k=3, los datos se agruparon en clústeres bien definidos, concluyendo que, aumentar el número de clústeres no aporta una mejora significativa en la reducción de la inercia.

Con este valor óptimo k=3, el algoritmo K-Medias generó el algoritmo buscó similitudes entre los datos, generando tres agrupamientos bien definidos con las instancias candidatas "N-BOS-N" y "AG-MP-T". Bajo esta orden de ideas, la función del algoritmo consistió en medir la

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

distancia entre los puntos, agrupando aquellos que se parecen en las dos instancias candidatas. Caso contrario si hay una diferencia muy alta, los puntos se van a distintos grupos



***lustración 4. Distribución de los Clústeres** (elaboración propia, 2025).*

La información representada en la Ilustración 4 muestra que el Clúster 1 agrupa puntos en el cual prevalecen zonas Naturales Boscosas No Protegidas, el Clúster 2 reúne áreas destinadas para actividades Agrícolas de Mediana Productividad Temporal y finalmente el Clúster 3 contiene valores con baja densidad de zonas agrícolas y predominio de áreas naturales. Esta representación revela patrones en los datos que no son visibles a simple vista, permitiendo entender cómo se distribuyen los tipos de suelo o zonas según sus características compartidas. Además, la similitud de cada grupo, así como la separación entre ellos refuerzan la validez del agrupamiento.

Aunque el método del codo identificó que el valor óptimo de k=3 agrupa correctamente la diversidad estructural de los datos, surge una situación importante: en este caso solo existen 2 clases reales, las instancias candidatas, pero se generaron 3 clústeres predichos. Este desfase es aceptable para un análisis descriptivo no supervisado, sin embargo, impide el uso directo de métricas supervisadas como "Exactitud", "Precisión", "Sensibilidad" etc. ya que dichas métricas

exigen que el número de clases reales coincida con el número de clases predichas (es decir, una matriz cuadrada NxN).

Bajo esta coyuntura, el estándar ISO/IEC 25024 citado en Bishop (2006), la precisión se calcula sobre casos correctamente clasificados, lo cual también supone igualdad entre clases reales y predichas; sin embargo, es totalmente válido utilizar un valor menor de k cuando se desea aplicar métricas supervisadas. El valor óptimo de k que sugiere el método del codo no es una regla rígida, sino una guía que puede ajustarse según el objetivo del análisis (Jain, 2010; Müller & Guido, 2016).

Para dar cumplimiento a los fines del presente estudio, se utilizó el valor de k=2; esto permitió construir una matriz de confusión válida y aplicar métricas supervisadas sin ambigüedad. Los resultados obtenidos de la aplicación del añgoritmo k-means arrojo los siguientes resultados:

- Clúster 0 y 2 = se agruparon como Clase 1, correspondiente a "N-BOS-N"
- Clúster 1 = se asignó como Clase 0, correspondiente a "AG-MP-T"

Aunque el algoritmo agrupó los datos inicialmente en 3 clústeres (k=3), se observó que la mayoría de los registros pertenecientes a "N-BOS-N" se concentraban en los clústeres 0 y 2, mientras que los registros de "AG-MP-T" estaban principalmente en el clúster 1.

Por tanto, se realizó una reasignación supervisada de clústeres, fusionando los clústeres con comportamiento similar, reduciendo así a 2 grupos, permitiendo generar una matriz de confusión de 2x2:

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

Real\Predicho	Clase A (Clústeres 0 y 1)	Clase B (Clúster 2)	Total
N-BOS-N (Clase 1)	9+1 = 10 (VP)	2 (FN)	12
AG-MP-T (Clase 2)	4+1 = 5 (FP)	2 (VN)	7
Total	15	4	19

Con base a la matriz de confusión elaborada, se calcularon las métricas supervisadas siguientes

Exactitud: el modelo clasifica bien el 63% de los casos (sean "N-BOS-N" o "AG-MP-T").

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN} \quad \frac{VP + VN}{VP + VN + FP + FN} = \frac{10 + 2}{10 + 2 + 5 + 2} = \frac{12}{19} \approx 0.6316$$

Precisión: cuando el modelo predice "N-BOS-N", acierta el 66% de las veces.

$$Precisión = \frac{VP}{VP + FP} \quad \frac{VP}{VP + FP} = \frac{10}{10 + 5} = \frac{10}{15} = 0.6667$$

Sensibilidad (Recall a+): de todos los verdaderos "N-BOS-N", el modelo identifica el 83%.

$$Sensibilidad = \frac{VP}{VP + FN} \quad \frac{VP}{VP + FN} = \frac{10}{10 + 2} = \frac{10}{12} \approx 0.8333$$

