UNIVERSIDAD AUTÓNOMA DE ZACATECAS

"Francisco García Salinas"

Predicción de Expansión Urbana y Detección de Cambio de Uso de Suelo Mediante Segmentación de Imágenes Satelitales y Técnicas de Machine Learning

Tesis para obtener el grado de:

Maestro en Ciencias del Procesamiento de la Información

Presenta

Ing. Alma Carmina Llamas Valenzuela

Director:

Dr. José Ismael de la Rosa Vargas

Codirectores:

Dr. Efrén Gonzales Ramírez

Dr. Gamaliel Moreno Chávez

Asesores:

Dr. José de Jesús Villa Hernández

Dr. José María Celaya Padilla

Zacatecas, Zac., 21 de enero de 2025

Zacatecas, Zac., 25 de Noviembre de 2024.

C. Alma Carmina Llamas Valenzuela
Estudiante de la MCPI
PRESENTE

At´n: Dr. Huizilopoztli Luna García
Responsable de la MCPI

Nos es grato comunicarle que después de haber sometido a revisión académica la propuesta de Tesis titulada "Predicción de Expansión Urbana y Detección de Cambio de Uso de Suelo Mediante Segmentación de Imágenes Satelitales y Técnicas de Machine Learning", presentada por el estudiante Ing. **Alma Carmina Llamas Valenzuela** y habiendo efectuado todas las correcciones indicadas por este Comité Tutorial, se **AUTORIZA** el documento de tesis para su impresión.

Sin más por el momento reciban un cordial saludo.

COMITÉ TUTORIAL
PROCESAMIENTO Y ANÁLISIS DE DATOS

Dr. José Ismael de la Rosa Vargas

Dr. Efrén González Ramírez

Dr. José de Jesús Villa Hernández

Dr. José María Celaya Padilla

Dr. Gamaliel Moreno Chávez

c.c.p. Interesado.

c.c.p. Responsable de la Maestría en Ciencias del Procesamiento de la Información.

Carta de similitud núm. 776/IyP
Zacatecas, Zacatecas 26/noviembre/2024

Dr. Huizilopoztli Luna García
Responsable de la MCPI – UAZ
Presente

Estimado Dr. Huizilopoztli,

Después de saludarlo, sirva el presente oficio para notificar que el documento

"Predicción de Expansión Urbana y Detección de Cambio de Uso de Suelo Mediante Segmentación de Imágenes Satelitales y Técnicas de Machine Learning"
de Alma Carmina Llamas Valenzuela

Fue analizado con el software Copyleaks, con la intención de detectar similitudes; el resultado en cuestión fue

1.8 % de similitud

De acuerdo a lo anterior, el porcentaje se considera **ACEPTABLE** de acuerdo a los estándares internacionales.

Atentamente
"Somos Arte, Ciencia y Desarrollo Cultural"

Dr. Carlos Francisco Bautista Capetillo
Coordinador de Investigación y Posgrado
Universidad Autónoma de Zacatecas

Zacatecas, Zacatecas 01 de diciembre de 2024

Asunto: Carta de Cesión de Derechos

Universidad Autónoma de Zacatecas
Unidad Académica de Ingeniería Eléctrica
Maestría en Ciencias del procesamiento de la información

Por la presente, yo, **Alma Carmina Llamas Valenzuela**, con numero de matrícula **42205712**, en mi calidad de autora de la tesis **"Predicción de Expansión Urbana y Detección de Cambio de Uso de Suelo Mediante Segmentación de Imágenes Satelitales y Técnicas de Machine Learning"**, declaro:

1. Que la mencionada obra es original y ha sido realizada como requisito para obtener el grado académico de Maestro en Ciencias del Procesamiento de la Información en la Universidad Autónoma de Zacatecas, bajo la dirección del Dr. José Ismael de la Rosa Vargas.
2. Que, en pleno ejercicio de mis derechos de autor, cedo de manera gratuita y voluntaria los derechos patrimoniales de la obra a la Universidad Autónoma de Zacatecas, con el fin de que esta pueda reproducirla, distribuirla, comunicarla públicamente, almacenarla en repositorios digitales y realizar cualquier otro uso permitido por la legislación aplicable para fines académicos, científicos y de difusión cultural.
3. Que esta cesión es de carácter no exclusivo, lo que significa que conservo la posibilidad de publicar o utilizar la obra en otros contextos, siempre respetando las disposiciones legales aplicables.

Sin más que agregar, firmo la presente carta como constancia de mi conformidad y aceptación de lo aquí expuesto.

Atentamente

Alma Carmina Llamas Valenzuela
Correo electrónico: llamasvlzcarmina@gmail.com

AGRADECIMIENTO ESPECIAL

Deseo expresar mi más sincero agradecimiento a la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) y al Programa de Becas para Estudios de Posgrado en México, por su valioso apoyo durante el desarrollo de mis estudios de posgrado. Gracias a su generosa beca, pude disponer de los recursos necesarios para dedicarme plenamente a mi formación académica y a la realización de este proyecto de investigación. Su respaldo fue fundamental para alcanzar mis metas académicas y profesionales, y estoy profundamente agradecido por la oportunidad que me brindaron.

Agradecimientos

Quiero expresar mi más profundo y sincero agradecimiento a todas las personas que, de una u otra manera, contribuyeron a que este logro fuera posible.

En primer lugar, quiero agradecer a mis amigos de Saquen el café. Su compañía, apoyo y palabras de ánimo fueron un refugio constante durante los momentos más desafiantes. Cada charla, cada sonrisa y cada momento compartido con ustedes me llenó de energía y motivación para continuar adelante. Gracias por estar ahí, siempre dispuestos a escuchar y acompañarme.

A mi esposo, quien ha sido mi roca, mi compañero y mi mayor apoyo. Tu paciencia, comprensión y amor incondicional han sido fundamentales en este camino. Gracias por creer en mí incluso en los días más difíciles, por animarme a no rendirme y por recordarme siempre mis capacidades. Este logro es tanto mío como tuyo, porque has sido mi mayor fuente de inspiración.

A nuestras familias, tanto la mía como la de mi esposo, quienes nos brindaron su respaldo incondicional. Su cariño, apoyo y palabras de aliento me impulsaron a dar lo mejor de mí. Gracias por comprender mis ausencias, por celebrar conmigo cada pequeño avance y por ser un pilar firme en este proceso. Su amor y confianza fueron una fuerza invaluable.

A mis asesores y directores de tesis, cuyo compromiso y dedicación fueron clave en mi formación académica. Gracias por compartir conmigo su conocimiento y experiencia, por guiarme con paciencia y por apoyarme para participar en congresos que enriquecieron mi perspectiva profesional. Su orientación no solo fue una guía para mi investigación, sino también una inspiración para seguir aprendiendo y creciendo.

Finalmente, quiero agradecer a todas las personas que, de alguna manera, dejaron su huella en este recorrido. Cada gesto, por pequeño que fuera, contribuyó a que hoy pueda cerrar este capítulo con gratitud y satisfacción.

A todos ustedes, gracias por ser parte de esta historia. Este logro es un reflejo del apoyo, amor y confianza que siempre me han brindado.

Dedicatoria

A mi esposo.

Resumen

Las ciudades son espacios físicos en los cuales se establece la población. En México, tres de cada cuatro personas viven en una ciudad. La expansión rápida y sin planeación trae consigo consecuencias indeseables para el desarrollo social y económico. El monitoreo de áreas urbanas y las modificaciones en el uso del suelo se lleva a cabo desde hace muchos años. Este monitoreo se lleva a cabo usando técnicas de percepción remota y otros índices. La detección de uso de suelo indica los patrones o tendencias de expansión urbana o el cambio del suelo en el área de estudio. El objetivo de esta investigación es proponer una metodología basada en aprendizaje automático para predecir la expansión urbana de las ciudades de Zacatecas y Guadalupe, empleando imágenes satelitales segmentadas, datos demográficos y ambientales, y así realizar una mejor planeación. Para lograr este objetivo se utilizaron mapas de uso de suelo y cobertura terrestre correspondientes al periodo 2000 a 2020, así como la inclusión de variables socioeconómicas, topográficas y atributos culturales. Se implementaron los algoritmos de Máquinas de soporte vectorial (SVM) con margen suave, Bosques Aleatorios (RF) y Boosting Categórico (catBoost). El mejor algoritmo para este problema fue SVM con margen suave. Comparado con otros estudios, este incorpora variables culturales e integra los valores SHAP, con el objetivo de conocer la influencia de dichas variables en el modelo final. Los valores SHAP revelaron que las variables clave en común fueron los mapas de uso y cobertura del suelo, las distancias a las zonas urbanas y las distancias a los centros de las ciudades. De forma contraria, la dirección de la pendiente, densidad de población religiosa y los ingresos por vivienda tuvieron una influencia menor en la predicción.

Palabras clave: expansión urbana, máquinas de soporte vectorial, bosques aleatorios, Boosting categórico, detección de cambio de uso de suelo, segmentación de imágenes, valores SHAP.

Abstract

Cities are physical spaces in which the population is established. In Mexico, three out of four people live in a city. Rapid and unplanned expansion brings with it undesirable consequences for social and economic development. Monitoring of urban areas and changes in land use and land cover has been carried out for many years. This monitoring is carried out using remote sensing techniques and other indices. Land use detection indicates patterns or trends in urban expansion or land change in the area under study. The objective of this research is to propose a methodology based on machine learning to predict urban expansion in the cities of Zacatecas and Guadalupe using segmented satellite images and demographic and environmental data, and thus carry out better planning. In order to achieve this objective, land use and land cover maps corresponding to the period 2000–2020 were used, as well as the inclusion of socioeconomic, topographic, and cultural attributes. Soft Support Vector Machines (SVM), Random Forests (RF), and Categorical Boosting (catBoost) algorithms were implemented. The best algorithm for this problem was soft SVM. Compared to other studies, this one incorporates cultural variables and integrates SHAP values to determine the influence of these variables on the final model. SHAP values revealed that the key variables in common were land use and land cover maps, distances to urban areas, and distances to city centers. Conversely, slope direction, religious population density, and household income had a minor influence on the prediction.

Keywords: Urban expansion, support vector machines, random forests, categorical boosting, land use land cover change detection, image segmentation, SHAP values.

Contenido General

Índice de figuras

Índice de Tablas

Cápitulo 1. Introducción

La expansión urbana es un fenómeno en aumento, impulsado por diferentes factores. Este fenómeno aumenta la demanda de nuevos espacios habitacionales, industriales y comerciales. En las últimas décadas, ciudades como Zacatecas y Guadalupe han experimentado una expansión urbana significativa, lo que ha planteado desafíos en la planificación territorial, el uso sostenible de los recursos naturales y la gestión de infraestructuras. Este proceso de expansión descontrolada no solo altera el paisaje físico, sino que también afecta la calidad de vida de los habitantes y la estabilidad ambiental de la región.

El estudio de la expansión urbana ha sido abordado desde diferentes enfoques. Algunos trabajos han utilizado modelos basados en autómatas celulares para simular los patrones de expansión urbana. Planteando diferentes escenarios, con restricciones o sin restricciones. Otros han optado por algoritmos de aprendizaje automático para realizar predicciones. Sin embargo, la mayoría de estos estudios se han centrado en regiones fuera de México, lo que subraya la necesidad de investigaciones locales que consideren las particularidades geográficas, sociales y ecológicas del país.

En este contexto, el presente estudio tiene como objetivo proponer una metodología basada en aprendizaje automático para predecir la expansión urbana de las ciudades de Zacatecas y Guadalupe a partir de imágenes satelitales segmentadas, datos demográficos y ambientales de los años 2000, 2010 y 2020. Asimismo, de las imágenes segmentadas, se crearon mapas de uso de suelo que permitieron detectar los cambios ocurridos en este periodo. Para pronosticar la expansión urbana hacia el año 2030, se implementaron algoritmos de aprendizaje automático, tales como Máquinas de Soporte Vectorial, Bosques Aleatorios y Boosting Categórico, con el fin de ofrecer predicciones precisas sobre las áreas con mayor probabilidad de expansión.

Los resultados de esta investigación proporcionarán una base sólida para la toma de decisiones tanto en el ámbito público como privado, ayudando a gestionar de manera más eficiente los recursos urbanos y naturales. Además, la visualización de los mapas de expansión proyectados permitirá identificar las zonas con mayor riesgo de desarrollo descontrolado, lo que es esencial para la implementación de políticas urbanas sostenibles y la protección del entorno.

1.1 Antecedentes

Las ciudades son espacios físicos en los cuales se establece la población. El aumento de la población ocasiona la necesidad de crecimiento urbano. El crecimiento urbano convierte a las ciudades en centros generadores de empleo, de innovación y aumenta la calidad de vida de sus habitantes, servicios básicos y viviendas [1].

Al existir crecimiento, la expansión urbana es un fenómeno inherente. La expansión urbana debe asegurar el correcto cambio de uso de suelo, este cambio se conoce como la adecuación o cambio de la cobertura terrestre o natural a algo útil a la ciudad. Este cambio de empleo de suelo trae una pérdida de áreas agrícolas y vegetación autóctona. Esto trae consigo un incremento en superficies impermeables y una reducción de áreas verdes [2].

La expansión urbana rápida y sin planificación genera consecuencias negativas para el desarrollo social y económico. Algunas de las consecuencias son: incremento en la temperatura [2], el daño ecológico [3], insuficientes viviendas para la demanda, sistemas de transporte deficientes, aumento en la segregación social [4] y el incremento del tráfico [5]. Estas consecuencias se

pueden aminorar, realizando una planeación a largo plazo, apoyada de diferentes instrumentos [6].

Alguno de los instrumentos existentes es el monitoreo de áreas urbanas y las modificaciones en el uso del suelo. Estos monitoreos se llevan a cabo desde el lanzamiento del primer satélite de observación de la tierra [7]. Este monitoreo se ha efectuado utilizando técnicas de percepción remota, apoyándose en el uso de Sistemas de Información Geográfica (GIS, por sus siglas en inglés) [8], la cual emplea imágenes satelitales para su análisis. Algunos análisis se hacen con índices como el Índice de Diferencia Normalizada y Distancia de Áreas Construidas [9], segmentación de áreas construidas con Optimización por Enjambre de Partículas [10], empleo del Índice de Intensidad de Expansión Urbana [11], [12] o el Índice de Diferenciación de la Intensidad de Expansión Urbana [9][10][11], [12][13] Estos estudios ayudan al monitoreo de los diferentes usos de suelos de las áreas analizadas, las cuales pueden indicar los patrones o tendencias de expansión urbana o del cambio del uso de suelo.

Existen diferentes estudios donde se ha realizado monitoreo de áreas urbanas, en otros estudios se han utilizado diferentes modelos matemáticos para la simulación de expansión urbana. Estos modelos se apoyan de los monitoreos mencionados anteriormente. Algunos de estos incluyen los basados en agentes (ABM) [14] y en autómata celular (CA, por sus siglas en inglés) [15], [16], [17], [18]. El modelo comúnmente empleado para simular la expansión urbana es él basado en CA [19]. CA es una simplificación de la realidad, construida con elementos clave. Estos se han utilizado en diferentes estudios. Incluido en algunas ciudades mexicanas, como en las ciudades de: Morelia [20], Área Metropolitana de Toluca [5], [21], [22], Ciudad de México y Monterrey [23]. Es importante aclarar que estos esfuerzos de prever la expansión urbana no sustituyen el conocimiento de los desarrolladores inmobiliarios o académicos en la planeación de ciudades. Los instrumentos deben considerarse como un apoyo a la toma de decisiones.

Sin embargo, estos modelos asumen que las áreas urbanas son espacialmente homogéneas. Por esta razón, el modelado de las decisiones individuales e interacciones socioeconómicas es difícil de representar [24].

Con la intención de solucionar esta falta de representación, se comenzó a incorporar el aprendizaje automático, como: XGboost-SHAP [6], [25], máquinas de soporte (SVM) [26], [27], árboles de decisión [28], bosques aleatorios [29], entre otros. Sin embargo, estos estudios se realizaron fuera de México. Lo anterior y el hecho de que para el 2020 casi cien millones de mexicanos vivían en asentamiento urbanos no planificados [5] impulsaron la creación de este estudio.

1.1.1 Área de estudio

Esta investigación se centra en las ciudades de Zacatecas y Guadalupe (Figura 1.1). estas forman parte del estado de Zacatecas, en México, ubicado en la zona centro-norte del país, lo que lo convierte en un estado altamente conectado por transporte terrestre. Esta situación se debe a su proximidad con ciudades como Aguascalientes, San Luis Potosí y Durango [30]. Zacatecas representa el 3.8% de la superficie total del país [31], lo cual ha contribuido al desarrollo del área metropolitana Zacatecas-Guadalupe, una de las 48 áreas metropolitanas reconocidas en México. Esta región se destaca por su rápida expansión urbana, diversidad cultural y la alta demanda de recursos y servicios naturales [32].

El estudio abarca de 2000 a 2020, con una duración de 20 años en los que la región ha experimentado cambios significativos. Según [31], el número de viviendas particulares pasó de 299,249 en el año 2000 a 442,263 en 2020, mientras que el promedio de ocupantes por vivienda

disminuyó de 4.5 a 3.7 en el mismo período. Además, [33] reporta una tasa de crecimiento promedio anual del 2.4% entre 2000 y 2015, mientras que [32] menciona un crecimiento del 2.03% de 2010 a 2020 en el área metropolitana de las ciudades Zacatecas-Guadalupe.

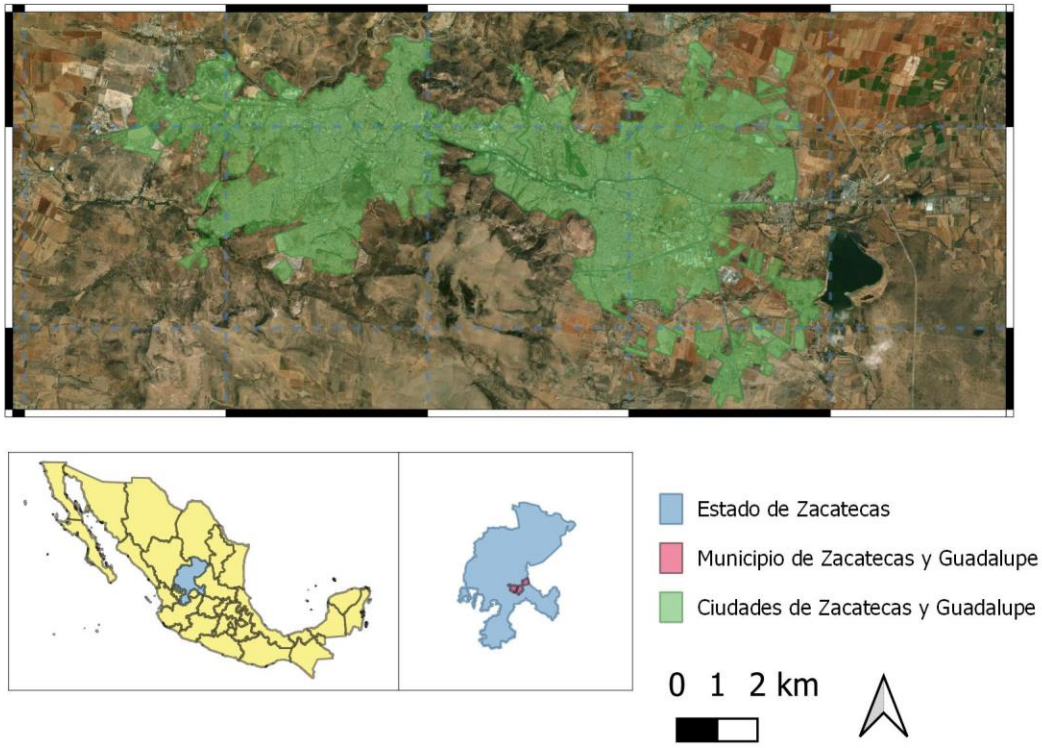

Figura 1.1. Área de estudio, ciudades de Zacatecas y Guadalupe.

1.2 Planteamiento del problema de investigación

El crecimiento urbano es un fenómeno complejo que depende de múltiples factores y variables, que pueden variar según el contexto geográfico, histórico, cultural y político de cada ciudad [19]. Las ciudades se enfrentan a problemáticas y oportunidades dependiendo de su escala, temporalidad y región, internamente enfrentan problemáticas como cobertura de servicios públicos básicos, seguridad, empleo, calidad de vida, equidad, entre otros [34].

Las problemáticas que enfrenta una ciudad pueden ser severas si no se tiene un control en el crecimiento de la ciudad y adicionalmente aumentar el daño ecológico derivado de la expansión [4]. Además, se puede aumentar el tráfico y falta de acceso a servicios básicos (agua, saneamiento, entre otros). Adicionalmente, el incremento demográfico y el desarrollo económico trae consigo la expansión urbana [35].

La expansión urbana es la transformación del uso de suelo a áreas construidas [36]. Esta transformación irreversible llevada a cabo por los seres humanos [37] ocasiona una pérdida de áreas agrícolas y vegetación autóctona. De acuerdo con [38] el futuro de la humanidad es urbano, pronosticando que para el año 2050 el 68 % de la población mundial vivirá en una ciudad urbana.

En 1950, México era el segundo país más poblado de Latinoamérica y el caribe, con 28 millones de habitantes, con el 43 % viviendo en asentamientos urbanos [1]. En 1990 las personas que vivían

en asentamientos urbanos sé incremento a un 71 % [5]. Mientras que en 2012 tres de cada cuatro personas, que corresponde al 72 %, viven en una ciudad [34]. Por otro lado, en 2018 en México la población creció a 131 millones y el 80 % vivía en un asentamiento urbano [1], muchos de estos asentamientos son ciudades grandes no planificadas [5].

El aumento de la población en las últimas décadas ha motivado a un progreso positivo que ayuda a mantener una población en constante aumento. A pesar de mejorar el acceso a servicios sociales para la población trabajadora, en la estrategia enfocada en el desarrollo estable no se consideró adecuadamente las disparidades regionales. La carencia de planificación territorial provocó tres desafíos primordiales: la disparidad en el desarrollo regional, el crecimiento urbano desordenado y la explotación rápida de los recursos naturales y la degradación de los ecosistemas [39].