Especificidad (Tasa negativa verdadera a-):

$$\frac{VN}{VN + FP} = \frac{2}{2 + 5} = \frac{2}{7} \approx 0.2857$$

Media Geométrica:

$$Media\ geométrica = \sqrt{Sensibilidad \cdot Especificidad} = \sqrt{0.8333 \cdot 0.2857} \approx 0.4879$$

F1 Score:

$$F1\ Score = 2\left(\frac{(Precisión)(Sensibilidad)}{Precisión + Sensibilidad}\right) \quad 2 \cdot \frac{0.6667 \cdot 0.8333}{0.6667 + 0.8333} = 2 \cdot \frac{0.5555}{1.5} \approx 0.7407$$

De este modo se tiene:

Métrica	Valor	Interpretación
Exactitud	0.6316	Acertó en 63.16%
Precisión	0.6667	De los positivos predichos, 66.67% eran correctos
Sensibilidad	0.8333	Detectó 83.33% de los reales positivos
Especificidad	0.2857	Solo 28.57% de negativos fueron bien clasificados
F1 Score	0.7407	Buen balance general
Media Geométrica	0.4879	Afectada por baja especificidad

lustración 5. Resultados de las métricas empleadas para las variables "N-BOS-N" y "AG-MP-T" *(elaboración propia, 2025).*

Calidad del agrupamiento

Con la utilización de la validación cruzada con 5 particiones se evaluó el rendimiento del modelo con el objetivo principal de garantizar que no esté sobre ajustado, es decir, que no aprenda patrones específicos del conjunto de datos en lugar de generalizar bien a datos nuevos.

La forma en que se utilizó la validación cruzada fue aplicando "K-Fold Cross Validation", dividiendo los datos en N subconjuntos o particiones, ejecutando K-medias en cada uno.

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

Algoritmo 3 *Validación Cruzada K-Fold para K-medias*

Inputs: Datos normalizados $X = \{x_1 \dots x_n\}$, No. de clústeres k, No. máximo de iteraciones, No. particiones K (K-Fold)
Outputs: Promedio del índice de silueta
Dividir los datos X en K subconjuntos (folds)
Para cada fold i=1 hasta *K*:
Usar K-1 folds como entrenamiento x_{train}, 1 como prueba x_{test}
Aplicar K-Means sobre x_{train} con k clústeres:
Predecir etiquetas de clúster en x_{test}:
Si hay más de un clúster en x_{test}:
    Calcular índice de silueta y guardarlo
Calcular el promedio de todos los índices de silueta válidos
Retornar el promedio como métrica de calidad de agrupamiento

Para medir la calidad del agrupamiento se calculó el coeficiente de silueta, el cual está dado por:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

Donde:

a (i): Distancia promedio entre el punto i y todos los demás puntos del mismo clúster

b (i): Distancia promedio entre el punto i y todos los puntos del clúster más cercano (al que no pertenece)

Los valores del coeficiente silueta pueden variar de -1 a 1, si el valor es cercano a 1, los clústeres están bien definidos y separados, sin embargo, un valor cercano a 0 indica que están muy mezclados. Por el contrario, un valor negativo indica que hay datos asignados al clúster incorrecto. Si s(i) ≈ 1, el punto está bien agrupado y lejos de otros clústeres

Si s(i) ≈ 0, el punto está entre dos clústeres

Si s(i) < 0, el punto puede estar mal asignado al clúster

En el agrupamiento obtenido se obtuvo un coeficiente de silueta promedio de 0.913, esto indica que los datos están bien agrupados, sólidos y definidos de manera consistente, refiriendo a una calidad excelente en la separación de los clústeres. Este valor refleja que los terrenos agrupados comparten similitudes claras entre sí, así como su correcta clasificación en diferentes grupos.

En el contexto actual donde se busca distinguir bosques naturales no protegidos de zonas agrícolas, este resultado implica que el modelo tiene una alta capacidad de identificar patrones ambientales reales, reduciendo así la posibilidad de clasificaciones erróneas. Por tanto, el agrupamiento es válido, útil y confiable para buscar la conservación de bosques y gestión de zonas de conservación.

Discusión

Al revisar los resultados previstos en la sección anterior, se observó que los clústeres 0 y 2 estaban conformados mayoritariamente por instancias de la clase “N-BOS-N", por lo que se reagruparon como Clase 1. El clúster 1, en cambio, se asoció con instancias de la clase “AG-MP-T", asignándolo como Clase 0, ante este escenario el Algoritmo k-medias ha evidenciado su utilidad para identificar patrones y agrupar instancias de los usos del suelo de áreas no urbanizables en estos dos grupos; no obstante, los resultados de las métricas aplicadas exigen una interpretación meticulosa especialmente cuando se trata de diferenciar entre las diversas clases existentes en escenarios donde existe desbalance.