La desigualdad territorial surge como una consecuencia de la centralización económica, que ha conferido una prominencia destacada a regiones específicas del país, como Ciudad de México, Monterrey y Guadalajara. No obstante, esta focalización en dichas áreas ha tenido un efecto adverso en el sursureste del país, donde se ha concentrado una proporción significativa de la población en situación de pobreza [39]. Este fenómeno de desigualdad territorial también se manifiesta dentro de las ciudades mismas.

Por otro lado, el crecimiento urbano se desencadenó por la migración rural-urbana [40], este incentivo la creación de conurbación [39]. El ritmo de crecimiento fomentó una expansión urbana desordenada y sin planeación. La expansión urbana desorganizada se vio agravada por políticas públicas improvisadas, que tienen efectos negativos en el desarrollo social y económico [5].

Por último, la expansión urbana y rural, como parte del proceso de industrialización, ocasionaron consecuencias directas sobre el ecosistema [39]. Esto ha traído como tema central a la seguridad hídrica. Creando presión sobre este recurso, el cual está en riesgo por la falta de previsión y planeación de la expansión urbana [21].

Estas problemáticas relacionadas con la expansión urbana sin control traen consigo consecuencias negativas. Algunas de estas son: incremento en la temperatura [2], daño ecológico [4], aumento de superficies impermeables y paisajes fragmentados [41], aumento en las emisiones de carbono y disminución de la captura de carbono [42], insuficientes viviendas para la demanda, sistemas de transporte deficientes, aumento en la segregación social [5] y el incremento del tráfico [6].

Por esta razón, se requieren métodos y modelos que puedan captar la diversidad y la incertidumbre de la expansión urbana, y que puedan adaptarse a las características y necesidades específicas de cada caso de estudio. Para abordar los impactos desfavorables y mejorar los beneficios de la expansión urbana a largo plazo, es importante contar con una planificación a largo plazo respaldada por diversos instrumentos [8]. Se han ideado numerosos enfoques para modelar la expansión, algunos de los cuales se basan en datos obtenidos mediante sensores remotos, utilizando imágenes satelitales para determinar el uso y la cobertura del suelo. Además, estos modelos cuentan con un amplio respaldo de Sistemas de Información Geográfica (SIG) [15].

Diferentes modelos matemáticos utilizados para la expansión urbana incluyen Autómata Celular [5], [16], [17], [18]. Sin embargo, suponen que las áreas urbanas son espacialmente homogéneas, lo que dificulta representar con precisión las decisiones individuales y las interacciones socioeconómicas.

Para abordar diferentes limitaciones, los investigadores han empleado no solo métodos estadísticos, como diferentes tipos de regresiones [5], [43], sino también diversas técnicas de aprendizaje automático, como XGBoost-SHAP [6], [25], máquinas de vectores de soporte (SVM)

[26], [27], árboles de decisión [28], bosques aleatorios [29] y otros. Sin embargo, estos estudios se realizaron fuera de México.

En México se han llevado a cabo diferentes estudios utilizando los Autómata Celulares en las ciudades de: Morelia [20], Área Metropolitana de Toluca [5], [21], [22], Ciudad de México y Monterrey [23]. Estos esfuerzos de simular la expansión urbana no sustituyen el conocimiento de expertos. Esto instrumentos deben de considerarse como un apoyo a la toma de decisiones.

En esta investigación se consideró a cada ciudad como un conjunto de patrones de expansión únicos. En el presente se decidió tomar como zona de estudio a las ciudades de Zacatecas y Guadalupe, esto es debido a sus características geográficas, hidrográficas y sociales.

Estos cambios demográficos y de vivienda tienen implicaciones directas en el grado de rezago social en la región. El aumento en el número de viviendas y la disminución en el promedio de ocupantes por vivienda pueden ser indicativos de una mejora en las condiciones de vida, reflejando una mayor accesibilidad a la vivienda. Sin embargo, el crecimiento poblacional constante también puede ejercer presión sobre los recursos y servicios disponibles, lo que podría agravar el rezago social si no se gestionan adecuadamente.

Como se mencionó anteriormente, la expansión urbana es un fenómeno multifactorial que tiene consecuencias en diferentes ámbitos. Para aminorar las consecuencias y lo que ocasiona la expansión urbana descontrolada, es importante llevar a cabo estudios que ayuden a entender cómo ocurre dentro del contexto de las áreas de estudio.

Consideramos que integrar herramientas de detección de cambio de uso de suelo y aprendizaje automático, serán capaces de comprender las tendencias de expansión. Este estudio no sólo pretende comprender el fenómeno en el presente, sino también ofrecer una herramienta que nos muestre diferentes escenarios de expansión. Con la intención de que se usen para la toma de decisiones.

Adicionalmente, consideramos que lo desarrollado es adaptable a otras regiones geográficas con características similares. La validación de estos métodos será fundamental para garantizar la aplicabilidad y efectividad en contextos diversos, permitiendo un enfoque más integral y sostenible en la planificación urbana futura.

1.3 Justificación del problema de investigación

La identificación de los patrones de expansión y la predicción que promueva una planificación a largo plazo ayudarían a reducir las consecuencias negativas de la expansión urbana descontrolada [5]. Anticipar y gestionar la expansión urbana de manera adecuada permite mejorar la calidad de vida de los residentes y tener una expansión sostenible [44], tomando en cuenta las características específicas de los asentamientos y sus patrones de expansión histórica.

Las ciudades de Zacatecas y Guadalupe enfrentan desafíos únicos en términos de expansión urbana. Esto se debe a que son ciudades históricas y sus alrededores deben manejarse cuidadosamente para preservar su identidad cultural, mientras se adapta a las crecientes necesidades de infraestructura y vivienda [45]. Además, la expansión urbana puede afectar la gestión de recursos naturales críticos como el agua [21] y el suelo, así como la movilidad urbana y la calidad del aire. La planificación urbana efectiva no solo busca facilitar un crecimiento ordenado y equitativo, sino también mitigar los impactos negativos asociados, como el aumento en la demanda de viviendas, la presión sobre los servicios básicos y la fragmentación del paisaje urbano y rural [4].

El trabajo de investigación se basa en la hipótesis de que la expansión urbana depende de factores y tendencias que se pueden medir y modelar matemáticamente. Con imágenes satelitales y técnicas de segmentación, se pueden obtener mapas de suelo que conozcan la expansión urbana. La ventaja de esta propuesta es que permite obtener resultados sin necesidad de realizar un trabajo de campo extenso, utilizando técnicas indirectas de medición como la percepción remota y el aprendizaje automático. No obstante, es importante reconocer que esta metodología no considera aspectos subjetivos, como las preferencias o percepciones de los habitantes, que también pueden influir en la expansión urbana.

En estudios previos, [6], [25] modelaron la expansión urbana en Corea del Sur, y lograron resultados precisos en la predicción de patrones futuros de expansión urbana. En este trabajo se incluyen variables ecológicas. De manera similar, [26], [28] utilizaron algoritmos de aprendizaje automático, como máquinas de soporte vectorial y árboles de decisión en ciudades de Estados Unidos. No obstante, estos trabajos no abordan las particularidades del contexto mexicano, algo que este estudio tiene como objetivo incorporar.

Aunque estos estudios ofrecen avances significativos, ninguno de ellos se ha llevado a cabo en las ciudades de Zacatecas y Guadalupe. La presente investigación busca llenar este vacío al integrar técnicas de segmentación y algoritmos de aprendizaje automático para predecir la expansión urbana en Zacatecas y Guadalupe. Los resultados proporcionarán información crucial para la planificación urbana y la gestión sostenible de los recursos hídricos, contribuyendo tanto a la academia como a las políticas públicas.

1.4 Preguntas de investigación

¿Qué tan precisas son las técnicas de segmentación al evaluar las áreas urbanizadas en las imágenes satelitales de los años 2000, 2010 y 2020 en las ciudades de Zacatecas y Guadalupe?

¿Qué cambios en el uso de suelo se pueden observar en las ciudades de Zacatecas y Guadalupe al comparar los mapas de 2000, 2010 y 2020?

¿Qué áreas han mostrado las tasas más altas de cambio de uso de suelo en la zona metropolitana?

¿Cuál de los algoritmos de aprendizaje automático, como Máquinas de Soporte Vectorial, Bosques Aleatorios y Boosting Categórico, ofrece mejores resultados en la predicción de la expansión urbana en Zacatecas y Guadalupe?

¿Qué variables son más influyentes en la predicción de la expansión urbana para el año 2030?

1.5 Objetivo general

Proponer una metodología basada en aprendizaje automático para predecir la expansión urbana de las ciudades de Zacatecas y Guadalupe, empleando imágenes satelitales segmentadas, datos demográficos y ambientales.

1.6 Objetivos específicos

- Segmentar y evaluar las áreas urbanizadas en cada imagen de los años 2000, 2010 y 2020, utilizando técnicas supervisadas
- Elaborar mapas de uso de suelo de las ciudades de Zacatecas y Guadalupe para los años 2000, 2010 y 2020, basados en los resultados obtenidos de la segmentación de

imágenes.

- Detectar y analizar los cambios en el uso de suelo en la zona metropolitana Zacatecas-Guadalupe entre los años 2000, 2010 y 2020, utilizando técnicas de análisis geo espacial.
- Implementar algoritmos de aprendizaje automático, como Maquinas de Soporte Vectorial, Bosques Aleatorios y Boosting Categórico, para predecir la expansión urbana de las ciudades de Zacatecas y Guadalupe, con proyecciones para el año 2030.

1.7 Hipótesis

La expansión urbana de las ciudades de Zacatecas y Guadalupe se puede predecir con un modelo basado en técnicas de aprendizaje automático, utilizando datos demográficos, ambientales y mapas de uso de suelo obtenidos mediante segmentación de imágenes satelitales. El modelo logrará un rendimiento con un F1-Score superior al 80%.

1.8 Estructura de la tesis

El documento consta de cinco capítulos principales en los que se describe el trabajo realizado. El Capítulo 1 contiene una descripción de los antecedentes, los objetivos, hipótesis y preguntas de investigación planteadas en esta investigación, así como la presentación de la problemática y el propósito general del trabajo.

En el Capítulo 2, se presentan las teorías base y conceptos que sustentan a la investigación, adicionalmente se presentan los trabajos relacionados.

En el Capítulo 3 se describe la metodología propuesta y seguida en esta investigación, así como los materiales e instrumentos utilizados para el desarrollo del trabajo.

Capítulo 4, en este apartado se presentan los experimentos elaborados en el trabajo, así como los algoritmos utilizados y sus respectivos parámetros. Igualmente se presentan los resultados de los experimentos, validación del modelo propuesto y algunas de las limitaciones de la investigación.

En el capítulo 5, se recogen las conclusiones de esta investigación, se habla sobre el trabajo futuro propuesto y la respuesta a las preguntas de investigación.

Por último, este documento contiene un apartado de Referencias y Anexos, en la primera sección mencionada contiene las fuentes consultadas en el desarrollo de este trabajo de investigación, en el segundó se encuentran los anexos, y los diferentes instrumentos utilizados.

Cápitulo 2. Marco Teórico

En este capítulo se describirán las teorías básicas que dan sustento a este trabajo de investigación, así mismo se describen los conceptos empleados para la elaboración de lo realizado.

2.1 Imágenes satelitales multiespectrales

De acuerdo con [46] una imagen satelital multiespectral es una imagen digital de una escena resultante de lo adquirido por sensores ópticos abordo de un satélite. Estas imágenes albergan la radiancia de los elementos encontrados en la superficie terrestre en un número determinado de bandas espectrales.

La existencia de un número determinado de bandas convierte a nuestra imagen en un cubo de tres dimensiones (x, y, λ) [47]. En este cubo se tienen dos dimensiones espaciales (x para las filas e y para las columnas) y una tercera dimensión que corresponde a la longitud de onda del espectro electromagnético (Figura 2.1). Cada banda espectral corresponde a un rango determinado del espectro electromagnético de una misma escena.

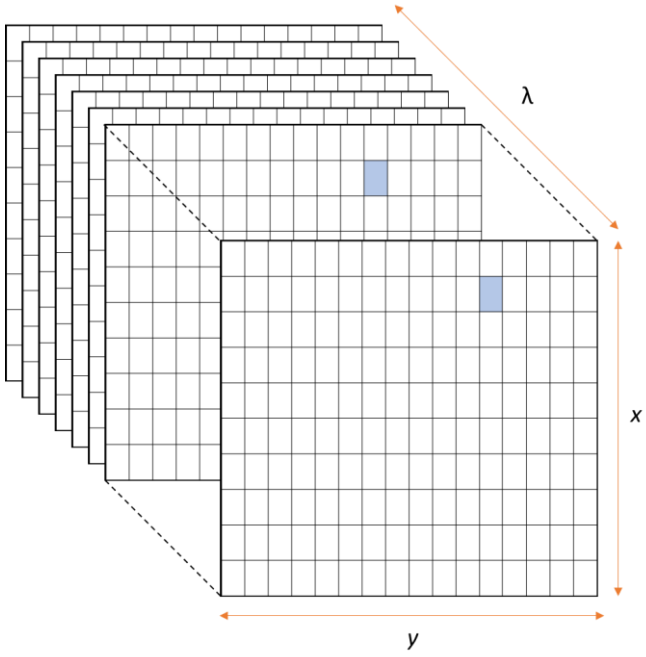

Figura 2.1. Dimensiones de las imágenes satelitales multiespectrales (x, y, λ).

El número de bandas y los rangos del espectro electromagnético dependen del sensor y del objetivo de la constelación de satélites. Lo que lo convierte en una imagen 3D discreta. Actualmente existen constelaciones de satélites que proveen este tipo de imágenes. Un ejemplo de estos son las misiones Landsat, Sentinel, Modis, entre otros. Las bandas espectrales más utilizadas son las correspondientes al rango visible y el infrarrojo del espectro electromagnético.

2.1.1 Landsat

Landsat es una familia de satélites que empezaron a operar en 1972. Estos satélites adquieren imágenes de la superficie terrestre y se utilizan para diferentes investigaciones. Algunas de las aplicaciones más importantes de estas investigaciones están en la agricultura, geología, planeación regional, educación, investigación forestal, entre otras.

Landsat es un proyecto que desarrolló la U.S. Geological Survey (USGS) y la National Aeronautics and Space Administration (NASA). El primer satélite se lanzó el 23 de julio de 1972 y recibió el nombre de Landsat 1. Los lanzamientos de Landsat 2, Landsat 3 y Landsat 4 fueron en 1975, 1978 y 1982, respectivamente. Landsat 5 se lanzó en 1984; este satélite estuvo en funcionamiento por 28 años y 10 meses, siendo el satélite de observación con más tiempo de operación. En 1993 el lanzamiento de Landsat 6 falló. El lanzamiento de Landsat 7 en 1999 fue un éxito y actualmente sigue dando datos globales. Landsat 8 se lanzó en febrero de 2013. Por último, el lanzamiento de la misión Landsat 9 se realizó en el 2021 y actualmente proporciona datos junto con la anterior.

Las imágenes multiespectrales generadas por la constelación Landsat tienen diferentes bandas dependiendo de la misión. Landsat 8 y 9 están equipados con diferentes sensores: el Operational Land Imager (OLI) y el Thermal Infrared Sensor (TIRS). Mientras que el Landsat 7 está equipado con el Enhanced Thematic Mapper Plus (ETM+), sin embargo, por un fallo, un porcentaje de las imágenes generadas por este sensor se perdieron. El Landsat 4 al 5 están provistos con el sensor Multispectral Scanner (MSS) y él Thematic Mapper (TM). La diferencia de sensores a bordo de los satélites ocasiona que las bandas estén compuestas por una longitud de banda diferente o estas imágenes contengan más bandas que otras misiones. En la Tabla 2.1 se muestran las longitudes de onda del espectro electromagnético de las imágenes multiespectrales [48].

Tabla 2.1. Longitud de onda de las bandas espectrales de las imágenes satelitales multiespectrales Landsat en micrómetros.

Banda	Longitud de onda		
	Landsat 8 - 9	Landsat 7	Landsat 4 -5
Azul	0.45 – 0.51 (B2)	0.45 – 0.52 (B1)	0.45 – 0.52 (B1)
Verde	0.53 – 0.59 (B3)	0.52 – 0.60 (B2)	0.52 – 0.60 (B2)
Rojo	0.64 – 0.67 (B4)	0.63 – 0.69 (B3)	0.63 – 0.69 (B3)
Infrarrojo cercano	0.85 – 0.88 (B5)	0.77 – 0.90 (B4)	0.76 – 0.90 (B4)
Infrarrojo de onda corta 1	1.57 – 1.65 (B6)	1.55 – 1.75 (B5)	1.55 – 1.75 (B5)
Infrarrojo de onda corta 2	2.11 – 2.29 (B7)	2.09 – 2.35 (B6)	2.08 – 2.35 (B6)

Además de las bandas espectrales existen otras características a considerar, resolución espacial y resolución temporal. La resolución espacial se refiere a la superficie terrestre que cubre un píxel de nuestra imagen. Las misiones Landsat tienen una resolución de 30 m x 30 m por píxel en las bandas visibles, infrarrojo cercano e infrarrojo de onda corta.

La resolución temporal está ligada al tiempo en el que los satélites se encuentran con la escena nuevamente y los sensores son capaces de capturar la reflectancia, esta resolución depende de la órbita de los satélites, en la constelación Landsat la resolución temporal es de cada 16 días.

A pesar de las limitaciones ligadas a este tipo de imágenes tienen una gran variedad de aplicaciones [49], algunas de estas son el monitoreo de agricultura, deforestación, geología, creación de mapas de uso de suelo, manejo de recursos (agua) y estudios en la costa. Estas

aplicaciones son posibles por la variedad de usos que se le puede asignar a cada una de las bandas [48].

2.1.2 Geo-mediana Nacional

La geo-mediana Nacional es un producto generado y publicado por el INEGI, generados a partir del Cubo de Datos Geoespaciales de México (CDGM). El CDGM usa imágenes satelitales provenientes de los satélites Landsat. Hoy se tienen datos históricos desde 1984. El CDGM está compuesto de imágenes del Landsat 4 hasta el Landsat 9, en estos se incorporan imágenes de las bandas espectrales del rango visible (azul, verde, rojo), infrarrojo cercano, infrarrojo de onda corta 1 e infrarrojo de onda corta 2, todas conservando la resolución espacial original de las imágenes de 30 metros [50].

Existen diferentes formas de crear una imagen a partir de series de tiempo. Sin embargo, no todas las metodologías preservan las relaciones espectrales, lo cual es lo deseable en la percepción remota. Por esta razón, se emplea el método basado en la creación de imágenes a composición de nivel píxel. Las nuevas imágenes se generan a partir de imágenes que pueden encontrarse en diferentes escenas de Landsat. Uno de los métodos de composición de píxel es la geo-mediana. La imagen resultante es un mosaico representativo de un tiempo específico.

La geo-mediana es un enfoque de composición a nivel pixel para obtener una imagen representativa de un conjunto de imágenes Landsat. Esta nueva imagen mantiene la relación espectral y minimiza los valores atípicos y el ruido. Lo que la convierte en una herramienta útil para monitorear fenómenos ambientales, socioeconómicos y demográficos que se reflejan en el territorio de manera persistente, como el crecimiento urbano, los cambios de uso de suelo, la extensión de cuerpos de agua, la degradación del suelo, entre otros. En la Figura 2.2 se muestra el proceso seguido por INEGI propuesto por [51].

El algoritmo propuesto por [51] emplea la mediana geométrica, conocida como mediana espacial. Este es un método iterativo, en el cual cada iteración mejora respecto a la anterior. En este método se aprovechan las bondades de la mediana geométrica, ya que esta estadística permite preservar las relaciones entre las bandas espectrales.

La mediana geométrica en el contexto de la percepción remota se define de la siguiente forma. Dado un conjunto finito $\mathbb{X}$ que corresponde a una serie de tiempo asociada a un pixel de *p*-bandas, como vector $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n} \in \mathbb{R}^p$, por lo tanto, la mediana geométrica es:

$$m := \underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \|x - x_i\|. \tag{2.1}$$

En la ecuación (2.1) se concluye que la geo-mediana es la minimización de la suma de las distancias Euclidianas de un conjunto de observaciones. Convirtiéndolo en un punto mínimo en el espacio multidimensional de p-bandas. En la referencia [6] propone un algoritmo iterativo, que corresponde a un algoritmo iterativo de mínimos cuadrados. El algoritmo propuesto se encuentra en la Tabla 2.2.

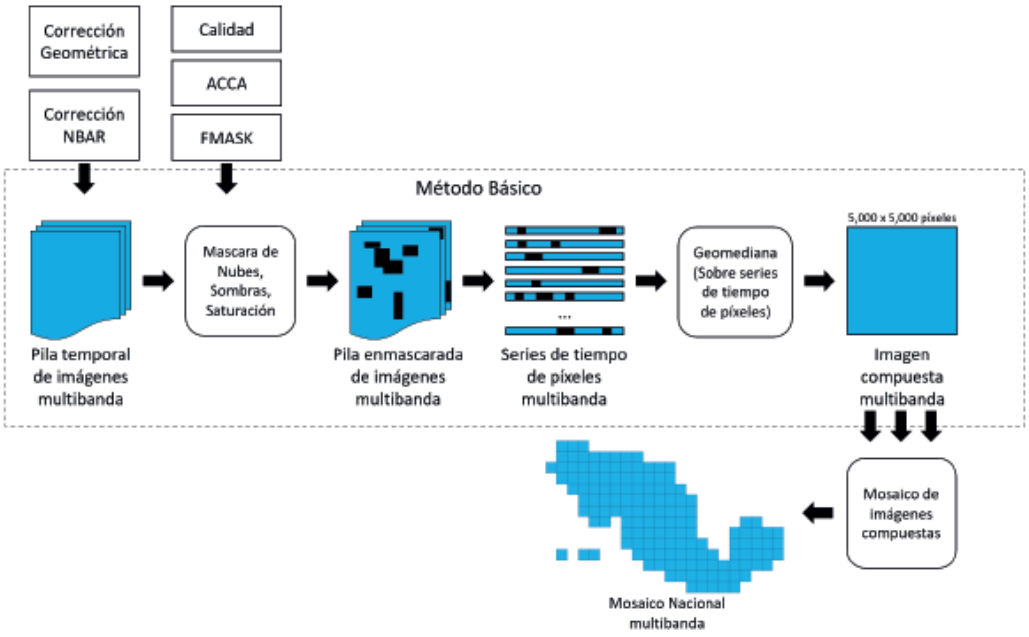

Figura 2.2. Diagrama del proceso de la generación de la geo-mediana tomado de [5] inspirado por el diagrama de [51].