Particularmente, la Sensibilidad ha evidenciado que el modelo implementado ha tenido un alto porcentaje de efectividad al identificar casos verdaderos en la Clase N-BOS-N, lo cual es un

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

elemento crucial para para emplear políticas de preservación y/o conservación de áreas naturales no protegidas.

Por su parte, el resultado obtenido de la especificidad refleja que el modelo tiene mayor tendencia a clasificar usos del suelo N-BOS-N, escenario que bajo el contexto urbano puede representar un gran desafío para identificar adecuadamente las áreas agrícolas en el territorio, ya que, considerando la información prevista en el PMDU de Almoloya de Juárez, existen zonas cuyo destino agrícola se encuentran altamente sometidas a la presión de la expansión urbana al ser categorizadas espacialmente a largo plazo como Áreas Urbanizables (SEDUI, 2022), dicha situación pone de manifiesto la necesidad de analizar el uso del suelo bajo un enfoque complementario al normativo que permita distinguir los diversos usos destinados a las áreas no urbanizables en el territorio.

A partir de ello, se observan las siguientes fortalezas y debilidades en la aplicación del algoritmo k-medias:

Fortalezas:

- El modelo muestra una alta sensibilidad (83%) detectando correctamente la mayoría de los terrenos de tipo "N-BOS-N".
- Buen F1 Score (74%) indicado un equilibrio razonable entre “Precisión” y "Exactitud” motivo que muestra un modelo funcional.

Debilidades:

- La especificidad es baja (28%), por lo que el modelo presenta dificultades para detectar correctamente los terrenos agrícolas, confundiéndolos como si fueran bosques no protegidos.
- La media geométrica (48%), confirma que el rendimiento no es parejo entre ambas clases.

Para comprobar la consistencia del modelo, se aplicó Validación Cruzada K-Fold, dividiendo el conjunto de datos en cinco grupos. En cada iteración, se entrenó el modelo con cuatro grupos y se validó con el restante. La métrica de calidad utilizada fue el "Coeficiente de Silueta", obteniendo un valor de 0.91, lo que confirma que los clústeres están bien definidos y separados.

El modelo ofrece un desempeño aceptable, en especial para detectar correctamente los terrenos N-BOS-N, lo cual es crítico para la protección de áreas naturales no protegidas, mostrando desequilibrio entre sensibilidad y especificidad, este análisis adquiere mayor sentido al interpretar los errores de la naturaleza de los datos. Considerando lo anterior, los resultados permiten proporcionar información clave sobre los problemas que requieren mayor atención.

Conclusiones

El presente estudio ha logrado demostrar efectividad en la implementación del algoritmo k-medias, logrando clasificar los diversos usos del suelo previstos en la Tabla de Usos del Suelo de Almoloya de Juárez, permitiendo la comprensión en la distribución de los usos del suelo a nivel normativo. Así mismo, los resultados arrojan que la sensibilidad y especificidad presentan limitaciones importantes dada la falta de equilibrio entre sus valores, reflejando un posible desbalance de clases.

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

El análisis realizando, logró identificar patrones en los diversos usos del suelo de las áreas no urbanizables de Almoloya de Juárez, estos usos del suelo consideran una amplia variedad de categorías; no obstante, tras realizar las tareas de clasificación, los clústeres generados arrojaron resultados más relevantes en los usos AG-MP-T y N-BOS-N; la exclusión de los otros usos del suelo como Mina a Cielo Abierto, Cuerpos de Agua, Ecoturístico, entre otros, reflejan patrones de distribución que no son suficientemente representativos por el algoritmo, cuyas características normativas están inmersas dentro de las dos clases señaladas anteriormente (AG-MP-T y N-BOS-N).

De los resultados obtenidos se puede afirmar que, el modelo presenta un rendimiento razonable, sin embargo, existen limitaciones para clasificar normativamente las zonas destinadas para uso Agrícola, sin embargo, en términos de extensión territorial, pese la gran extensión que presenta el uso del suelo AG-MP-T, el modelo subestima la importancia de dicha clase.

Considerando lo anterior, para mejorar la eficacia en la aplicación del modelo, particularmente en la capacidad para discriminar entre los dos usos del suelo identificados, se sugiere emplear algunas otras técnicas adicionales bajo un enfoque de análisis multidimensional, como por ejemplo, a través del procesamiento de imágenes de percepción remota, ya que como se observó a inicios del presente documento, la literatura manifiesta que el uso de los SIG’s y las técnicas de Teledetección, otorgan una visión precisa sobre la distribución de los usos del suelo, permitiendo enriquecer los resultados obtenidos del análisis normativo de los instrumentos de PLU a partir de la clasificación de técnicas de aprendizaje no supervisado como k-medias.