El algoritmo tiene dos criterios para su conclusión, el máximo de iteraciones y el valor épsilon que determina la precisión de la geo-mediana. Al finalizar se tiene una imagen compuesta de 5000 x 5000 píxeles en una imagen con formato TIFF [50].

Tabla 2.2. Algoritmo propuesto por [51] para el cálculo de la geo-mediana.

Algoritmo Mediana Geométrica

$\mathbb{X} \leftarrow [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}]$

Procedimiento GEO-MEDIANA

$\mathbb{X}$, $\boldsymbol{\epsilon = 10^{-7}}$, iteraciones = 1000

$\boldsymbol{y_0 \leftarrow \frac{1}{n}\sum_{i=1}^{n} x_i}$, $\boldsymbol{k \leftarrow 0}$

while $\boldsymbol{k < iteraciones}$ **do**

$\boldsymbol{y_{k+1} \leftarrow \left(\sum_{i=1}^{n} \frac{x_i}{\|x_i - y_k\|}\right) / \left(\sum_{i=1}^{n} \frac{1}{\|x_i - y_k\|}\right)}$

if $\boldsymbol{\|y_{k+1} - y_k\| / \sqrt{p} < \epsilon}$ **then**

break

$\boldsymbol{k \leftarrow k + 1}$

return $\boldsymbol{y_k}$

Las ventajas de usar este tipo de productos es que tiene menos ruido, debido a que se emplean las máscaras de calidad para detectar las nubes y los píxeles de baja calidad. Estos píxeles son repuestos con los pixeles de las otras imágenes creando un resumen de los píxeles disponibles. El algoritmo emplea las seis bandas de las imágenes a la vez, por lo que conserva la razón entre los valores. Por esto se pueden aplicar diferentes operaciones para calcular índices con las bandas.

2.2 Segmentación

El procesamiento digital de imágenes consiste en un conjunto de técnicas y procesos, con el objetivo de resaltar o descubrir información contenida en nuestra imagen. La segmentación forma parte de estas técnicas. La tarea consiste en dividir los píxeles de la imagen en función de características similares [52].

En la referencia [53] se define a la segmentación de manera formal usando la teoría de conjuntos. Sea R un conjunto no vacío que representa una imagen. Se puede considerar que la segmentación de R consiste en crear múltiples subconjuntos (subregiones) no vacíos, $\{R_i,\ i = 1, 2,\ \dots, n\}$ los cuales cumplen las siguientes cinco condiciones:

1. $\bigcup_{i=1}^{n} R_i = R$
2. Las subregiones no se superponen entre sí, es decir, $R_i \cap R_j = \emptyset$ para todo i y j, $i \neq j$.
3. Cada pixel contenido en una subregión debe de tener algo en común o ser parecidos (color, textura similar, ser parte del mismo objeto).
4. No permite la combinación de dos regiones para crear una región homogénea, si combinamos dos o más subregiones, los píxeles que la forman ya no serán parecidos o no tendrán nada de similitud.
5. Para cada $i = 1, 2, \dots, n$, R_i es una región conectada, es decir, existe un camino entre cualquier par de píxeles dentro de la región.

La segmentación puede realizarse basada en los píxeles, contornos, regiones y el aprendizaje profundo. Cada técnica tiene sus ventajas y desventajas, la elección depende de la imagen, o del problema de segmentación [52].

Las técnicas basadas en pixel son aquellas que dividen la imagen usando los valores de los píxeles [54]. En estas incluyen las técnicas de umbral y los métodos de agrupamiento. Las técnicas de umbral se basan en identificar un valor de umbral, donde los píxeles se dividen en grupos dependientes de los valores de intensidad de la imagen. Esta técnica es fácil de implementar, sin embargo, estas tienen mejor rendimiento cuando la imagen es bimodal. Los métodos de agrupamiento pretenden colocar los pixeles en grupos usando el color, la intensidad o la localización, con una medida de distancia [7]. Un ejemplo de estos son k-means, c-means, fuzzy c-means, agrupamiento jerárquico, entre otros.

La segmentación basada en contornos realiza su tarea buscando los límites de las diferentes regiones en la imagen. Estos métodos usan el gradiente de la imagen, algunos de estos son: Sobel, operador Canny, Prewitt, entre otros. Por último, los métodos basados en región se inicializan con una semilla. Los píxeles que son similares a la semilla se unen a ella, posteriormente se actualiza la semilla usando estos pixeles hasta que se llegue a la convergencia. Principalmente, este tipo de técnicas incluyen crecimiento de regiones, división de píxeles y super-píxel [54].

2.2.1 Técnicas de segmentación usando agrupamiento

El agrupamiento es un procedimiento que se centra en dividir las observaciones en diferentes subconjuntos. Esto se realiza mediante una medida de similitud. Lo que ocasiona que los datos que tienen alguna similitud se agrupen.

Suponga que $\boldsymbol{X} = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^{n \times D}$ es el conjunto de datos a procesar. Este conjunto se

dividirá en c subconjuntos, los cuales serán, $X_1, X_2, \dots, X_c$, estos subconjuntos deben de cumplir las siguientes condiciones:

$$\boldsymbol{X_1} \cup \boldsymbol{X_2} \cup \dots \cup \boldsymbol{X_c} = \boldsymbol{X}, \tag{2.2}$$

$$X_i \cap X_j = \emptyset, 1 \leq i \neq j \leq c, \tag{2.3}$$

$$X_i \neq \emptyset, X_i \neq X \text{ , } 1 \leq i \leq c. \tag{2.4}$$

La condición (2.2) nos indica que la unión de todos los subconjuntos debe formar el conjunto de datos o imagen en su totalidad, mientras que (2.3) nos indica que los conjuntos son excluyentes, es decir, no existe traslape entre los subconjuntos, por último (2.4) indica que no pueden existir subconjuntos vacíos, al menos debe existir una observación o píxel en el subconjunto creado [55].

Generalmente, estos algoritmos incluyen tres pasos principales: extracción de características, calcular la medida de similitud y agrupamiento. Al igual que otros algoritmos usados para segmentación, se tienen que considerar algunas cosas antes de utilizarlos, una de ellas es que podría ser una técnica que no necesita etiquetas. Convirtiéndolo en un algoritmo autónomo, sin embargo, se necesita proporcionar algunos parámetros de inicialización [52]. Existen muchos algoritmos de agrupamiento enfocados en la segmentación de imágenes, estos pueden ser clasificados en dos principales categorías: jerárquico y particional.

El agrupamiento Jerárquico, obtiene los grupos usando diferentes niveles de similitud creando un esquema en tipo de árbol. Estos algoritmos pueden ser clasificados por su enfoque: unión (particional) o por división (divisivo). Algunos ejemplos de estos algoritmos son agrupación divisiva (DVCLUS-T), agrupación usando representantes (CURE), chameleon, entre otros. Sin embargo, estos algoritmos son complejos y tienen una alta complejidad, por estas razones los algoritmos basados en particiones son más utilizados en conjunto de datos grandes [55].

En el enfoque particional los datos se agrupan de acuerdo a una función objetivo, que determina la similitud de los datos a cada uno de los grupos. Para lograrlo, la similitud de cada elemento a cada grupo tiene que calcularse. Al igual que los métodos anteriores la cantidad de grupos tiene que ser definida a priori. Los métodos basados en este enfoque se aseguran de que los datos sean destinados a un grupo, esto ocasiona distorsión en la forma de los grupos o ruido. La agrupación particional puede ser categorizada por: suave o rígido.

Los agrupamientos rígidos son métodos iterativos que crean conjuntos que no se sobreponen. Regularmente su función objetivo es la suma del cuadrado de la distancia Euclidiana. En estos algoritmos un elemento de nuestro conjunto de datos solo puede pertenecer a un grupo, es decir su grado de pertenencia es 0 o 1. Dentro de estos se tiene el K-means, basados en histograma, entre otros.

De manera contraria en los agrupamientos suaves un elemento puede pertenecer a más de un grupo con un grado de pertenencia. Esta pertenencia está en el intervalo de [0,1] y depende de la función objetivo. Algunos algoritmos populares en esta categoría son el fuzzy c-means (FCM), fuzzy c-shells, entre otros [55].

El uso de estos algoritmos en imágenes satelitales es extenso y puede abordarse de dos maneras principales. En un enfoque, se proporcionan parámetros de inicialización mediante muestras; en el otro, se permite que el algoritmo se adapte automáticamente a la imagen.

La selección del algoritmo depende de nuestro problema y nuestra imagen. En el campo de las imágenes satelitales en las imágenes con baja resolución, los pixeles de la imagen pueden representar diferentes tipos de cobertura terrestre. Por esta razón el píxel tiene un nivel elevado

de incertidumbre. Esto ocasiona que las técnicas de agrupamiento con un enfoque suave o difuso sean más adecuadas para realizar una segmentación [52].

2.2.1.1 Fuzzy C-Means

El algoritmo de agrupamiento fuzzy c-means (FCM) fue propuesto por Dunn y extendido posteriormente por Bezdek. Es un algoritmo de agrupamiento suave que evoluciono del k-means. En comparación con su sucesor el FCM establece una medida de pertenencia del grupo a todos los datos, esto ocasiona que se tenga un mejor entendimiento de cómo son nuestros datos [54].

La medida de pertenencia implementada en este algoritmo proporciona un mecanismo para manejar la superposición entre los grupos, lo que lo convierte en una técnica adecuada para abordar problemas del mundo real, donde suele ser difícil delimitar los grupos de manera precisa [56].

Este método se basa en minimizar la siguiente función objetivo (2.5) según se muestra en [57]

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2 . \tag{2.5}$$

En esta fórmula m es el exponente difuso ($1 \leq m \leq \infty$). Mientras que u_{ij} es el grado de membresía de que x_i pertenezca al grupo j, este es un elemento de la matriz de membresía $U = \left[u_{ij}\right]^{N\times C}$, adicionalmente en esta matriz se tiene que cumplir $\sum_{j=1}^{C} u_{ij} = 1$. Donde x_i es la i-ésima observación del conjunto de datos, c_j es el centro del grupo j. Por último, $\|\,.\,\|$ es cualquier medida de similitud entre los datos y el centro del grupo.

El método FCM se lleva a cabo iterativamente, al optimizar la expresión (2.5) se obtiene la ecuación (2.6), la cual actualiza el grado de membresía de los pixeles a cada uno de los grupos:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}} . \tag{2.6}$$

Al actualizarse la matriz de membresía también se tienen que actualizar los valores centrales de cada uno de los grupos, eso se realiza con le expresión (2.7)

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}\cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}} . \tag{2.7}$$

El algoritmo completo se lleva acabo de la siguiente manera:

Paso 1: determine el umbral de convergencia ε, el número de grupos c, el número de iteraciones, asignar el exponente de difusión m e inicializar la matriz de membresía.

Paso 2: Calcular los valores centrales de los grupos usando (2.7).

Paso 3: Calcular la distancia de cada punto a los valores centrales de los grupos.

***Paso 4*:** Actualizar la matriz de membresía usando (2.6).

Paso 5: Repetir pasos del 2 al 4, hasta que se cumpla el máximo de iteraciones o se cumpla la expresión en (2.8)

$$\max_{ij}\{\left|u_{ij}^{(k+1)} - u_{ij}^{k}\right|\} < \varepsilon. \quad (2.8)$$

En [57] nos menciona cuatro desventajas que debemos considerar al implementar este algoritmo, pese a las ventajas de esta técnica. Las cuales se discuten a continuación:

- Sensible al ruido: Los resultados obtenidos en imágenes con mucho ruido son pobres.
- Tiempo computacional largo: debido a que es un algoritmo iterativo, se calculan repetitivamente las distancias desde los pixeles a los centros. Por ello, el número de operaciones a realizar incrementan con el de observaciones dentro de nuestros datos. Es decir, si d es una imagen que tiene N pixeles, los cuales son usados para encontrar c grupos, por esta razón la complejidad del FCM se especifica por $O(N \times c \times d \times t)$.
- Susceptible a mínimos locales: este inconveniente surge ya que se inicializan los centros de manera aleatoria, lo que puede traer como consecuencia ser atrapado en un mínimo local. Al final de las iteraciones se pueden obtener grupos vacíos, por lo que encontrar el mínimo global es un desafío para el FCM.
- Estimación del número de grupos: como indican los pasos descritos antes, para inicializar el algoritmo se debe proporcionar a priori el número de grupos en que se designaran los pixeles. Por lo tanto, encontrar el número de grupos puede ser un desafío ya que este puede cambiar de imagen a imagen y nuestro flujo de trabajo deja de ser automático.

Con el objetivo de subsanar estos inconvenientes que tiene el FCM se han desarrollado diferentes variantes de este, algunas de estas son Enhanced FCM (EnFCM), Fast Generalized FCM, Fast and Robust FCM (FRFCM), entre otros.

2.2.1.2 Fast and Robust FCM (FRFCM)

En [58] se propone una metodología en la cual remplaza el filtro de media o mediana por la reconstrucción morfológica. Esta técnica forma parte del análisis morfológico, herramientas que ayudan a la extracción de componentes de la imagen, útiles para la representación y descripción de regiones.

El análisis morfológico simplifica las imágenes conservando las principales características. Estas también se utilizan para tareas de pre y post procesamiento. En este algoritmo la reconstrucción morfológica nos ayuda a preservar los objetos dentro de nuestra escena, y puede corresponder a un excelente pre-procesamiento de la imagen [59].

Por lo tanto, la reconstrucción morfológica es una operación que usa dos imágenes y un elemento estructural. Las imágenes se definen como marcador (marker), en la cual se realizarán las operaciones, la segunda imagen es la máscara (mask), ésta limita la reconstrucción. Adicional a las imágenes se tiene que definir el elemento estructural el cual definirá la conectividad, este puede tener diferentes figuras.

La reconstrucción morfológica es una operación iterativa en la cual definiremos al marcador como f, la máscara como g. Por lo tanto, la reconstrucción de g a partir de f es $R(f)$. El proceso iterativo es el siguiente:

1. Inicializar $h_1$como la imagen marcadora f, la cual tiene que cumplir con $g(f \subseteq g)$.
2. Crear el elemento estructural.

3. Realizar iterativamente: $h_{k+1} = (h_k \oplus B) \cap g$, hasta que $h_{k+1} = h_k$.

En el algoritmo FRFCM se utiliza la reconstrucción morfológica de cierre la cual se define como: $R^C(f)$. Esto nos indica que se realizara primeramente una operación morfológica de dilatación y posteriormente una erosión [60].

Teniendo lo anterior en consideración en la propuesta de [58] se tiene la función objetivo encontrada en (2.9)

$$J_m = \sum_{l=1}^{q} \sum_{k=1}^{c} \gamma u_{kl}^m \|\xi_l - v_k\|^2. \tag{2.9}$$

Donde u_{kl} representa el grado de pertenencia difuso de los valores de gris l, respecto a cada uno de los grupos k, también se tiene:

$$\sum_{l=1}^{q} \gamma_l = N. \tag{2.10}$$

Definiremos a ξ como la imagen reconstruida a través de la reconstrucción morfológica, a su vez $\xi_l$es un valor dentro de la escala de grises, $1 \leq l \leq q$, q simboliza el número de niveles de grises que contiene ξ. El cual es por lo general mucho más pequeño que N. ξ se define como:

$$\xi = R^C(f). \tag{2.11}$$

Donde R^C es la reconstrucción morfológica de cierre y f representa a la imagen original.

Incorporando la técnica del multiplicador de Lagrange en (2.9), se convertirá nuestro problema de optimización a un problema de optimización sin restricción el cual minimizará la siguiente función objetivo:

$$\tilde{J}_m = \sum_{l=1}^{q} \sum_{k=1}^{c} \gamma u_{kl}^m \|\xi - v_k\|^2 - \lambda \left(\sum_{k=1}^{c} u_{kl} - 1 \right). \tag{2.12}$$

En la ecuación (2.12) λ se convierte en el multiplicador de Lagrange. Sin embargo, el problema de minimizar esta función objetivo se convierte el de encontrar el punto de meseta de la función Lagrange y esto se puede lograr calculando la derivada del Lagrangiano $\tilde{J}_m$ respecto a los parámetros, por ejemplo u_{kl} y v_k.

Por lo tanto, minimizando la función objetivo (9), se obtienen las siguientes soluciones:

$$u_{kl} = \frac{\|\xi_l - v_k\|^{-2/(m-1)}}{\sum_{j=1}^{c} \|\xi_l - v_j\|^{-2/(m-1)}}, \tag{2.13}$$

$$v_k = \frac{\sum_{i=1}^{q} \gamma u_{kl}^m \xi_l}{\sum_{i=1}^{q} \gamma u_{kl}^m}. \tag{2.14}$$

De acuerdo con (2.13), se obtiene la matriz de pertenencia $U = [u_{kl}]^{c \times q}$. Para obtener una matriz U estable, se necesita realizar iterativamente la ecuación (2.13) y (2.14), hasta que $\max \{U^{(t)} - U^{(t+1)}\} < \eta$, donde η es nuestro umbral de error mínimo. Debido a que $u_{kl}^{(t)}$ es una membresía

difusa de algún valor de gris en l, respecto al grupo k, una nueva matriz de membresía $U' = [u_{kl}]^{c \times N}$ la cual corresponde a la imagen original f.

Finalmente, para obtener una mejor matriz de pertenencia y hacer más rápida la convergencia del algoritmo. Se modifica u_{ki} usando un filtro de la mediana, representado de la siguiente manera:

$$U" = med\{U'\} \tag{2.15}$$

Resumiendo lo mencionado con anterioridad el algoritmo de FRFCM es el siguiente:

Paso 1. Definir el número de grupos c, el parámetro de difusión m, el tamaño de la ventana del filtro mediana w y el criterio de finalización η.

Paso 2. Calcular una nueva imagen ξ usando (2.11), teniendo esta imagen se debe de calcular el histograma de ξ.

Paso 3. Inicializar aleatoriamente la matriz de pertenencia $U^{(0)}$.

Paso 4. Colocar el contador del bucle en $t = 0$.

Paso 5. Actualizar los centros de los grupos usando (2.14).

Paso 6. Actualizar la matriz de pertenencia $U^{(t+1)}$ usando (2.13).

Paso 7. Revisar nuestro criterio de finalización del algoritmo, si $\max\{U^{(t)} - U^{(t+1)}\} < \eta$ entonces se detiene, de forma contraria se coloca $t = t + 1$ y se regresa al Paso 5.

Paso 8. Implementar el filtro de la mediana en la matriz de pertenencia U' usando (2.15).

2.2.2 Evaluación de la segmentación

La segmentación es un problema muy importante y difícil, en diferentes campos. Esto la convierte en una etapa fundamental del procesamiento de imágenes. A consecuencia, de la importancia y el incremento de técnicas, se hace necesario contar con las técnicas de validación.

De acuerdo con [61] los métodos de evaluación se dividen en dos, evaluación subjetiva y objetiva. La evaluación subjetiva, también llamada método de observación, se lleva a cabo por un ser humano como evaluador. La desventaja de este acercamiento es la subjetividad que acompaña al evaluador. Esto ocasiona que muchas personas tengan que llevar a cabo una especialización para tener una evaluación robusta.

Las métricas de evaluaciones objetivas nos permiten medir la calidad de los resultados de la segmentación. Contrario a las subjetivas, las objetivas ya no dependen de la observación humana. Esto las convierte en métricas cuantitativas, que son empleadas para comparar los resultados de algoritmos, optimizar los parámetros de los algoritmos y validar los resultados empleando una imagen ya segmentada o datos verdaderos. Las métricas objetivas de evaluación se dividen en dos categorías: supervisadas y no supervisadas.

Las métricas no supervisadas hacen referencia a que las métricas son autónomas, debido a que no necesitan información adicional. La ventaja de estas métricas es que no se realiza una comparación directa con otro elemento. Esto beneficia cuando se quiere efectuar segmentación de muchas imágenes y aplicaciones en tiempo real. Estas métricas ayudan a la adecuación de los parámetros de algoritmos no supervisados. Algunas de estas métricas son: Raíz del Error Cuadrático Medio (RMSE), Error Medio Cuadrático (MSE), entre otros [62].

Por otro lado, las métricas supervisadas permiten evaluar la calidad del resultado utilizando

imágenes de referencia. Las imágenes de referencia son generadas por expertos o por mapas existentes. Estas métricas ejecutan una comparación o correlación del resultado con la imagen de referencia. Algunas de estas métricas son: Tasa de falsos positivos (FPR), Cohen Kappa, Índice Jaccard, entre otros [62].

2.2.2.1 Métricas supervisadas

Las métricas supervisadas pueden proporcionar evaluaciones más precisas y confiables. Esto se debe a que los resultados se obtienen comparando el resultado de la segmentación y una imagen de referencia. La imagen de referencia representa el resultado verdadero o deseado del proceso de segmentación. Esto ocasiona que los resultados de la validación sean más consistentes y comparables entre ellos, debido a que utilizan un estándar o criterio común. Adicionalmente, no son afectados por las variaciones o incertidumbre de los subgrupos creados [61].

A partir de esta clase de métricas los resultados de la comparación pueden ser positivos y negativos. Los cuales son los siguientes [63]:

- **Verdaderos Positivos (TP)**: Estos son los casos en que el modelo incluyo en el objeto los pixeles correctamente. Es decir, lo pixeles del objeto coinciden con los encontrados en la imagen de referencia como parte del objeto a segmentar.
- **Verdaderos Negativos (TN)**: Los pixeles fuera de la imagen fueron clasificados de forma correcta. En otras palabras, el modelo excluyo los elementos fuera del objeto a segmentar correctamente.
- **Falsos Positivos (FP):** En este caso en el objetivo a segmentar se incluyeron pixeles que no se encuentran en la imagen de referencia. También se conocen como errores de tipo 1.
- **Falsos Negativos (FN):** Al contrario del anterior, en este caso se omitieron pixeles que se encuentran en la imagen de referencia. En otras palabras, se asignaron como píxeles negativos que deberían de ser positivos. A este también se le conoce como errores tipo II.

En este trabajo se decidió utilizar el Índice Jaccard, coeficiente Cohen Kappa y el coeficiente de correlación de Matthews (MCC). Se incluyeron debido a su uso en diferentes trabajos consultados.

2.2.2.1.1 Índice Jaccard

El índice Jaccard se ha utilizado desde 1901, propuesto por Jaccard. En el contexto de la segmentación está relacionado a la métrica Intersección sobre Unión (IoU) [64]. Esta métrica es usada para evaluar la segmentación de objetos. Mide la similitud entre la imagen de referencia y el resultado de la imagen. En [65] definen este índice como:

$$JI = \frac{A(G \cap S)}{A(G \cup S)}. \tag{2.16}$$

En (2.16) $A(\cdot)$ corresponde a la operación de conteo, G corresponde a la imagen de referencia (ground truth), mientras que S es el resultado de nuestra segmentación. En (2.16) el numerador corresponde del número de píxeles correspondientes en los dos conjuntos, los verederos positivos (TP). El denominador representa la suma de los píxeles que corresponden a la imagen de referencia (TP), los píxeles que en el resultado de la segmentación son considerados como parte de la imagen, pero en la imagen de referencia no son parte del objeto (FP) y por último los que la

segmentación no se considera dentro del objeto, mientras que en la imagen de referencia si lo son (FN). Considerando esto último, la ecuación anterior también puede escribirse de la siguiente manera:

$$JI = \frac{TP}{TP + FN + FP}. \tag{2.17}$$

El resultado está en el rango de 0 a 1, donde el 1 representa una perfecta similitud, entre la segmentación y la imagen de referencia. Mientras que el 0 significa que no existe similitud entre ella.

2.2.2.1.2 Coeficiente Cohen Kappa

El coeficiente Cohen kappa es otra métrica supervisada que mide el grado de acuerdo entre dos segmentaciones. Se basa en la proporción de píxeles que coinciden y la proporción de píxeles que se esperaría que coincidieran por azar. Este inicialmente se propuso para cuantificar el acuerdo entre dos observadores sobre el mismo conjunto de datos, con dos o más clases [21].

La métrica también se suele usar cuando se tienen dos clases. En este contexto la ecuación para este coeficiente es la siguiente:

$$k = \frac{2 \cdot (TP \cdot TN - FP \cdot FN)}{(TP + FP) \cdot (FP + TN) + (TP + FN) \cdot (FN + TN)}. \tag{2.18}$$

El rango de valores resultantes de esta métrica es de -1 a 1, en donde -1 significa nada parecido, mientras que 1 es el mejor valor para nuestra evaluación [63]. Mientras que si se puede un valor de 0 significa que todo es consecuencia del azar. Aunque fue diseñado para las ciencias sociales se usa en trabajos de segmentación y clasificación de percepción remota [12].

2.2.2.1.3 Coeficiente de Correlación de Matthews (MCC)

El coeficiente de correlación es utilizado desde 1975. En un inicio estaba enfocado en el área de la bioquímica. Después se amplió su uso y popularidad [64]. Este coeficiente es utilizado en el aprendizaje automático como medida de calidad cuando se tiene clasificación binaria o multiclase. Esta métrica usa los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. La expresión para este coeficiente está definida de la siguiente forma:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}. \tag{2.19}$$

El valor mínimo del MCC es -1, el cual indica que la referencia y lo segmentado no tienen ningún parecido. El valor máximo es 1, indica una similitud exacta entre los elementos. Si el valor es 0, significa que la segmentación es aleatoria.

2.3 Detección de cambio de uso de suelo

La detección de cambios en el uso del suelo se refiere a la identificación de modificaciones en la cobertura terrestre y su uso a lo largo del tiempo. Este análisis se realiza comparando mapas de diferentes intervalos de tiempo [66]. Principalmente, se apoya en Sistemas de Información Geográfica (SIG). Sin embargo, para llevar a cabo este análisis, es fundamental comprender los conceptos y la diferencia entre cobertura terrestre y uso del suelo.

La cobertura terrestre hace referencia a los elementos físicos visibles en la superficie de la Tierra, como la vegetación, cuerpos de agua, suelo desnudo y la infraestructura artificial creada por el ser humano. Por otro lado, el uso del suelo describe el propósito o función de los elementos en la cobertura terrestre en relación con las actividades humanas. Esta categoría incluye usos como el residencial, agrícola, hospitalario, entre otros [67].

A lo largo del tiempo se han utilizado diversas métricas para llevar a cabo este análisis, como el uso de la Entropía de Shannon [11], el cálculo de la superficie transformada [68], la tasa de cambio anual [69] o la comparación directa entre períodos. Este tipo de estudios son esenciales para comprender el impacto de las actividades humanas en el medioambiente y en su entorno, así como los diferentes fenómenos consecuentes de estos cambios [70].

En este estudio se utilizó la tabulación cruzada, una técnica que contrasta los mapas de uso del suelo. Este método consiste en superponer los mapas con el objetivo de identificar los cambios ocurridos en las diferentes clases [71]. Al realizar esto se calcula un cambio absoluto usando la ecuación (2.20), la cual es la diferencia entre las categorías en las dos fechas, regularmente es una superficie expresada en hectáreas o en kilómetros cuadrados [72].

$$CA = A_2 - A_1 \tag{2.20}$$

Donde A_1 y A_2 son las áreas de categoría en cuestión en las fechas 1 y 2, respectivamente. Este cálculo del cambio absoluto se complementa con el cambio dinámico utilizando la ecuación (2.21). Este mide las variaciones de las categorías durante los intervalos analizados.

$$K = \frac{U_b - U_a}{U_a} \times \frac{1}{T} \times 100 \tag{2.21}$$

Donde K es el cambio dinámico del uso de suelo de una categoría en particular, U es el área de uso de suelo o cobertura terrestre, mientras que a y b representan el inicio y el final del intervalo elegido para el análisis, respectivamente. Por último, T es el número total de años comprendido en el estudio [66]. Un resultado positivo indica un aumento en la extensión de una categoría específica, mientras que un resultado negativo refleja una disminución en dicha categoría.

Adicional a los índices de cambio se obtiene una matriz de cambio, que se construye para representar la estadística de transición "de-a" durante los intervalos analizados [66]. Esta muestra las ganancias y las pérdidas de las categorías una con otra. La matriz se compone como se muestra en la Tabla 2.3, en la cual la diagonal principal representa las áreas que no han cambiado en la categoría en los años analizados. Las celdas fuera de la diagonal indican las transiciones entre categorías, se da el cambio "de-a".

Complementando los resultados obtenidos en los cambios y la matriz de cambio, se profundizó en el análisis de la expansión urbana utilizando indicadores como el Índice de Intensidad de Expansión Urbana (IIEU). Mientras que la matriz de cambio ofrece una visión general de las transiciones entre las categorías de uso de suelo, el IIEU se calculó debido a que proporciona una medida cuantitativa de la intensidad y rapidez de la expansión urbana en periodos específicos.

Tabla 2.3. Matriz de cambio que muestra la transición "de-a" de las categorías.

Tiempo 2 / Tiempo 1	Categoría 1	Categoría 2
Categoría 1	De categoría 1 a categoría 1	De categoría 1 a categoría 2
Categoría 2	De categoría 2 a categoría 1	De categoría 2 a categoría 2

2.3.1 Índice de Intensidad de Expansión Urbana (IIEU)

El proceso de expansión urbana es complejo y en algunas ocasiones es impulsado por diferentes factores. Esta complejidad ocasiona que el desarrollo no sea uniforme, revelando una preferencia de expansión urbana en una dirección [73]. Para comprender esta falta de uniformidad se han desarrollado diferentes índices como: índice de intensidad de expansión urbana [11], índice de diferenciación de expansión urbana [74] y el índice de dimensión fractal [75]. En este estudio se seleccionó el índice de intensidad de expansión urbana. Esto se debe a su utilidad para analizar cuantitativamente las diferencias de expansión en diferentes zonas y comprender las preferencias de expansión en periodos determinados [73]. El cálculo del IIEU se llevó a cabo usando la ecuación (2.22), en esta el área de estudio se divide en i zonas [11].

$$IIEU_i = \frac{\left(\frac{ULA_{i,b} - ULA_{i,a}}{t}\right)}{TLA_i} \times 100\ . \tag{2.22}$$

Donde $IIEU_i$ representa el índice de intensidad de la expansión promedio anual de la zona i, $ULA_{i,a}$ y $ULA_{i,b}$ son el área construida inicial y final, respectivamente de la zona espacial i. TLA_i es el área total de la zona espacial i. El resultado del índice se clasifica como se indica en la Tabla 2.4 [73].

Tabla 2.4. Division de los resultado de IIEU.

IIEU	Velocidad
>1.92	Desarrollo de alta velocidad
1.05 – 1.92	Desarrollo rápido
0.59 – 1.05	Desarrollo de velocidad media
0.28 – 0.59	Desarrollo de baja velocidad
0 – 0.28	Desarrollo lento

2.4 Aprendizaje Automático (ML)

La Inteligencia Artificial (AI) es un tema que recientemente ha cobrado relevancia, no sólo para los investigadores en el área. A pesar, de su amplio uso en diferentes campos, existen muchas definiciones y conceptos de AI en diferentes trabajos [76].

En [77] nos presentan diferentes conceptos de AI. Una de ellas define a la inteligencia artificial

como: la habilidad de las computadoras o máquinas de emular o imitar el comportamiento humano, con el objetivo de analizar su entorno y realizar un objetivo específico.

A pesar de los múltiples conceptos dados a la AI, en muchas ocasiones se confunde con el Aprendizaje Automático (ML). Sin embargo, la AI se refiere a un rango más grande de capacidades que intentan emular las funciones cognitivas humanas, como el aprendizaje, razonamiento y reconocimiento propio. Por otro lado, el ML es una rama dentro de la AI que se centra en la capacidad de las máquinas para aprender de los datos y mejorar su desempeño de manera autómata. Mientras que la AI abarca cualquier técnica que permita a las maquinas imitar la inteligencia humana, el ML se refiere a los algoritmos que permiten aprender y tomar decisiones basadas en la experiencia. En este contexto no se programan si no se instruyen [78].

De manera general el algoritmo del Aprendizaje Automático es el siguiente:

Paso 1. Recolectar un conjunto de entrenamiento, este consiste en una colección de características de entrada para nuestro modelo.

Paso 2. Seleccionar el tipo de modelo que queremos entrenar.

Paso 3. Entrenamiento del modelo presentándole el conjunto de entrenamiento y ajustar los parámetros del modelo.

Paso 4. Repetir el paso tres hasta que el desempeño del modelo sea satisfactorio.

Paso 5. Proporcionar al modelo entrenado un conjunto de datos desconocido para que este pueda producir nuevas respuestas.

Dentro del ML, existen diferentes enfoques para enseñar a las máquinas a aprender de los datos. Los dos más importantes son el aprendizaje supervisado y no supervisado. El aprendizaje supervisado se basa en proporcionar un conjunto de características junto con la etiqueta. Por lo tanto, la máquina aprende basándose en ejemplos. Por otro lado, el aprendizaje no supervisado trabaja con datos sin etiquetas. Esto permite a la máquina identificar patrones ocultos y agrupar la información de manera autónoma [79].

2.5 Algoritmos de Aprendizaje Supervisado

De acuerdo con [80] el objetivo del aprendizaje supervisado es aprender de las entradas x, llegando a un objetivo y. El problema se formula a partir de un conjunto de pares ordenados expresado como $\mathcal{D} = \{(\mathrm{x}_i, y_i)\}_{i=1}^{N}$. Donde, $\mathcal{D}$ se llama el conjunto de entrenamiento, x_i representa los valores de entrada, y_i la salida o etiqueta correspondiente, mientras que N es el número de ejemplos.

En este problema x_i es un vector $\mathcal{D}$-dimensional, que representa características o atributos de nuestro conjunto de datos. Este vector puede ser una estructura compleja como una imagen, una oración, una serie de tiempo, un grafo, entre otras.

De manera similar al vector x_i, la salida y_i puede ser en principio cualquier cosa. Sin embargo, se asume que es una variable categórica o nominal de un conjunto finito, expresado como $y_i \in \{1, \dots, C\}$. Cuando y_i, es un valor real, el problema es una regresión. En caso contrario, la variable es categórica el problema se conoce como clasificación.

En la expresión $y_i \in \{1, \dots, C\}$ donde C representa el número de clases. Si $C = 2$, se convierte en un problema de clasificación binaria. Comúnmente se asume que $y_i \in \{1,0\}$. Cuando $C > 2$ este se llama clasificación multiclase.

Formalizando el problema se puede ver como una aproximación de funciones. Suponemos que $y = f(x)$ para una función desconocida f. El objetivo del aprendizaje es el de estimar la función f dada un conjunto de entrenamiento con etiquetas. Posteriormente, se utiliza la estimación para realizar predicciones mediante $\hat{y} = \hat{f}(x)$. El objetivo principal es hacer estimaciones sobre entradas que desconocemos. A este proceso le conoceremos como generalización [80].

En el problema de clasificación existen diferentes algoritmos, como los Árboles de Decisión, el k-Nearest Neighbors (k-NN) y las Máquinas de Soporte Vectorial (SVM), que asignan categorías o etiquetas a los datos. Las SVM se destacan por crear fronteras de decisión óptimas entre clases y son especialmente útiles en espacios de alta dimensionalidad. También están los Modelos de Ensamble, como los Bosques Aleatorios o el Boosting, que combinan varios modelos para mejorar las predicciones.

2.5.1 Maquinas de Soporte Vectorial (SVM) de margen suave

Las máquinas de soporte vectorial fueron propuestas por Vapnik and Lerner (1963). Se introdujeron inicialmente para la clasificación binaria supervisada [81]. Esta técnica se fundamenta en la construcción de un Hiperplano de Separación Óptimo (HSO) y la construcción de los vectores de soporte.

Supongamos que tenemos un conjunto de entrenamiento $\{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_N, y_N)\}$, donde, $(\vec{x}_i, y_i)$ corresponden a las muestras para $i = 1, 2, \dots, N$ está constituido por un vector n de características y una etiqueta que indica las clases $\{\pm 1\}$ a la que pertenece a cada una de las muestras [81].

Como las clases son +1 y -1, el HSO lo define el margen máximo de separación entre las clases. Con base en esto, los vectores de soporte se definen como: $\vec{w} \cdot \vec{x} + b = +1$ y $\vec{w} \cdot \vec{x} + b = -1$, los cuales son paralelos al HSO el cual se expresa por $\vec{w} \cdot \vec{x} + b = 0$. El margen máximo de los vectores de soporte está expresado por $2/\|\vec{w}\|$. Esto se aprecia en la Figura 2.3.

Sin embargo, inicialmente las SVM fueron propuestas de manera rígida, es decir, que los datos son completamente separables. No obstante, existen datos que no lo son. En estos casos se introduce una variable de holgura ξ_i, la cual permite errores en la clasificación, pero estos errores son penalizados.

Al incluir este nuevo parámetro la SVM se suaviza. El problema primario cambia de forma y ahora se busca minimizar la expresión:

$$\min_{w,b,\xi} \left(\frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i \right). \tag{2.23}$$

Sujeta a la condición:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \dots, n\,, \tag{2.24}$$

$$\xi_i \geq 0, \quad i = 1, \dots, n\,. \tag{2.25}$$

Donde ξ_i es el parámetro de holgura y c es el parámetro de penalización de los errores de clasificación [26]. En este nuevo enfoque, el objetivo es encontrar el HSO que minimice los errores de clasificación y maximice la separación entre los vectores de soporte. Mientras el valor de c se incrementa, se obtiene un margen estrecho, esto minimiza el número de clasificaciones erróneas. Por otro lado, si c disminuye, se están permitiendo más clasificaciones erróneas [82].

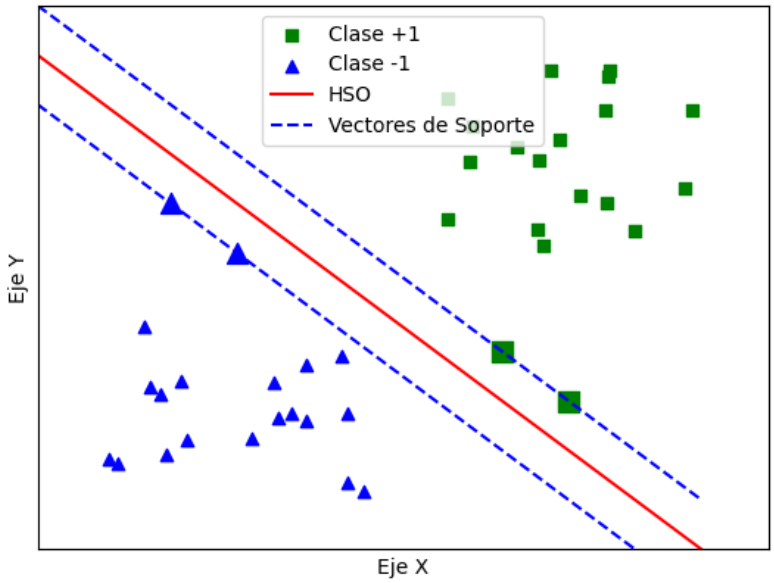

Figura 2.3. Hiperplano de Separación Óptimo y vectores de soporte de las Maquinas de Soporte Vectorial.

Para abordar la no linealidad, los datos son transportados a un espacio de alta dimensionalidad. Esto se logra realizando una proyección matemática de los datos usando un kernel. Dicho kernel se puede definir como: $k(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$. Algunas funciones de kernel populares son la lineal $k(x,y)=x_i\cdot x_j$, función de base radial (RBF) $k(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|^2)$, y la función polinomial $k(x,y)=(1+x_i\cdot x_j)^q$. Donde γ y q son parámetros correspondientes al kernel [26].

2.5.2 Bosques Aleatorios (RF)

Los Bosques Aleatorios (RF) es un método de aprendizaje por ensamble para clasificación y regresión [83]. Este algoritmo crea múltiples arboles de decisión de forma aleatoria [78] durante la fase de entrenamiento, posteriormente combina los resultados para obtener un modelo más robusto comparado con cada árbol de decisión individualmente [84].

El algoritmo de los Bosques Aleatorios para el problema de clasificación, es el siguiente [85]:

Paso 1. Para cada árbol $b = 1$ a B:

a. Extraer una muestra Bootstrap $\mathbf{Z}^*$ de tamaño N de los datos de entrenemiento.

b. Crear un árbol de bosque aleatorio T_b a partir de los datos bootstrap, repetir recursivamente los siguientes pasos para cada nodo terminal del árbol, hasta que se alcance el tamaño mínimo de nodo n_{min}:

   i. Seleccionar m variables al azar de las p variables.

   ii. Elegir la mejor variable/punto de división en las m.

   iii. Dividir el nodo en dos nodos hijos.

Paso 2. Generar el conjunto de árboles $\{T_b\}_{b=1}^{B}$

Para hacer una predicción en un nuevo punto x:

Sea $\hat{C}_b(x)$ la prediccion de clase del arbol numero b del bosque aleatorio. Entonces $\hat{C}_{rf}^{B}(x) = voto\ mayoritario\ \{\hat{C}_b(x)\}_{b=1}^{B}$.

Como se puede observar en el algoritmo, por cada árbol se crea una muestra Bootstrap. La muestra se crea seleccionando instancias de la población de forma aleatoria con reemplazo. Esto implica que existen instancias que se seleccionarán más de una vez y otras que nunca se seleccionarán. A estas últimas se les conoce como “out of the bag". Estas se usan para validar el modelo [84].

Las muestras Bootstrap forman parte de la técnica de ensamble llamada Bagging. Se basa en generar múltiples modelos paralelos con este tipo de muestras. Los árboles se benefician de esta técnica debido a que estos contienen mucho ruido. Su uso minimiza la varianza, esto ayuda a prevenir el sobre ajuste y mejorar la capacidad de generalización [85]. La intención del RF al usar la técnica de bagging es descorrelacionar los árboles generados. Para mejorar la descorrelación, el algoritmo también selecciona un subconjunto de variables en cada división [85].

Este algoritmo es ampliamente utilizado en los estudios de percepción remota, debido a su buen desempeño en conjuntos de datos grandes, su facilidad de entrenamiento, y su aplicabilidad tanto en regresión como en clasificación. Sin embargo, entre sus desventajas se encuentra su susceptibilidad al sobreajuste y la falta de control sobre su funcionamiento interno, lo que lo convierte en una caja negra [84].

2.5.3 Boosting Categórico (catBoost)

Boosting Categórico o CatBoost es un algoritmo de aprendizaje automático desarrollado por Yandex [86]. Es un algoritmo de ensamble con base en la metodología del Boosting del gradiente usando árboles de decisión. Los algoritmos Boosting combinan diferentes modelos débiles, un poco mejores que el azar, para formar al modelo fuerte de manera iterativa [83].