Prashant Banerjee (2021), en su trabajo "*K-Means Clustering with Python*", implementó el algoritmo k-medias para detectar grupos en datos sin etiquetas, un enfoque que guarda similitud con el procedimiento realizado en este estudio. Ambos casos destacan la importancia de

seleccionar el número óptimo de clústeres y evaluar la calidad de las agrupaciones para obtener resultados confiables. La validación cruzada aplicada en este trabajo respaldó la estabilidad del modelo, mostrando que los clústeres definidos reflejan patrones coherentes en los datos analizados.

Finalmente, es importante señalar las líneas abiertas de investigación del presente documento, las cuales considerando el estado del arte se centran en complementar el presente estudio clasificando los usos del suelo utilizando imágenes satelitales bajo un análisis multiescalar y temporal para contar con una perspectiva detallada sobre los patrones de ocupación y reconfiguración de los usos del suelo en áreas no urbanizables para la conservación del territorio con potencialidades ambientales y agrícolas. Por otro lado, resulta fundamental abordar el análisis de los usos del suelo y su clasificación utilizando Algoritmos de Inteligencia Artificial Explicable (conocido por sus siglas en inglés como XAI) esto con el objeto de poder comprender con precisión la complejidad del mundo urbano mejorando la confianza en la aplicación de los modelos de Aprendizaje Automático.

Agradecimientos

Este trabajo ha sido financiado por la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (Secihti) bajo la beca [821307].

Vol. 10 No. 30 Septiembre-diciembre (2025)
ISSN: 2448-5497

Referencias

Arthur, David & Vassilvitskii, Sergei. (2007). K-Means++: The Advantages of Careful Seeding. *Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms, 8*, 1027-1035. https://dx.doi.org/10.1145/1283383.1283494.

Banerjee, P. (2021). K-Means Clustering with Python. *Kaggle, 2021*. https://www.kaggle.com/code/prashant111/k-means-clustering-with-python.

Bishop, Christopher. (2006). *Pattern Recognition and Machine Learning*. https://dx.doi.org/10.1117/1.2819119.

Camelot. (s/f). Python library to extract tables from PDFs. https://camelot-py.readthedocs.io/.

Jain, Anil. (2010). Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters, 31*, 651-666. https://dx.doi.org/10.1016/j.patrec.2009.09.011.

Liu, Lingbo & Zhenghong, Peng & Wu, Hao & Hongzan, Jiao & Yu, Yang & Zhao, Jie. (2018). Fast Identification of Urban Sprawl Based on K-Means Clustering with Population Density and Local Spatial Entropy. *Sustainability, 10*, 2683. https://dx.doi.org/10.3390/su10082683.

Peña Zamalloa, Gonzalo Rodolfo. (2021). CLASIFICACIÓN ESPACIAL DEL SUELO URBANO POR EL VALOR ESPECULATIVO DEL SUELO E IMÁGENES MSI SATELITALES USANDO K-MEANS, HUANCAYO, PERÚ. *Urbano (Concepción), 24*(44), 70-83. https://dx.doi.org/10.22320/07183607.2021.24.44.06.

Ramirez Martell, C. A., Cárdenas Chujandama, M. P., & Alegría Lazo, K. M. (2024). Uso de Suelo Urbano y la Conservación del Inmueble en el Barrio San Pedro En Chazuta. Ciencia Latina

Revista Científica Multidisciplinar, 8(5), 3421–3444. https://doi.org/10.37811/cl_rcm.v8i5.13829

Müller, A. y Guido, S. (2016). *Introduction to Machine Learning with Python*. Sebastopol, CA, USA: O'Reilly Media.

Secretaría de Desarrollo Urbano e Infraestructura [SEDUI]. (2022). Plan Municipal de Desarrollo Urbano de Almoloya de Juárez 2022: Tabla de Usos del Suelo. https://legislacion.edomex.gob.mx/sites/legislacion.edomex.gob.mx/files/files/pdf/gct/2022/sep082.pdf.

Secretaría de Medio Ambiente y Recursos Naturales [SEMARNAT]. (s/f). Cambios en el uso del suelo. *SEMARNAT*. https://paot.org.mx/centro/ine-semarnat/informe02/estadisticas_2000/informe_2000/02_Vegetacion/2.2_Cambios/index.htm.

Scikit-learn. (s/f). *Sklearn.metrics.confusion_matrix*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html.

Steurer, Miriam & Bayr, Caroline. (2020). Measuring urban sprawl using land use data. *Land Use Policy, 97*, 104799. https://dx.doi.org/10.1016/j.landusepol.2020.104799.

Zhou, Xiangbing & Gu, Jianggang & Shen, Shaopeng & Ma, Hongjiang & Miao, Fang & Zhang, Hua & Gong, Huaming. (2017). An Automatic K-Means Clustering Algorithm of GPS Data Combining a Novel Niche Genetic Algorithm with Noise and Density. *ISPRS International Journal of Geo-Information, 6*, 392. https://dx.doi.org/10.3390/ijgi6120392.