Algunas de las características que diferencian al catBoost es el manejo de las variables categóricas. Las variables que son categorías son cambiadas por valores numéricos usando las Estadísticas Ordenadas del Objetivo para evitar la fuga de información del objetivo. Esto sucede cuando los datos de entrenamiento contienen información futura o que no estaría disponible al momento de hacer predicciones, lo que puede ocasionar una falta de generalización en nuestro modelo. Adicionalmente, el algoritmo selecciona o realiza un conjunto de datos permutados, en cada iteración se toma un conjunto diferente [87]. Con lo anterior tenemos que, sea $\sigma = (\sigma_1, \dots, \sigma_n)$ la permutación de nuestros datos, entonces la categoría $x_{\sigma_p,k}$ se sustituye con la expresión [86]:

$$\frac{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k} = x_{\sigma_p,k}\right] Y_{\sigma_j} + a \cdot P}{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k} = x_{\sigma_p,k}\right] + a}. \tag{2.26}$$

Donde $x_{\sigma_p,k}$ representa la $k - ésima$ observación permutada en la posición p, Y_{σ_j} es el valor de la variable objetivo correspondiente a la observación permutada en la posición j, a es un parámetro de suavizado y P es la media global de la variable objetivo. Por último, $\sum_{j=1}^{p-1}\left[x_{\sigma_j,k} = x_{\sigma_p,k}\right]$ cuenta cuantas observaciones anteriores tienen la misma categoría que la observación actual.

Para evitar el sobreajuste y mejorar el rendimiento, CatBoost sigue los principios del Boosting del Gradiente, simplificando el algoritmo mediante la creación de árboles simétricos. Esto

significa que el segundo nodo es idéntico en ambos lados del árbol. Estos árboles, al estar equilibrados, son menos susceptibles al sobreajuste y además permiten acelerar considerablemente la ejecución durante la fase de predicción [86].

Conociendo lo anterior de manera general el algoritmo de CatBoost es el siguiente:

Entradas. $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, I, α, L, s

Paso 1. Crear $s + 1$ permutaciones $[\sigma_0, \sigma_1, \dots, \sigma_s]$ del conjunto de datos, se guarda en σ_r.

Paso 2. Se inicializa el modelo $F_0(x)$ (con la permutacion σ_0) con valor 0 para cada instancia i.

Paso 3. Para $t = 1$ hasta I:

i. Calcular los residuos para cada instancia x_i,

$Residuos_i = -\frac{\partial L(y_i, F_{t-1}(x_i))}{\partial F_{t-1}(x_i)}$.

ii. Construir un nuevo árbol basado en los residuos T_t.

iii. Actualizar las predicciones:

1. Para cada hoja j en el árbol T_t calcular:

$Prediccion_t(j) = \frac{Residuos\ en\ la\ hoja}{\#Instancias\ en\ la\ hoja}$.

2. Actualizar las predicciones del modelo:

$F_t(x) \leftarrow F_{t-1}(x) + \alpha \cdot T_t(x)$, donde $T_t(x)$ es la prediccion del nuevo árbol para instancia x.

Paso 4. El modelo final se obtiene como la suma de todos los árboles ajustados:

$F(\mathrm{x}) = \sum_{t=1}^{I} \alpha \cdot T_t(x)$.

En este algoritmo se tiene que $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ es el conjunto de datos, I numero de árboles, α tasa de aprendizaje, L función de pérdida y s numero de conjuntos de datos permutados.

CatBoost es un algoritmo robusto con alta precisión. Es capaz de manejar grandes conjuntos de datos con características complejas. Además, reduce el sobreajuste mediante regularizaciones avanzadas y admite tanto regresiones como clasificaciones.

Este algoritmo emplea una serie de modelos base que han sido entrenados en varias permutaciones aleatorias de los datos, excepto para la instancia que se está evaluando. Cada iteración de boosting genera modelos que estiman los gradientes para las iteraciones posteriores. La implementación optimiza este procedimiento al manejar todas las combinaciones con un solo modelo por iteración. CatBoost facilita el tratamiento de características categóricas convirtiéndolas en números que representan el valor esperado para cada categoría [83]. Esto se realiza mediante la alteración aleatoria de los datos de entrenamiento para evitar el sobreajuste.

2.6 Evaluación

Al igual que la segmentación, los modelos de aprendizaje automático deben de ser evaluados. La evaluación consiste en medir la capacidad del modelo entrenado para realizar su tarea. La elección de las métricas de evaluación depende del contexto del problema, así como el conjunto de datos [88]. La evaluación de los modelos se realizó usando las métricas supervisadas detalladas

en la Sección 2.2.2.1.

Las métricas se calculan usando la matriz de confusión, está compuesta por los verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN). Existen diferentes métricas como; exactitud, precisión, curvas ROC, entre otros.

2.6.1 Exactitud

La exactitud es una métrica ampliamente usada en el aprendizaje automático. Mide la proporción de predicciones correctas realizadas por un modelo, respecto al total de muestras. Es útil en conjuntos de datos balanceados, en escenarios donde las clases están desbalanceadas, la exactitud puede ser engañosa [88]. Se calcula con a ecuación (2.27).

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.27}$$

2.6.2 Precisión

Métrica que indica el porcentaje de las predicciones positivas realizadas por el modelo que fueron correctas. Se calcula con la ecuación (2.28). Es especialmente útil cuando el costo de los falsos positivos es alto, ya que se enfoca en la calidad de las predicciones positivas [88].

$$Precisión = \frac{TP}{TP + FP}. \tag{2.28}$$

2.6.3 Recall

Mide la proporción de las instancias verdaderamente positivas fueron correctamente identificadas por el modelo. Esta métrica es crucial cuando es más importante identificar todos los casos positivos, incluso si se producen algunos falsos positivos. Se calcula con la ecuación (2.29). Se usa para otras métricas, esa es una razón para su cálculo [88].

$$Recall = \frac{TP}{TP + FN}. \tag{2.29}$$

2.6.4 F1-Score

El F1-Score es la media armónica entre la precisión y el recall. Métrica que equilibra ambas, siendo útil cuando se desea un compromiso entre la precisión y la sensibilidad, especialmente en escenarios de clases desbalanceadas. Un F1-Score alto indica que el modelo tiene tanto una buena precisión como un buen recall [88]. Se calcula con la expresión:

$$F1 - Score = 2 \cdot \frac{Precisión \cdot Recall}{Precisión + Recall}. \tag{2.30}$$

2.7 Inteligencia Artificial Explicable

En los últimos años, se ha observado un aumento en la popularidad del aprendizaje profundo

o Deep Learning. Sin embargo, estos algoritmos suelen considerarse un "problema de caja negra". Esto significa que conocemos las entradas y las salidas, pero es difícil comprender completamente el proceso que realiza el algoritmo para tomar decisiones o llegar a esas salidas.

Con el objetivo de enfrentar este problema, surgió la inteligencia artificial explicable. Estas técnicas buscan comprender los algoritmos de caja negra sin sacrificar el rendimiento ni las ventajas que ofrecen [89]. No obstante, debido a que todavía son técnicas en desarrollo, se deben de tomar como sugerencia.

La inteligencia artificial explicable se divide en dos enfoques post-hoc y ante-hoc. Los Post-hoc son las técnicas que analizan e interpretan la decisión de los modelos. Esto quiere decir que se aplican después del proceso de entrenamiento y predicción. Ayuda a tener una visión de cómo el modelo utilizó las características para emitir una decisión en las salidas [90]. Algunas de las técnicas más conocidas son SHAP, LIME y análisis de la importancia de características.

Por otro lado, los métodos ante-hoc se enfocan en incorporar interpretabilidad directamente en la arquitectura de los algoritmos de aprendizaje automático. La inclusión permite un entendimiento del proceso completo de entrenamiento y salida [90]. Algunos de estos son los modelos simples de Árboles de Decisión pequeños y regresión lineal.

2.7.1 Shapley Additive exPlanations (SHAP)

Shapley Additive exPlanations o simplemente SHAP por sus siglas en inglés, propuesto por Lundberg y Lee [91], se utiliza para asignar un valor de importancia a cada característica en relación con una predicción específica. Este enfoque se basa en la teoría de los valores Shapley, que proviene de la teoría de juegos cooperativos. Los valores Shapley miden la contribución marginal de cada característica al resultado final [6], se calcula con la ecuación (2.31):

$$\Phi_i = \sum_{S \in N} \frac{|S|!\,(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]\,. \tag{2.31}$$

Donde N es el número de instancias que tienen n características, S representa un sub-conjunto de las características n y v representa las características de entrada del conjunto S.

2.8 Principales estudios relacionados

En la revisión de la literatura, se identificaron diferentes enfoques para la predicción de la expansión urbana. Uno de estos enfoques es el uso de modelos matemáticos, como los autómatas celulares, y otro es la aplicación de algoritmos de aprendizaje automático.

Algunos de los estudios corresponden a los realizados por [6], [25]. Estos estudios se llevaron a cabo en Corea del Sur; [25] se centró en la zona metropolitana de Seúl, mientras que [6] abarcó todo el país. Ambos trabajos emplearon el algoritmo basado en Boosting de gradiente, XGBoost. Posterior a la predicción, se utilizaron los valores SHAP para comprender cómo el algoritmo generó las predicciones.

Los estudios [6], [25] abarcan el periodo de 2007 a 2019 e incluyen variables de uso de suelo y cobertura terrestre, topográficas, socioeconómicas y ambientales. La evaluación del modelo en [25] se realizó utilizando métricas como exactitud, recall, precisión y F1-Score; sin embargo, sólo se reportó el resultado de la exactitud, que se situó entre el 90 % y el 95 % para todos los modelos entrenados. Por otro lado, el estudio [6] reportó una exactitud general del 82.54 %. En cuanto a

los valores SHAP, ambos estudios concuerdan en que las variables más importantes fueron las relacionadas con el uso de suelo y las características del terreno.

Por otro lado, se encuentran los trabajos [26], [28] realizados en diferentes ciudades de Estados Unidos. En [26] la predicción de expansión urbana se realizó en el condado de Guilford en Carolina del Norte. El periodo de estudio fue de 2001 al 2011. En este se utilizaron variables de uso de suelo y cobertura terrestre, densidad de población y distancias Euclidianas a diferentes elementos como: calles, carreteras, zona urbana, entre otros.

Posterior a la selección de variables, el estudio [26] utilizo el algoritmo de Máquinas de Soporte Vectorial para la predicción de expansión urbana. Se llevó a cabo una búsqueda exhaustiva de parámetros y diferentes técnicas de sub muestreo con el objetivo de encontrar el mejor modelo. El mejor modelo encontrado se reportó con una exactitud de un 85 %, utilizando el kernel radial basis function, valor de penalización de 1 y valor 2 en el parámetro del kernel.

El estudio [28] se efectuó en la región del Triángulo, en el estado de Carolina del Norte. Este se realizó en el periodo de 2001 al 2011. Las variables se obtuvieron para los años 2001, 2006 y 2011. Las variables seleccionadas fueron uso de suelo y cobertura terrestre, modelo digital de elevaciones, densidad de población, distancias Euclidianas a diferentes estructuras, entre otros.

Posterior a la selección y preprocesamiento de las variables, el algoritmo de árboles de decisión se seleccionó. En los resultados mencionados en el trabajo [28], el modelo alcanzó una exactitud de un 94 % y un coeficiente kappa de 0.80. Realizado el modelo se utilizó el algoritmo CART DT para conocer las variables que tienen más impacto en las decisiones. Este indica que la proximidad a áreas construidas, la proximidad a autopistas, el tipo actual de uso de suelo y cobertura terrestre, la elevación y la distancia a cuerpos de agua son las variables predictoras más significativas para la predicción de la expansión urbana en el área bajo estudio.

A pesar de que estos estudios son relevantes, se realizaron fuera de México. Por esta razón, se exploraron investigaciones que aplican el enfoque de autómatas celulares, con el objetivo de identificar posibles variables a utilizar. Los trabajos [5], [21] están enfocados en el área metropolitana de Toluca y tienen como objetivo apoyar la toma de decisiones relacionadas con la expansión urbana tanto en instituciones públicas como privadas.

El estudio [21] está basado en autómatas celulares y presenta dos escenarios. El primero refleja una expansión que sigue los patrones y tendencias observados hasta el momento, considerando además valores ecológicos. El segundo escenario también sigue las tendencias pasadas, pero incorpora nuevos valores ecológicos, tomando en cuenta la protección de recursos hídricos.

El periodo analizado en el estudio [21] abarca desde 2003 hasta 2017. La validación de los escenarios se realizó utilizando el coeficiente Kappa de Cohen y el índice Jaccard. De las 256 simulaciones realizadas, la seleccionada obtuvo un índice Kappa de 0.62 y un índice Jaccard de 0.72.

El estudio [5] se realizó en la misma área de estudio; sin embargo, en este se añadió una regresión geográficamente ponderada. Se generaron escenarios de expansión que priorizan la zona centro, el pericentro y la periferia. Además, se introdujeron restricciones a la expansión urbana, que impiden ocupar áreas ya desarrolladas, espacios públicos, carreteras, zonas protegidas, entre otras. Estas restricciones aseguran que la ciudad se expanda únicamente sobre suelo vacante sin restricciones.

Cápitulo 3. Método y propuesta de investigación

3.1 Modelo de investigación

Para comprender mejor la expansión urbana, se pueden emplear diversos métodos y técnicas de investigación. En este estudio, se adoptó un paradigma pragmático, el cual se enfoca en la utilidad y las consecuencias de las acciones, con el objetivo de resolver problemas prácticos y generar cambios positivos. Este enfoque nos permitió integrar diferentes tipos de datos y técnicas de análisis, centrándonos en la solución de problemas reales como la expansión urbana desordenada, la planificación urbana sostenible, y los impactos ambientales y sociales.

El alcance de esta investigación es explicativo, ya que busca explicar por qué y cómo se producen ciertos fenómenos [24]. Esto nos permitió adoptar una hipótesis causal y un diseño cuantitativo no experimental longitudinal.

El diseño cuantitativo no experimental es una estrategia metodológica que nos permite manejar y analizar datos medibles. Sin embargo, este tipo de diseño se basa en la observación y descripción de variables de interés sin intervención ni manipulación. Dado que no podemos manipular las variables, ya que se trata de eventos pasados, las variables de este estudio serán los cambios de expansión observados a lo largo del tiempo. Además, se incluirán datos de censos previos, reforzando el carácter longitudinal del diseño al recolectar datos de diferentes momentos, específicamente en distintos años [92].

La población se compone de todas las áreas urbanas del estado de Zacatecas. Sin embargo, las zonas no tienen la misma probabilidad de ser seleccionadas en la muestra, esto se debe a la disponibilidad de datos. La zona seleccionada fue la zona metropolitana Zacatecas – Guadalupe.

El área bajo estudio se seleccionó considerando las necesidades y desventajas que conlleva un área metropolitana. Adicionalmente, es la zona más habitada del estado de Zacatecas. Lo cual la convierte en un área importante en el estado.

Debido a la naturaleza del estudio, objetivos y preguntas de investigación, se definió una muestra no probabilística. Esto se debe a que nosotros seleccionamos la muestra en consecuencia de la disponibilidad de los datos y consulta bibliográfica. No obstante, es fundamental considerar que para la generalización de este trabajo se tiene que pasar por una validación dentro de la nueva área de estudio.

En esta investigación, la recolección de datos se relaciona con los períodos seleccionados, específicamente los años 2000, 2010 y 2020. La selección se fundamentó en la disponibilidad de los datos recopilados por el Instituto Nacional de Estadística y Geografía (INEGI) como consecuencia de los censos.

En virtud de lo anterior, nuestras principales fuentes de obtención de información fueron las imágenes satelitales multiespectrales de uso libre de la misión Landsat, así como los diversos productos desarrollados y distribuidos por el INEGI. La elección de estos datos se debe a que son datos de uso libre y habitual en los trabajos relacionados. Además, se puede obtener información valiosa con escasos recursos.

Se realizó un preprocesamiento de las imágenes obtenidas y los datos generados por INEGI. Este procedimiento se llevó a cabo mediante la utilización de un Sistema de Información Geográfica (GIS, por sus siglas en inglés), de código abierto, denominado QGIS. En este proceso, nos

aseguramos de que las imágenes tengan la misma resolución espacial y el mismo sistema de coordenadas. La segmentación y el modelo predictivo se efectuaron en Python.

El análisis de los datos se realizó con herramientas de código abierto (QGIS y Python). En este proceso se aplicaron histogramas a las imágenes para obtener un conocimiento de lo que prevalece en nuestra escena y la cantidad de píxeles que tendrá nuestra imagen. Con esta información podemos conocer mejor nuestra imagen y decidir si se aplicara un filtro que nos ayude a mejorar el brillo, el contraste y la exposición de la imagen.

Se ha llevado a cabo el análisis de nuestras imágenes, se realizó el preprocesamiento y después su segmentación. Debido a que la resolución espacial de nuestras imágenes es de 30 m para el proceso de segmentación se utilizó un método de agrupamiento difuso. Teniendo nuestro conjunto de datos se aplicaron las técnicas de aprendizaje automático a nuestros datos, posteriormente se seleccionaron los mejores rendimientos y se realizó la predicción al año 2030.

3.2 Descripción de la propuesta

Para abordar la problemática planteada en esta investigación, se siguió el proceso ilustrado en la Figura 3.1. La cual se divide en tres fases, la primera es la obtención de datos (imágenes y otros datos), en esta etapa se realizó el preprocesamiento de los mismos. En la segunda fase se aplicaron los algoritmos de aprendizaje automático con el objetivo de construir los modelos predictivos de expansión urbana. Por último, en la fase tres se realizó una interpretación de la predicción de la expansión urbana usando técnicas de inteligencia artificial explicable y la comparación de los mejores modelos de cada uno de los algoritmos.

3.2.1 Primera Fase

3.2.1.1 Datos

Los datos usados fueron dos tipos diferentes de imágenes satelitales multiespectrales y datos tabulares resultados de los censos poblacionales. Las imágenes satelitales multiespectrales utilizadas en este estudio incluyen el producto geo-mediana nacional, elaborado por INEGI (según se indica en el Capítulo 2 de este trabajo). Estas imágenes ya se han procesado previamente mediante correcciones atmosféricas y geométrica. Poseen seis bandas (rojo, azul, verde e infrarrojos cercanos y medios) disponen de una resolución de 5000 x 5000 píxeles, contenidos en una imagen Tagged Image File Format (TIFF). El píxel tiene una medida de 30 x 30 m, por lo tanto, un píxel representa 900 metros cuadrados en el terreno. La imagen se encuentra en la proyección de Albers en grados, empleando el elipsoide GRS80. Con índice interno de INEGI: MX_014009. Los límites geográficos de la imagen son:

- Oeste: -102.97961°.
- Este: -101.51019°.
- Sur: 21.50714°.
- Norte: 22.85708°.

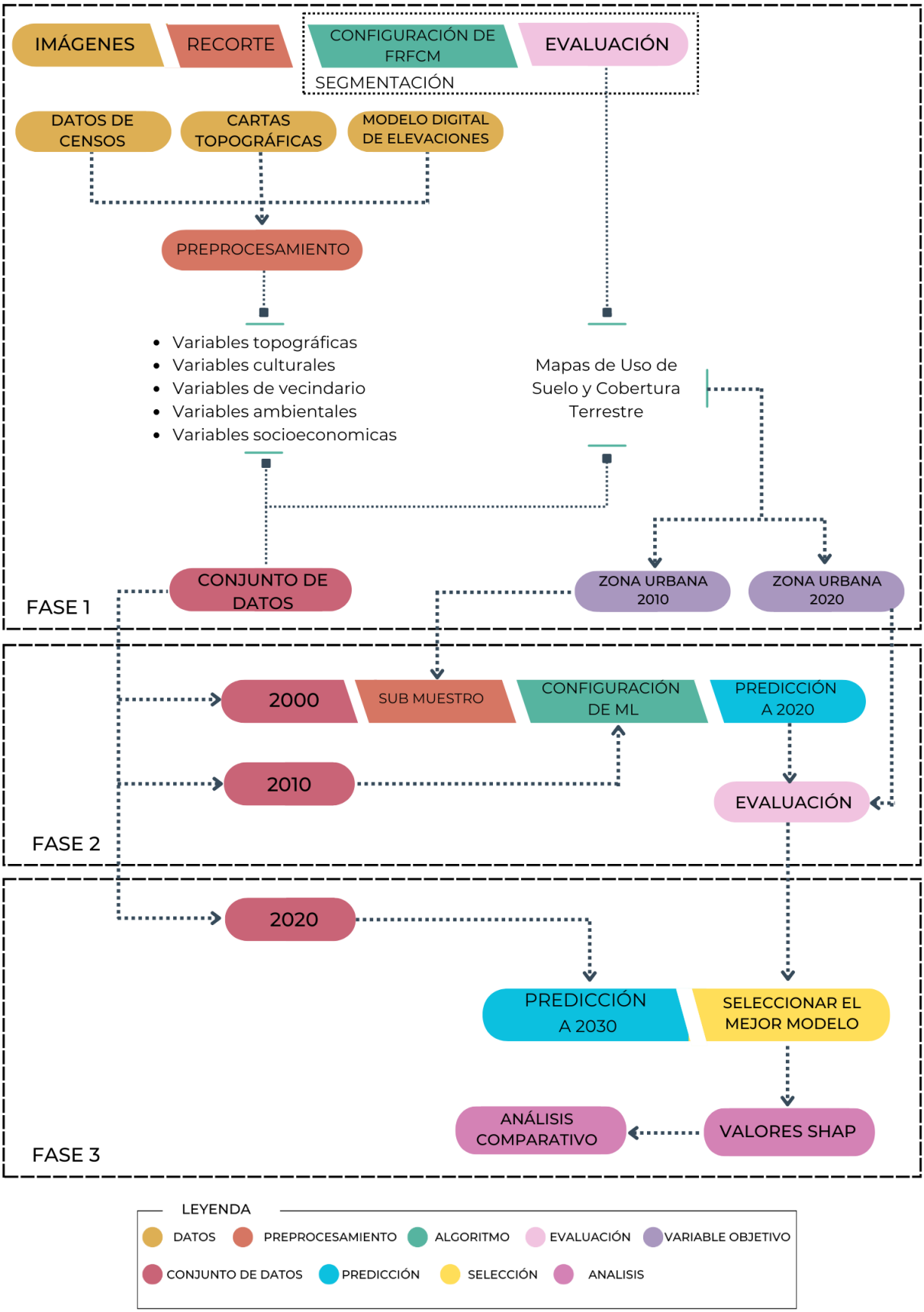

Figura 3.1. Modelo o esquema general de investigación.

Las ventajas en el uso de estas imágenes es que están libres de ruido y de nubes. El producto de geo-mediana se elabora de múltiples imágenes y misiones Landsat, en función del año seleccionado. En este estudio se emplearon los años 2000, 2010 y 2020 la descripción se encentra en la Tabla 3.1.

Tabla 3.1. Metadatos de las imágenes utilizadas en dependencia del año.

Año	Cantidad de imágenes usadas	Misiones
2000	110	Landsat 5 y 7
2010	107	Landsat 5 y 7
2020	158	Landsat 7 y 8

Adicionalmente, se recolectaron datos tabulares resultantes de los censos del año 2000 [93], 2010 [94] y 2020 [95]. Los datos colectados fueron la población total, total de viviendas, población religiosa, cantidad de personas ocupadas, viviendas habitadas, ingresos trimestrales por vivienda y el número de casas con jefatura femenina y masculina.

Con el objetivo de crear el conjunto de datos, se recolectaron cartas topográficas de escala 1:50000 creadas por el INEGI [96]. Las cartas topográficas nos brindarán información sobre diferentes datos espaciales presentes en la zona, como vías de comunicación, localidades, áreas verdes, instituciones educativas, entre otras. En este tipo de datos se priorizaron las cartas en formato Shapefile o TIFF. Estos recursos están generados en la proyección Universal Transversa de Mercator (UTM). La edición se trató de que fuera lo más cercano a los años de estudios. Los archivos descargados para este trabajo se encuentran en la Tabla 3.2. Estos archivos son puntos y líneas georreferenciadas que nos ayudaran a calcular las variables de distancias.

Tabla 3.2. Cartas topograficas utilizadas para la creación del conjunto de datos.

Clave	Titulo	Escala	Edición	Tipo de archivo
F13B58	Conjunto de datos vectoriales de información topográfica F13B58 Zacatecas escala 1:50 000 serie II [97]	1:50 000	1995	Shapefile
F13B58d	Carta topográfica. F13B58d [98]	1:20 000	2009	Shapefile
F13B58e	Carta topográfica. F13B58e [99]	1:20 000	2009	Shapefile
F13B58f	Carta topográfica. F13B58f [100]	1:20 000	2009	Shapefile
F13B58	Conjunto de Datos Vectoriales de Información Topográfica F13B58 Zacatecas escala 1:50 000, 2019 [101]	1:50 000	2019	Shapefile
F13B68	Conjunto de datos vectoriales de información topográfica F13B68 Guadalupe escala 1:50 000 serie II [102]	1:50 000	1995	Shapefile
F13B68a	Carta topográfica. F13B68a [103]	1:20 000	2009	Shapefile
F13B68b	Carta topográfica. F13B68b [104]	1:20 000	2009	Shapefile
F13B68	Conjunto de Datos Vectoriales de Información Topográfica F13B68 Guadalupe escala 1:50 000, 2019 [105]	1:50 000	2019	Shapefile

Por último, otros datos adicionales recolectados fueron los Modelos Digital de Elevación (DEM). El DEM seleccionado fue el ASTER GDEM v3 (Advanced Spaceborne Thermal Emission and Reflection Radiometer Global Digital Elevation Model) que es un modelo digital de elevación

global derivado del sensor ASTER de la misión Terra, una colaboración entre la NASA y el METI de Japón. Cubre el 99 % de la superficie terrestre, entre las latitudes 83° norte y 83° sur, con una resolución espacial de 30 metros aproximadamente y un error vertical medio que varía entre ±8 y 17 metros, dependiendo de la región. El modelo se distribuye en formato Geo TIFF, proyectado en coordenadas geográficas bajo el sistema de referencia WGS84. Los datos se recopilaron entre 1999 y 2011, con mejoras en versiones posteriores, y está disponible gratuitamente en plataformas como Earthdata y USGS EarthExplorer. La precisión vertical es mejor que ±20 metros en áreas con terreno sencillo y sin vegetación densa [106].

3.2.1.2 Preprocesamiento

Una vez recolectados los datos se realizó un preprocesamiento de los mismos. Las imágenes Geo medianas ya fueron previamente preprocesados por INEGI. Por lo tanto, en esta etapa se recortarán las imágenes. El recorte se realizará para adecuar la imagen al área de estudio. Para lograr esto se realizó un polígono en la herramienta de código abierto QGIS. En este software se realizó el recorte y se trasladó la imagen de la proyección Albers a la WGS 84 / UTM zona 13N.

La imagen recortada tiene un ancho y alto de 838 x 429, respectivamente. Las coordenadas finales del recorte se encuentran en la Tabla 3.3. Las imágenes resultantes son las encontradas en la Figura 3.2, Figura 3.2a corresponde al año 2000, mientras que Figura 3.2b y Figura 3.2c, a los años 2010 y 2020, respectivamente.

Tabla 3.3. Coordenadas de la imagen recortada a utilizar en coordenadas UTM zona 13N en metros.

Vértice	X	Y
1	738365.56502	2512539.43972
2	763511.34926	2512539.43972
3	763511.34926	2525412.40087
4	738365.56502	2525412.40087

El preprocesamiento de los datos cartas topográficas, datos de censos y DEM, requirió de una detección de datos atípicos, recorte de datos vectoriales. Los datos vectoriales como los derivados de las cartas topográficas se preprocesaron usando QGIS, en los cuales se detectaron las vías de comunicación, escuelas, áreas verdes, cementerios, entre otros. El DEM también se recortó usando QGIS, este recorte tiene un ancho y altura de 838 x 429. El recorte se realizó usando las coordenadas encontradas en la Tabla 3.3. A partir del DEM se calculó la pendiente usando QGIS, Los datos de censos fueron limpiados y se realizaron cálculos para realizar las variables de densidad. Los datos también fueron normalizados, con esto los datos se encuentran de 0 a 1.

3.2.1.3 Segmentación

Realizado el preprocesamiento, enseguida se llevará a cabo una etapa de segmentación. En esta etapa se utilizó el algoritmo FRFCM, descrito anteriormente. Este algoritmo es de agrupamiento difuso. La elección es debido a esta resolución espacial (30 x 30m). Con esta resolución existe la posibilidad de que el píxel contenga elementos de varios grupos. Los parámetros del algoritmo fueron los siguientes: el elemento estructural de la MR fue en forma de disco con radio de cuatro, se eligieron un total de ocho grupos, con un grado de membresía de dos. En la Tabla 3.4 se encuentran los parámetros del algoritmo.

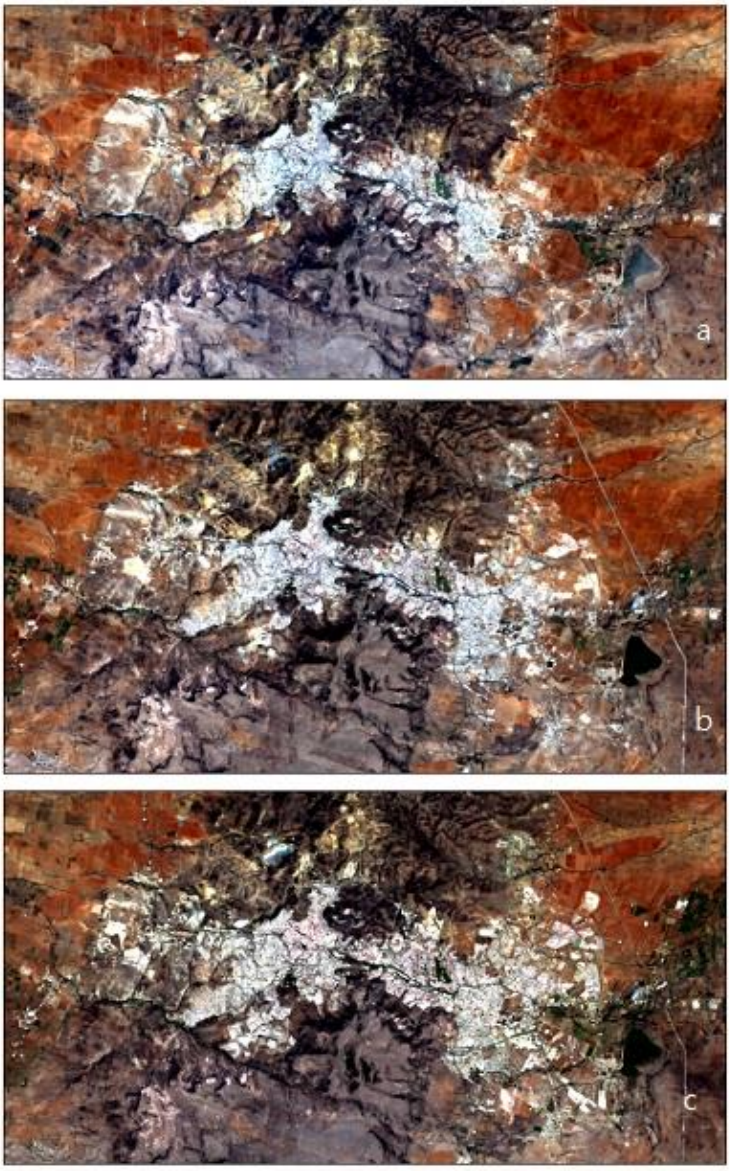

Figura 3.2. Imágenes recortadas en RGB de la zona metropolitana Zacatecas – Guadalupe, en los diferentes años a usar en el estudio. a) 2000, b) 2010, c) 2020.

Tabla 3.4. Parámetros del algoritmo FRFCM

Parámetro	**Valor**
Elemento estructural (MR)	Disco, radio cuatro
Grupos	8
Exponente de difusión	2
Criterio de finalización (error η)	1e-5
Máximo de iteraciones	100
Ventana del filtro mediana	5

La selección del elemento estructural se debe a que es lo recomendado por [58]. Igualmente, el exponente de difusión se elegio porque es el valor más utilizado. Al igual que el criterio de finalización del algoritmo (error η), el cual se utiliza en el cálculo de la ecuación $\max\{U^{(t)} - U^{(t+1)}\} < \eta$. Los demás parámetros (grupos, ventana del filtro mediana y el radio

del elemento estructural) se eligieron realizando una búsqueda exhaustiva e iterativa, en la cual se fueron modificando hasta encontrar los mejores parámetros.

3.2.1.4 Evaluación de la segmentación

La evaluación de la segmentación se realizó usando técnicas de evaluación supervisadas. Esto implica que se tiene una máscara segmentada de referencia de lo que se considera zona urbana para los años 2000, 2010 y 2020. Las imágenes de referencia se crearon a mano en el software QGIS. Como primer acercamiento para su creación, se obtuvo una imagen con resolución 15 x 15 m, para cada año. Estas nuevas imágenes se realizaron con imágenes Landsat con la fusión de la banda pancromática. Para mejorar la calidad de la imagen, se buscaron otros recursos según el año en cuestión. En el año 2000 se utilizó una ortofoto generada por INEGI, mientras que con el año 2010 y 2020, se utilizó el modelo digital de superficie. Generadas las imágenes de referencia, se calculan las métricas de evaluación. Las seleccionadas fueron el Índice de Jaccard, el coeficiente Cohen Kappa y el MCC. Teniendo la validación de la segmentación se realizaron los mapas de uso de suelo.

3.2.1.5 Conjunto de datos

Con el fin de examinar y ejecutar la predicción de la zona Zacatecas-Guadalupe, se elaboró un conjunto de datos con veinte variables compuesto por variables topográficas, socioeconómicas, ambientales, culturales y de vecindario de los años 2000 a 2020 (Tabla 3.5). Estas variables son imágenes ajustadas a una resolución espacial de 30m x 30m por pixel, esto se realizó para ser consistentes con las imágenes Geomediana usadas en la segmentación.

Las variables se seleccionaron tomando en cuenta algunos estudios relacionados [5], [6], [21], [22], [26], [28], estadísticas religiosas del estado [107] y otras características poblacionales. Las variables se realizaron con los datos recolectados y clasificados en seis categorías que describen a las mismas. Las categorías son topografía, uso de suelo, ambientales, socioeconómicas, culturales y de vecindario.

Las variables dentro de la categoría ambiental son las distancias a áreas verdes/parques. Dentro de la categoría socioeconómica se encuentran variables que describen aspectos sociales de la población. En esta categoría se tienen variables como densidad de población, densidad de población ocupada, densidad de jefatura femenina y jefatura masculina, ingresos trimestrales por vivienda y el número de viviendas por píxel.

En la categoría cultural se encuentran variables como distancias a templos/lugares de culto, distancias a cementerios, distancias a escuelas y densidad de población religiosa. Por último, se tiene la variable de vecindario, la cual es la probabilidad de composición del píxel. Este representa la probabilidad de cambio dependiendo de los píxeles en un vecindario de 3 x 3.

En la categoría de topografía se encuentran variables que describen el terreno. En esta se encuentra el DEM, pendiente del terreno y la dirección de la pendiente. Uso de suelo, en esta categoría se encuentran las clasificaciones de mapas de uso de suelo, la distancia a áreas construidas, distancia a calles y carreteras, distancia a vías férreas y distancias al centro de las ciudades. Las variables de distancias, son distancias Euclidianas calculadas en metros. Se calcularon a partir de los archivos vectoriales publicados por INEGI.

El conjunto de datos se creó para cada uno de los años analizados 2000, 2010 y 2020. Por lo tanto, se tienen las veinte variables para cada uno de los años. En la Figura 3.3 se pueden apreciar

variables del año 2000.

Tabla 3.5. Conjunto de datos generado a partir de las recolección de datos y segmentación.

Categoría	Variable
Topografía	(1) Modelo Digital de Elevaciones, (2) pendiente del terreno y (3) dirección de la pendiente
Uso de Suelo	(4) Mapas de uso de suelo, (5) Distancia a áreas construidas, (6) calles y correteras, (7) vías férreas y (8) centro de las ciudades.
Ambientales	(9) Distancia a áreas verdes/parques
Socioeconómicas	(10) Densidad de población, (11) población ocupada, (12) jefatura femenina y (13) jefatura masculina, (14) ingresos trimestrales por vivienda y (15) número de viviendas por celda.
Culturales	(16) Distancia a templos, (17) cementerios y (18) escuelas, (19) densidad de población religiosa,
Vecindario	(20) Probabilidad de cambio del pixel

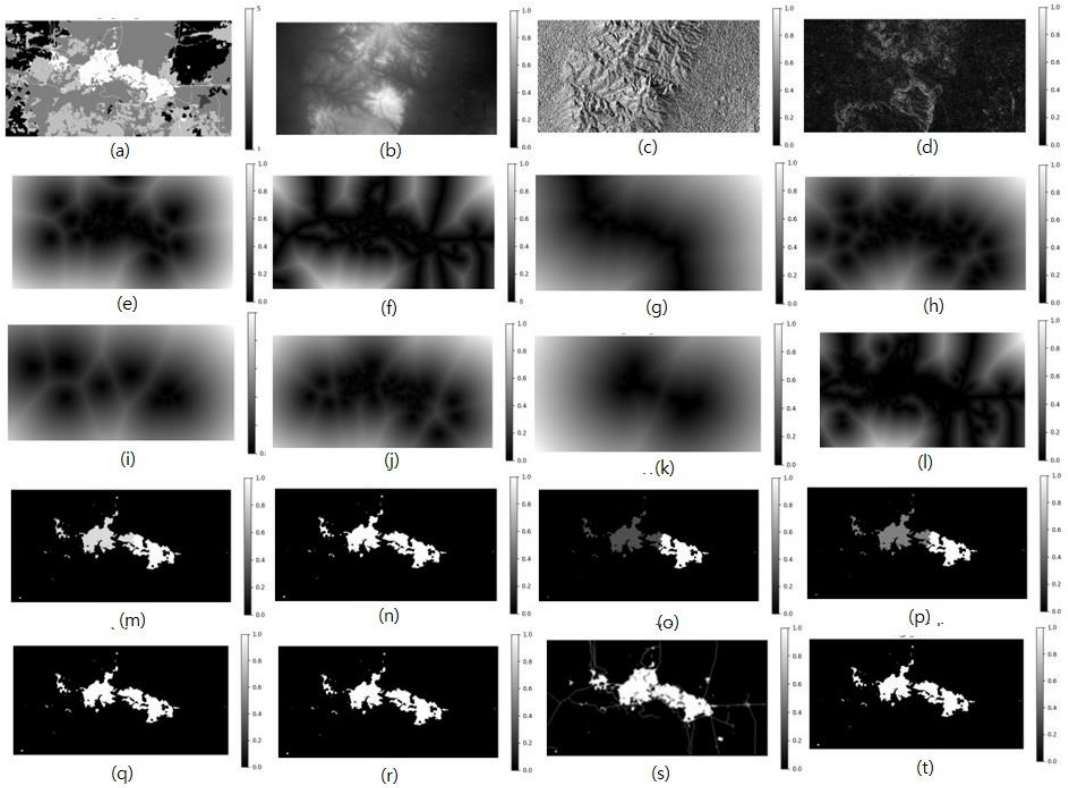

Figura 3.3. Datos de entrenamiento del año 2000 usadas en el estudio. (a) Uso de Suelo, (b) Modelo Digital de elevaciones, (c) Dirección de la pendiente, (d) Pendiente. (e)-(l) variables de distancia y (m)-(t) Variables de densidad.

Adicionalmente, al conjunto de datos se crearon las imágenes objetivas, donde se clasificó en Zona Urbana y Zona no Urbana de los años 2000 y 2020. Sin embargo, por la naturaleza de la zona bajo estudio, se observa que se tiene un desbalance en las clases, se puede ver en la Figura 3.4. Por esta razón, en la fase número dos antes de realizar los modelos predictivos se realizó un sub muestreo aleatorio donde se balancearon las clases.

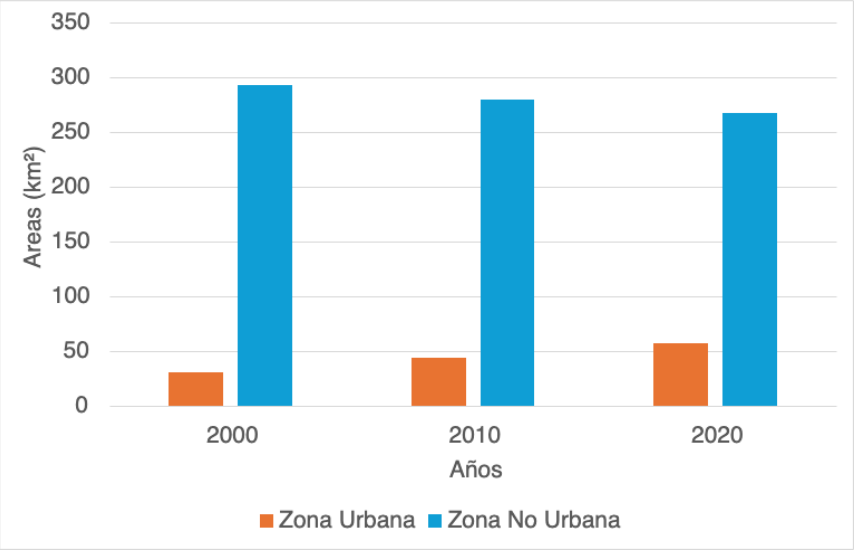

Figura 3.4. Áreas en km^2 de las Zonas Urbanas y No Urbanas.

3.2.2 Segunda Fase

Teniendo construidos los conjuntos de datos, se continuó con la segunda fase. En esta fase se separaron los datos en dependencia del año (2000, 2010 y 2020). En este estudio los datos se dividieron como se indica en la Tabla 3.6. Para el entrenamiento se utilizaron los datos del año 2000 como variables independientes y las dependientes corresponden a la clasificación binaria del 2010. Para la evaluación se seleccionaron los datos del año 2010 con variable objetivo del año 2020. Para la predicción se usaron como variables independientes los datos del año 2020. Debido a la disparidad de las clases, se realizó un sub muestreo aleatorio nivelando las clases. En la Tabla 3.7 se observa el porcentaje de píxeles correspondientes a la clase minoritaria. Para este proceso se seleccionaron ocho semillas aleatorias para realizar el sub muestro.

Tabla 3.6. División de datos en entrenamiento y evaluación.

Conjunto de datos	Variables independientes	Variable dependiente
Entrenamiento	Datos del año 2000	Clasificación binaria del 2010
Evaluación	Datos del año 2010	Clasificación binaria del 2020

Para evitar el sesgo en la predicción y errores de predicción de clase minoritaria derivados del desbalance de las clases. Se realizó un sub muestreo aleatorio nivelando las clases. La experimentación de los diferentes algoritmos de aprendizaje automático se llevó a cabo con este subconjunto. En este problema se decidió utilizar los algoritmos de Máquinas de soporte vectorial de Margen Suave, Bosques Aleatorios y Boosting categórico.

Tabla 3.7. Distribución de la variable objetivo.

	2000	**2010**	**2020**
Zona Urbana	9.5%	13.7%	17.6%
Zona no urbana	90.5%	86.3%	82.4%

La selección de los algoritmos se realizó tomando en cuenta los estudios relacionados, al igual que los valores de sus parámetros.

El modelo de SVM de margen suave se desarrolló utilizando diferentes valores de los parámetros del algoritmo. Los valores por variar en esta experimentación fueron: parámetro c, kernel, gama y grado del polinomio. Para el parámetro c se seleccionaron los valores 0.1, 1, 10, 100. Asimismo, se seleccionó el kernel radial (RBF) con el parámetro gama de 1, 2, 3. Finalmente, para el kernel polinomial se eligieron los valores de grado de polinomio de 1, 2, 3, 4 y 5. Las combinaciones se hicieron para obtener el modelo predictivo más adecuado. En la Tabla 3.8, se encuentran los diferentes valores de los parámetros.

Tabla 3.8. Parámetros de SVM se modificaron en la experimentación.

Parámetros	**Valores**
C	0.1, 1, 10, 100
Kernel	Polinomial, radial (RBF)
Gama	1, 2, 3
Grado	1, 2, 3, 4, 5

El segundo algoritmo utilizado corresponde a los bosques aleatorios (RF). El modelo se desarrolló usando diferentes parámetros del algoritmo. Los parámetros que se variaron en este trabajo fueron el número de árboles, profundidad máxima, semilla de aleatoriedad y el número máximo de características. Los diferentes valores para estos parámetros son los mostrados en la Tabla 3.9.

Tabla 3.9. Parámetros de bosques aleatorios a variar en la experimentación.

Parámetros	**Valores**
Números de árboles (n_estimators)	10, 50, 100
Profundidad máxima (max_depth)	3, 5, 10
Semilla de aleatoriedad (random_state)	0, 1, 42
Número máximo de características (max_features)	sqrt, log2, none

Por último, en la experimentación desarrollada para el algoritmo de Boosting Categórico (catBoost), se variaron los parámetros: iteraciones, tasa de aprendizaje, profundidad, iteraciones de estimación de hoja y regularización L2 de hoja. Los valores seleccionados en la experimentación se encuentran en la Tabla 3.10.

Tabla 3.10. Parámetros del boosting categórico a variar en la experimentación.

Parámetros	**Valores**
Iteraciones (iterations)	100, 500, 1000
Tasa de aprendizaje (learning_rate)	0.025, 0.05, 0.1, 0.2, 0.3
Profundidad (depth)	3, 6, 9
Iteraciones de estimación de hoja (leaf_estimation_iterations)	1, 10
Regulación L2 de hoja (l2_leaf_reg)	1, 3, 6, 9

Realizado el entrenamiento de los modelos con los diferentes parámetros, se obtuvo la predicción al año 2020 usando las veinte variables independientes del conjunto de datos del 2020. Con la predicción realizada y lo observado en la realidad del año 2020, se efectuó la evaluación usando métricas supervisadas. Las métricas de evaluación para la selección del mejor modelo fueron: exactitud (entrenamiento y prueba), F1-Score, precisión y recall. Estas métricas se calcularon usando la matriz de confusión de los píxeles construidos y no construidos. Realizado el

entrenamiento y las pruebas, se pasa a la tercera fase.

3.2.3 Tercera Fase

Usando las métricas de evaluación, en la tercera fase se seleccionó el mejor modelo de cada uno de los algoritmos. Debido al desbalance de las clases, la selección se centró en el mayor F1-Score. Al seleccionar los mejores modelos, se usó el conjunto de datos del año 2020 para realizar una predicción al año 2030.

Con el objetivo de comprender y analizar la predicción, se usaron los métodos de inteligencia artificial explicable; en este trabajo se utilizaron los valores SHAP. En este estudio se utilizó el 5 % de los datos. Esto se decidió debido a las exigencias computaciones de la técnica. El cálculo de estos valores nos ayudará a comprender como los algoritmos llegaron a sus resultados. Ayudando a mejorar la comprensión y análisis de los resultados.

Finalmente, se llevó a cabo un análisis comparativo de los resultados de la predicción, valores SHAP y se emitió una conclusión, además de los mapas de escenario de cada uno de los algoritmos.

Cápitulo 4. Resultados

4.1 Segmentación

La segmentación se llevó a cabo usando el algoritmo FRFCM, usando los parámetros encontrados en la Tabla 3.4. A partir de los resultados de la segmentación se obtuvieron imágenes binarias de las zonas urbanas encontradas en la Figura 4.1. Figura 4.1a es la correspondiente al 2000, Figura 4.1b del 2010 y la Figura 4.1c del 2020. En estas se puede observar el aumento de los pixeles urbanizados.

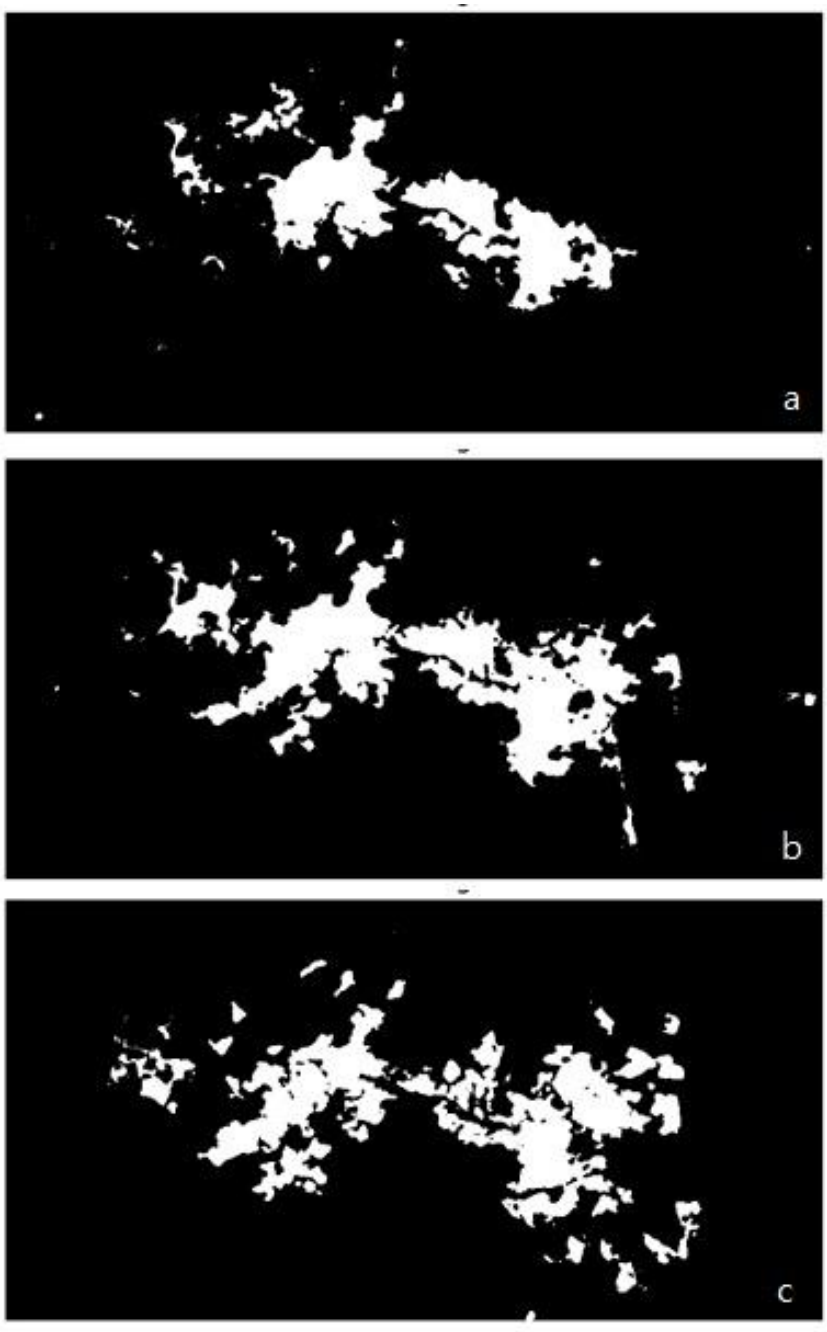

Figura 4.1. Imágenes binarias aisladas de la segmentación por año, a) 2000, b) 2010 y c) 2020.

La evaluación de estas imágenes binarias se llevó a cabo con métricas de evaluación supervisadas. Se utilizo el Índice de Jaccard, coeficiente Cohen Kappa y el MCC. Para lograrlo se crearon imágenes de referencia en el software QGIS. Las imágenes de referencia se encuentran en la Figura 4.2. En Figura 4.2a es la imagen de validación del año 2000, Figura 4.2b del año 2010 y del Figura 4.2c el año 2020.

En la Tabla 4.1 se encuentran los resultados de las métricas de evaluación, utilizando las ecuaciones (2.16), (2.17) y (2.18). En las tres métricas de evaluación mientras más cercano a uno la segmentación se realizó de buena manera. Los resultados obtenidos están alrededor del 0.7, lo cual consideramos un buen rendimiento. Con estos resultados se generaron los mapas de uso de suelo.

Tabla 4.1. Tabla de índices resultantes de la evaluación supervisada.

Año	Jaccard	MCC	kappa de Cohen
2000	0.77678	0.72834	0.72454
2010	0.78566	0.74608	0.74275
2020	0.78247	0.74304	0.73887

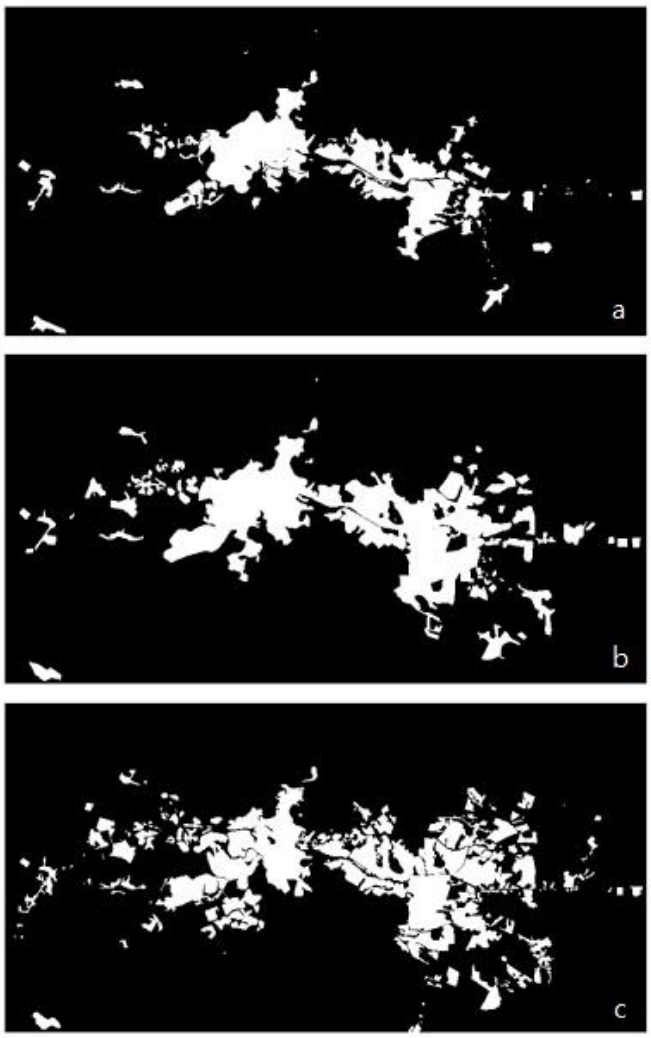

Figura 4.2. Imágenes de referencia que se usaron para la validación supervisada de la segmentación. a) 2000, b) 2010 y c) 2020

A los resultados de la segmentación se incorporaron datos vectoriales de vías de comunicación, cementerios, áreas verdes y recreativas a la categoría de zona urbana. Para crear un mapa de uso de suelo preciso. Los mapas de uso de suelo y cobertura vegetal se crearon etiquetando los resultados de la segmentación en diferentes categorías. Estos mapas nos pueden ser útiles para identificar y diferenciar las áreas según su función y características. Se usan en la planificación urbana y rural, la gestión de los recursos naturales y para entender la dinámica del paisaje en el tiempo. En este trabajo se utilizaron para identificar cómo el suelo se modificó para ser usado como zona urbana.

Los mapas se crearon para cada año, 2000, 2010 y 2020. Se utilizó la herramienta de software libre QGIS, los mapas están en la proyección UTM zona 13 N (Figura 4.3). Las etiquetas seleccionadas fueron: agricultura, cuerpos de agua, suelo desnudo, vegetación y zona urbana. En la Tabla 4.2 se encuentran las descripciones de estas categorías. Teniendo los mapas de cada año se procedió a realizar el análisis de detección de cambio de uso de suelo usando Python y la técnica de tabulación cruzada.

Tabla 4.2. Esquema de clasificación de las diferentes clases usados en este estudio.

Numero	Clase de uso	Descripción
1	Agricultura	Tierra de cultivo sembrada
2	Cuerpos de Agua	Lagos, ríos, entre otros.
3	Vegetación	Matorrales y pastizales
4	Suelo Desnudo	Áreas agrícolas después de cosechar, terrenos erosionados, superficie sin vegetación
5	Zona Urbana	Áreas residenciales, industrial, comercial, redes de transporte, parques y áreas recreativas.

4.1.1 Detección de cambio de uso de suelo

La detección de cambio de uso, suelo y cobertura terrestre se llevó a cabo calculando las áreas en kilómetros cuadrados de cada clase de 2000 al 2020. En la Figura 4.4 se puede observar las áreas en los años del estudio, en este se puede observar que la categoría de agua y vegetación no tienen muchos cambios. Por otro lado, la agricultura disminuyó en el año 2010. Esto puede ser consecuencia de que las imágenes se tomaron en temporadas en las cuales esas parcelas no se sembraron. Esta disminución se mira reflejada en el aumento del suelo desnudo en el año 2010. Mientras que en el 2020 aumentó la agricultura de nuevo y el suelo desnudo disminuyó. Finalmente, la categoría de zona urbana experimentó un aumento continuo.

Complementando la Figura 4.4, se presenta la Tabla 4.3, que muestra el cálculo de la Tasa de Cambio Dinámico. Esta tasa mide la velocidad o intensidad del cambio en el uso del suelo a lo largo del tiempo, y se expresa en un valor anual para los periodos 2000-2010 y 2010-2020.

La superficie destinada a la agricultura disminuyó en promedio un 1.82 % anual en el periodo de 2000 al 2010, reflejando una reducción significativa en esta categoría. Sin embargo, entre 2010 y 2020, se observó un crecimiento promedio anual del 2.05 %, lo que indica una recuperación o expansión de las tierras agrícolas. Durante el mismo periodo, el área cubierta por agua aumentó a un ritmo promedio del 3.79 % anual entre 2000 y 2010, señalando una expansión considerable en las superficies acuáticas. El crecimiento del cuerpo de agua continuó aumentando entre 2010 y 2020, la tasa de aumento fue ligeramente menor, con un 2.37 % anual, lo que sugiere una estabilización relativa.

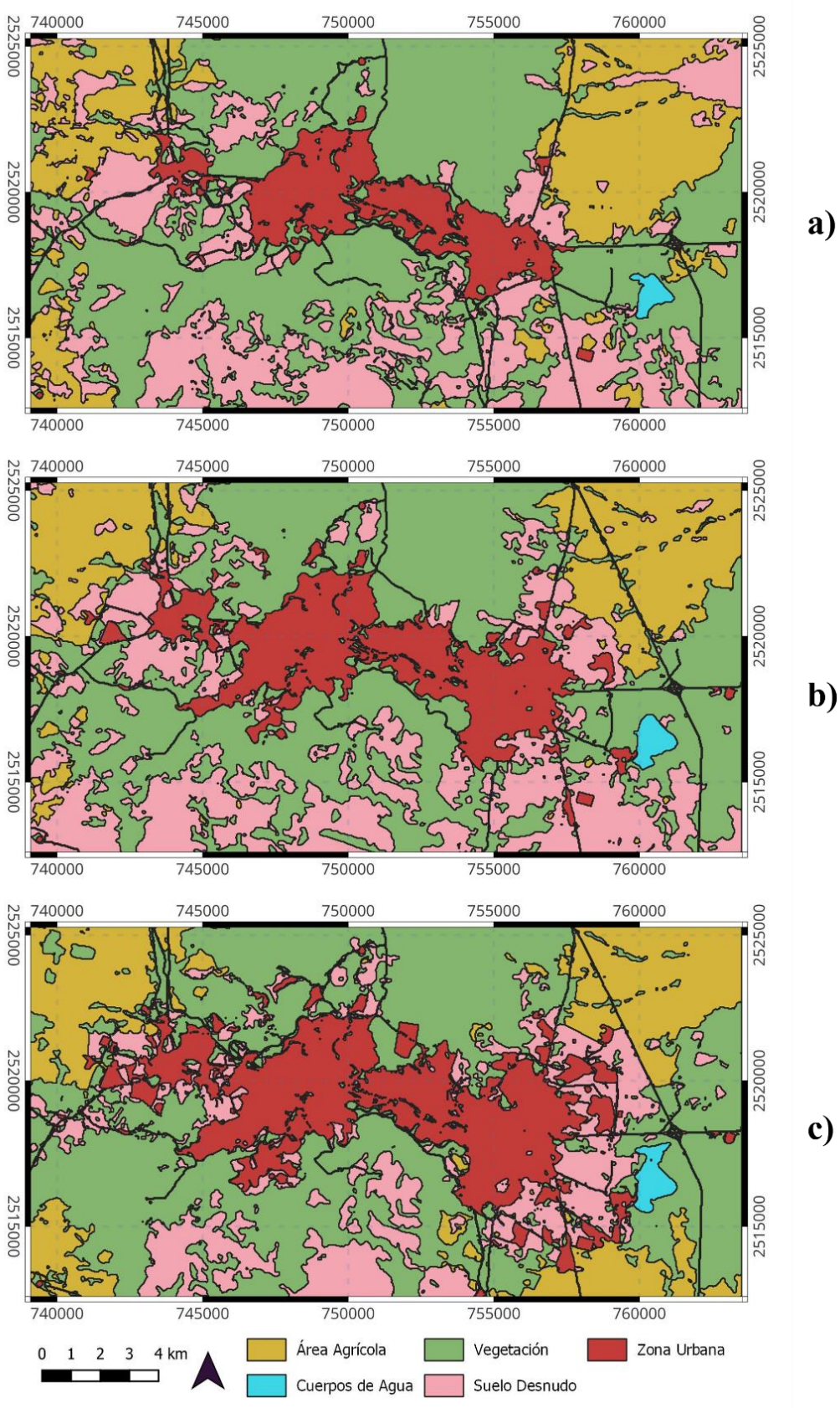

Figura 4.3. Mapas de uso de suelo realizados a partir de la segmentación. a) 2000, b) 2010 y c) 2020

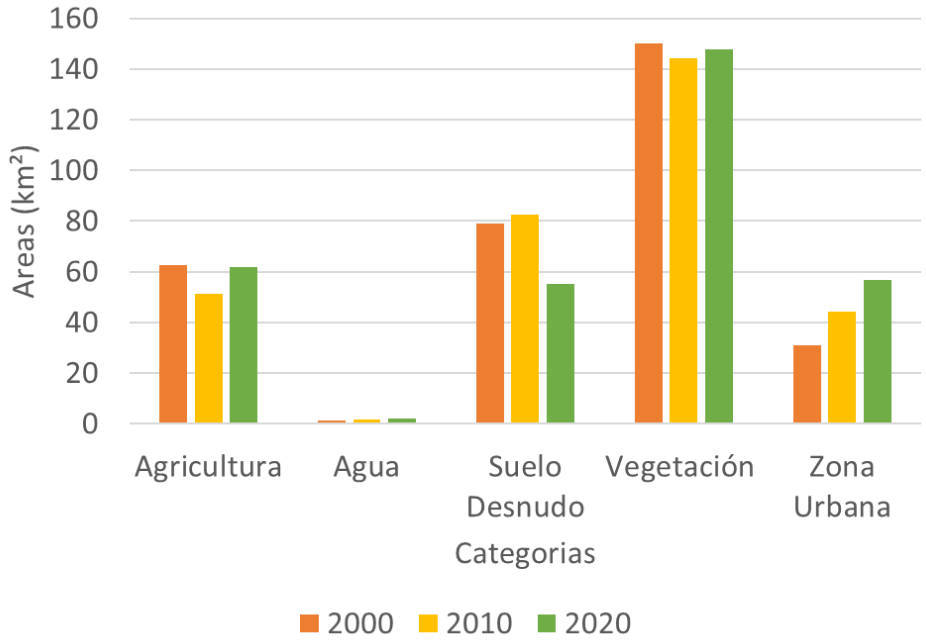

Figura 4.4. Área en km^2 por categoría en los mapas de uso de suelos.

Tabla 4.3. Áreas de usos de suelo y cobertura terrestre y la tasa de cambio dinámico (%) del 2000 al 2020.

Clases	**Áreas en km^2**			**Tasa de Cambio Dinámico (%)**	
	2000	**2010**	**2020**	**(2000-2010)**	**(2010-2020)**
Agricultura	62.61	51.18	61.69	-1.8	2.1
Cuerpos de Agua	1.17	1.62	2	3.8	2.4
Vegetación	149.95	144.07	147.75	-0.4	0.3
Suelo Desnudo	79.05	82.31	55.22	0.4	-3.3
Zona Urbana	30.78	44.38	56.89	4.4	2.8

En cuanto al suelo desnudo, entre 2000 y 2010 se experimentó un leve incremento del 0.41 % anual en promedio, reflejando una ligera expansión de esta categoría. Sin embargo, en la siguiente década, de 2010 a 2020, se produjo una disminución relevante del 3.29 % anual, lo que podría indicar una posible revegetación o cambio en el uso del suelo. La superficie vegetada, por su parte, experimentó una disminución moderada del 0.39 % anual entre 2000 y 2010, lo que alude una pérdida de vegetación. Esta tendencia se revirtió entre 2010 y 2020, con un leve aumento del 0.26 % anual en promedio, indicando una recuperación parcial de la vegetación.

Por último, las áreas urbanas crecieron significativamente, con una tasa promedio anual del 4.42 % entre 2000 y 2010, lo que refleja una rápida expansión urbana durante ese periodo. Aunque el crecimiento continuó en la siguiente década, lo hizo a un ritmo más lento, con un aumento del 2.82 % anual en promedio entre 2010 y 2020, lo que podría indicar una desaceleración en la expansión urbana.

Los cambios en el uso del suelo en las diferentes categorías estudiadas se analizaron, elaborando tablas de tabulación cruzada que permiten comparar y visualizar cómo las áreas de distintas clases de uso del suelo han evolucionado entre los periodos considerados. Estas tablas proporcionan una representación detallada de la transición entre las categorías, permitiendo identificar patrones de cambio, como la conversión de áreas agrícolas en zonas urbanas, la expansión de superficies

vegetadas, o la reducción de suelos desnudos.

Los resultados de la tabulación cruzada se encuentran en las Tabla 4.4 y Tabla 4.5. En estas matrices la diagonal principal representa las áreas que se conservaron en las mismas categorías, mientras que las otras secciones nos indica que categoría se trasladaron o cambiaron.

La Tabla 4.4 ilustra el cambio en el uso del suelo entre 2000 y 2010. Se observa que la mayoría de las áreas permanecieron en la misma categoría, lo que sugiere una cierta estabilidad en el uso del suelo. Sin embargo, también se evidencian transiciones significativas entre categorías.

El área correspondiente a 15.35 kilómetros cuadrados de tierras agrícolas pasaron a ser suelo desnudo, mientras que 5.96 kilómetros cuadrados de áreas agrícolas se transformaron en vegetación y 1.24 kilómetros cuadrados se convirtieron en zona urbana. En la categoría de suelo desnudo, 14.68 kilómetros cuadrados identificados en 2000 cambiaron a vegetación, y 7.46 kilómetros cuadrados se urbanizaron.

En cuanto a la vegetación, 6.03 kilómetros cuadrados se urbanizaron, mientras que 18.77 kilómetros cuadrados se transformaron en suelo desnudo. Adicionalmente, las transiciones más notables incluyen los 18.77 kilómetros cuadrados de vegetación que pasaron a ser suelo desnudo, 15.35 kilómetros cuadrados de terrenos agrícolas que se convirtieron en suelo desnudo, 14.68 kilómetros cuadrados de suelo desnudo que se transformaron en vegetación, y 7.46 kilómetros cuadrados de suelo desnudo que se urbanizaron.

Tabla 4.4. Matriz de cambio de uso de suelo y cobertura (superficie en km^2) de 2000 a 2010

2010 / **2000**	**Agricultura**	**Cuerpos de Agua**	**Suelo Desnudo**	**Vegetación**	**Zona Urbana**
Agricultura	40.07	0	15.35	5.96	1.24
Cuerpos de Agua	0	1.10	0	0.07	0
Suelo Desnudo	0	0	48.06	14.68	7.46
Vegetación	2.26	0.51	18.77	122.36	6.03
Zona Urbana	0.02	0	0.13	0.98	29.64

Tabla 4.5. Matriz de cambio de uso de suelo y cobertura (superficie en km^2) de 2010 a 2020

2020 / **2010**	**Agricultura**	**Cuerpos de Agua**	**Suelo Desnudo**	**Vegetación**	**Zona Urbana**
Agricultura	39.65	0	4.18	6.93	0.43
Cuerpos de Agua	0	1.55	0.001	0.07	0
Suelo Desnudo	18.71	0.00	32.96	21.98	8.65
Vegetación	3.27	0.45	17.86	118.32	4.17
Zona Urbana	0.06	0	0.23	0.44	43.64

Aunque hubo un aumento en la urbanización, el cambio en esta categoría no fue tan significativo en comparación con las transformaciones observadas entre otras categorías de uso del suelo, como la conversión de vegetación y tierras agrícolas a suelo desnudo.

En el segundo periodo, de 2010 a 2020, ilustrado en la Tabla 4.5, se observa un comportamiento similar al del periodo anterior, con las mayores áreas concentradas en la diagonal principal. Dentro de las áreas agrícolas, 4.18 kilómetros cuadrados pasaron a ser suelo desnudo, 6.93 kilómetros cuadrados se transformaron en vegetación, y solo 0.43 kilómetros cuadrados se convirtieron en

zona urbana.

En la categoría de suelo desnudo, 18.71 kilómetros cuadrados se convirtieron en terrenos agrícolas, 21.98 kilómetros cuadrados en vegetación, y 8.65 kilómetros cuadrados se urbanizaron.

En cuanto a las áreas de vegetación en la región, 3.27 kilómetros cuadrados se transformaron en terrenos agrícolas, 17.86 kilómetros cuadrados pasaron a ser suelo desnudo, y 4.17 kilómetros cuadrados se convirtieron en zona urbana.

Similar al periodo anterior, la transición de suelo desnudo a terrenos agrícolas (18.71) y vegetación a suelo desnudo (17.86), son mayores a la transformación a Zona Urbana, estando la mayor de suelo desnudo a Zona urbana, con unos 8.65 kilómetros cuadrados. Analizadas las matrices de cambio se realizó un mapa para cada uno de los periodos, estos mapas se encuentran en Figura 4.5.

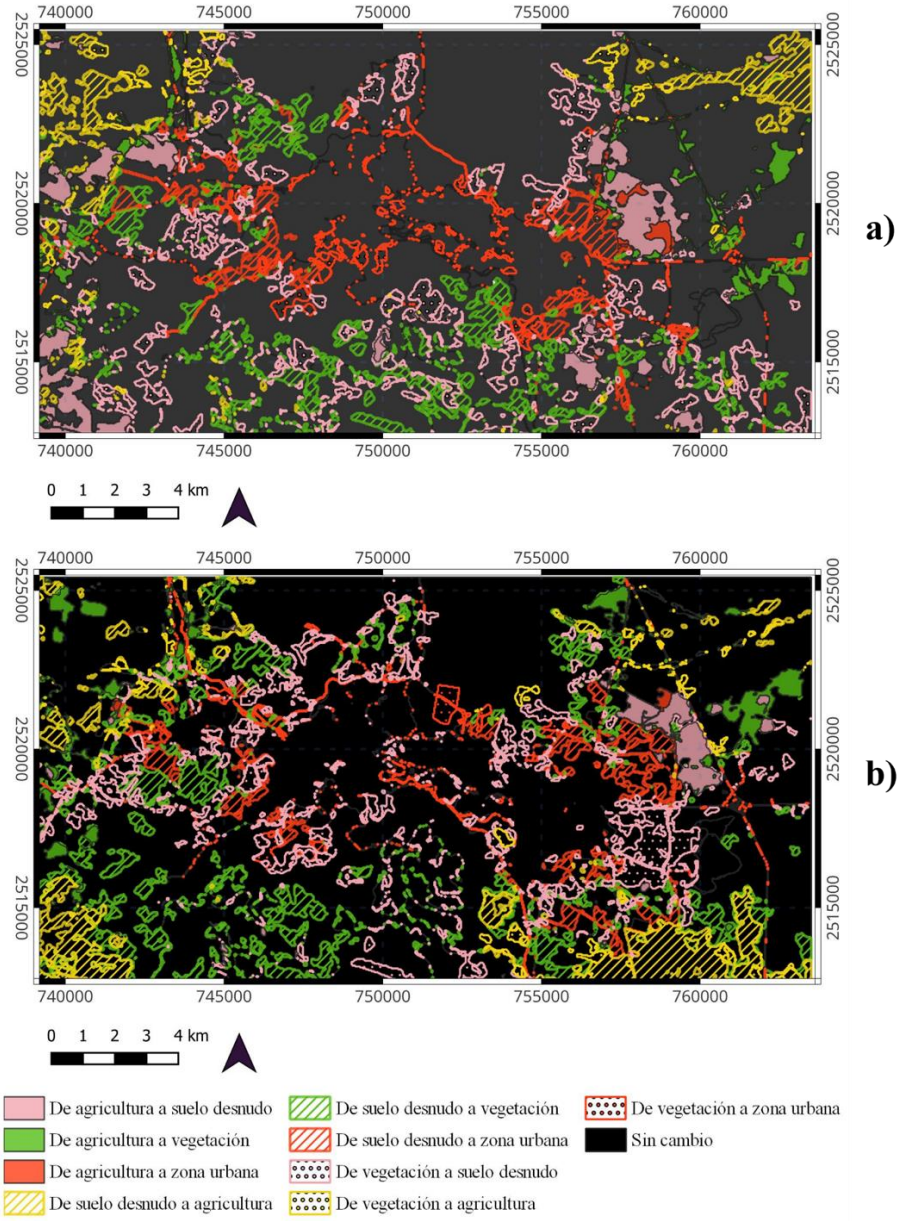

Figura 4.5. Mapas de cambio de uso de suelo de la zona de estudio, a) periodo 2000 a 2010 y b) periodo de 2010 a 2020.

El análisis de ambos periodos muestra que la transformación de áreas hacia zonas urbanas no es significativa en comparación con otros cambios de uso del suelo. Este comportamiento podría indicar un crecimiento urbano moderado en la región, reflejando una estabilidad en los usos del suelo o una mayor prioridad en la transformación hacia otros tipos de cobertura, como vegetación o suelo desnudo. Además, la tasa de urbanización podría sugerir la presencia de barreras geográficas y económicas que dificultan un desarrollo urbano más extenso.

Para comprender mejor cómo se distribuye y evoluciona la expansión urbana dentro de la región, se decidió dividir el área de estudio en cuatro partes: noroeste (NW), noreste (NE), suroeste (SW)

y sureste (SE). Esta división, ilustrada en la Figura 4.6, permitiendo un análisis más focalizado de la expansión urbana de la ciudad en diferentes sectores. Por lo tanto, adicional a la sectorización de nuestra imagen, se cambiaron las clases a Zona Urbana y Zona no urbana. Estando en esta última las categorías de zonas agrícolas, cuerpos de agua, suelo desnudo y vegetación.

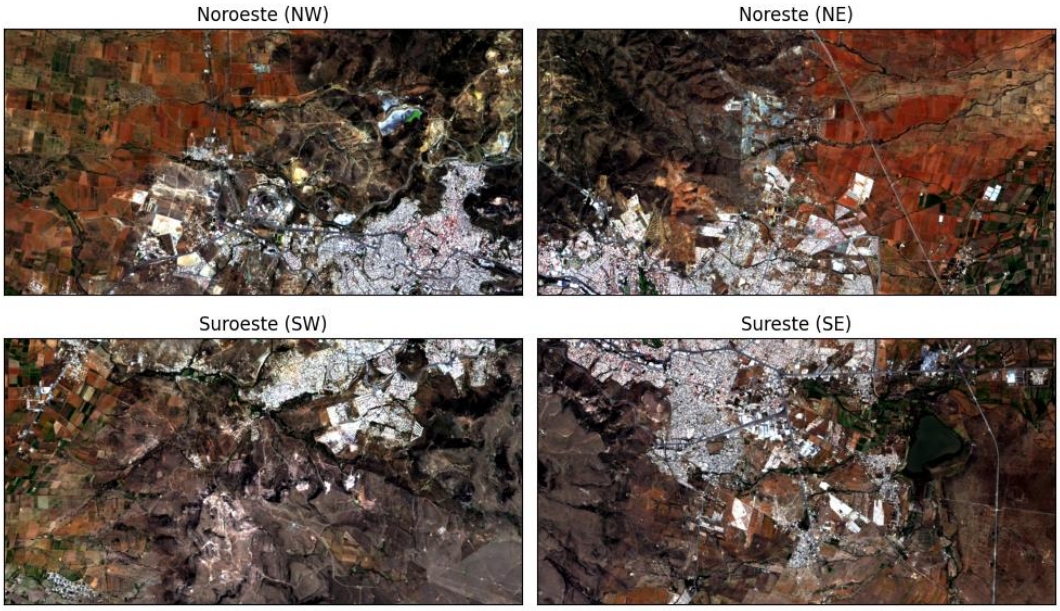

Figura 4.6. División de la zona de estudio en cuatro partes (NW, NE, SW y SE).

Realizada la división se creó la Figura 4.7. La Figura 4.7a nos muestra el área en kilómetros cuadrados de las clases en los diferentes sectores, se puede observar el incremento de las áreas de las zonas urbanas y las no urbanas a lo largo de los años analizados. Se puede observar que históricamente se tiene mayor área construida en la zona noroeste (NW), seguida de la sureste (SE). La zona NW en su mayoría, corresponde a la zona del municipio de Zacatecas, mientras que la SE corresponde al municipio de Guadalupe. Por otro lado, la Figura 4.7b se tiene la diferencia de áreas en las zonas entre 2000 a 2010 y de 2010 a 2020, para conocer qué sector tuvo más expansión en los periodos analizados. En el periodo de 2000 a 2010, el crecimiento mayor se concentra en la zona SE seguido de la SW. En el otro periodo de 2010 a 2020, el crecimiento se encuentra en la zona NE seguida del SE.

En la Tabla 4.6 se expresan la cantidad de pixeles urbanos y no urbanos encontrados en la imagen binaria resultante de la unión de la segmentación y lo datos vectoriales. Teniendo esta tabla se realizó el cálculo de la tasa de cambio global de los píxeles entre los años estudiados. Estos resultados se expresan en la Tabla 4.7. La tasa de cambio se mantuvo constante en los periodos analizados.

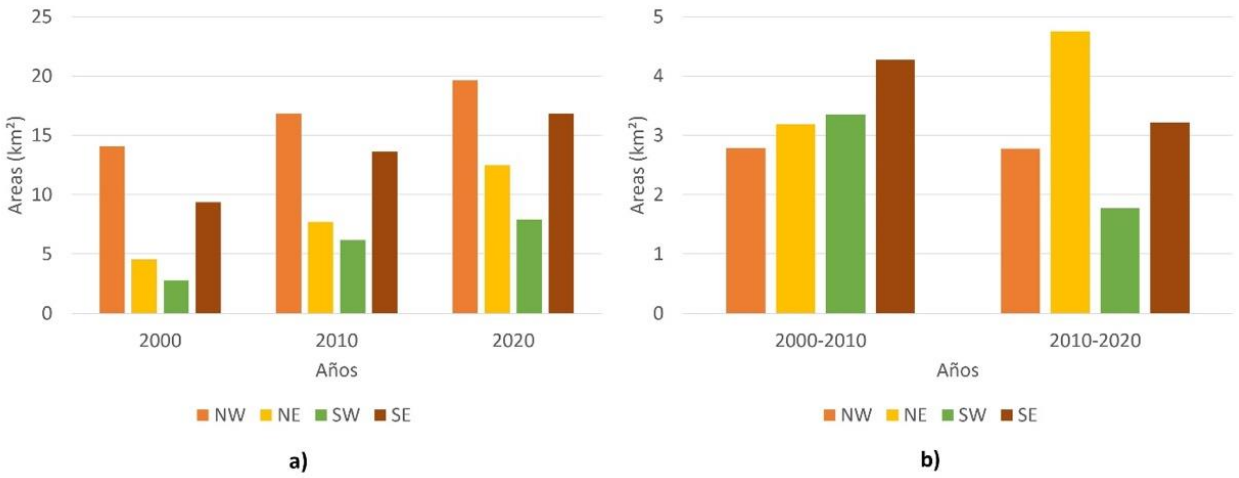

Figura 4.7. a) Áreas en km^2 de las zonas construidas en las divisiones NW, NE, SW y SE. b) Diferencias de las zonas construidas en los periodos 2000 a 2010 y 2010 a 2020 en km^2 de las divisiones NW, NE, SW y SE.

Tabla 4.6. Numero de pixeles construidos y no construidos en el área de la zona metropolitana Zacatecas-Guadalupe en los años 2000, 2010 y 2020.

	2000	**2010**	**2020**
Píxeles urbanos	34195	49306	63067
Píxeles no urbanos	325307	310196	296435

Tabla 4.7. Tasa de crecimiento global de la zona de estudio en los periodos 2000 a 2010 y 2010 a 2020.

	(2000-2010)	**(2010-2020)**
Cantidad de píxeles que cambiaron	16369	14578
Cantidad de píxeles que no cambiaron	343133	344924
Tasa de cambio	4.55%	4.06%

Teniendo el área de estudio dividida y para robustecer nuestro análisis, se calculó el Índice de Intensidad de Expansión Urbana (IIEU). En la Tabla 4.8 se aprecian los resultados del IIEU en el primer periodo analizado de 2000 a 2010. De manera general, el desarrollo de la zona urbana se encuentra en una velocidad baja, ya que se encuentra el tango de 0.29 a 0.59. En este periodo, el sector con el índice más alto es el SE.

Tabla 4.8. Valores del IIEU en el periodo 2000 a 2010.

Sector	**IIEU**	**Descripción**
NW	0.35	Desarrollo de velocidad baja
NE	0.39	Desarrollo de velocidad baja
SW	0.41	Desarrollo de velocidad baja
SE	0.53	Desarrollo de velocidad baja

En el segundo periodo, de 2010 a 2020, encontrado en la Tabla 4.9. Podemos observar que tres de los cuatro cuadrantes continúan con una velocidad de desarrollo bajo (NW, SW y SE). Mientras que el sector NE tuvo una velocidad de desarrollo media, debido a que este ya tiene un valor de 0.59. Es importante establecer que estos resultados concuerdan con los observados en la Figura

4.7b.

Tabla 4.9. Valores del IIEU en el periodo 2010 a 2020.

Sector	IIEU	Descripción
NW	0.34	Desarrollo de velocidad baja
NE	0.59	Desarrollo de velocidad media
SW	0.22	Desarrollo de velocidad baja
SE	0.40	Desarrollo de velocidad baja

4.2 Modelos predictivos

Una vez efectuado el análisis del cambio de uso de suelo, es crucial considerar cómo estas variaciones se reflejan en los modelos de predicción. Para considerar las variaciones se realizaron la Figura 3.4, con las categorías Zona Urbana y Zona no Urbana en la cual se observa un desbalance en las clases. En la Tabla 3.7 se observa el porcentaje de pixeles correspondientes a la clase minoritaria.

Para evitar el sesgo en la predicción y errores de predicción de clase minoritaria derivado del desbalance de las clases. Se realizo un submuestreo aleatorio nivelando las clases. La experimentación de los diferentes algoritmos de aprendizaje automático se llevó a cabo con este subconjunto. En este problema se decidió utilizar los algoritmos de Maquinas de Soporte Vectorial de Margen Suave, Bosques Aleatorios y Boosting categórico.

4.2.1 Máquinas de Soporte Vectorial de Margen Suave

La experimentación de las SVM de margen suave se llevó a cabo variando los parámetros del algoritmo. Los valores a los diferentes parámetros se encuentran en la Tabla 3.8. Después de realizar el entrenamiento, la fase de pruebas se realizó con el conjunto de datos del 2020. Esta se llevó a cabo usando las métricas de exactitud (entrenamiento y validación), F1-Score y precisión. Estas métricas se calcularon usando la matriz de confusión de los píxeles construidos y no construidos.

Los resultados obtenidos de todos los modelos tienen una exactitud de entrenamiento y prueba mayores al 90 %. En la Figura 4.8 se pueden apreciar los resultados. En la exactitud de entrenamiento, los valores más altos son los modelos con kernel RBF. Por otro lado, en la exactitud de validación, los mejores valores se encuentran en los modelos con kernel polinomial con grado de polinomio de cinco. Sin embargo, debido al desbalance de las clases, se decidió apoyarse del F1-Score.

Los resultados más elevados en el F1-Score se encuentran en el kernel polinomial y grado del polinomio cinco, similar a la exactitud de validación. Estos resultados están alrededor del 80 %. Los resultados en la precisión son similares al F1-Score, los mejores se encuentran en el kernel polinomial en los valores del 80.

Para los valores de c los mejores resultados están en los valores de 100 y 0.1. En la exactitud presentada en el entrenamiento y precisión, los mejores valores están en el valor c de 100. Mientras que la exactitud de validación y F1-Score se encuentran en el valor de 0.1.

El modelo que se eligió como el mejor, se debe a que tienen uno de los mejores F1-Score y un

balance con las demás métricas. Este modelo es de kernel polinomial, valor c de 0.1 y un grado de polinomio de cinco, la exactitud de entrenamiento es de un 92.4%, exactitud de validación 93% y F1-Score de un 86.3%. En la Figura 4.8 en el recuadro rojo se encuentra el modelo seleccionado.

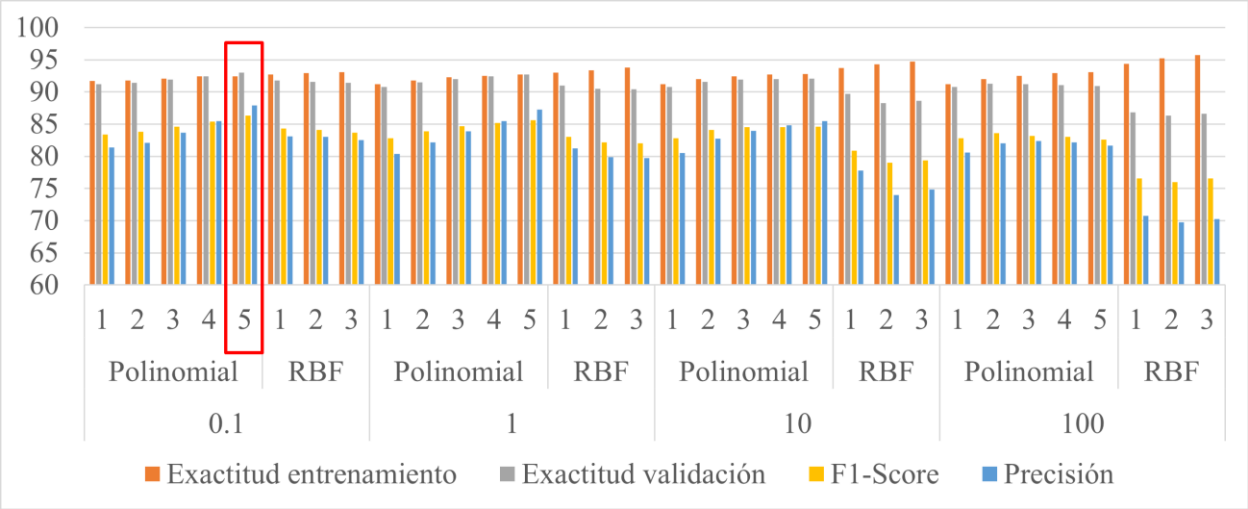

Figura 4.8. Resultado de la experimentación de SVM, en el cuadro rojo se encuentra el que se considera el mejor modelo.

En la Figura 4.9 se encuentra la comparación del resultado del modelo respecto a los pixeles construidos y no construidos en el año 2020 con las reales.

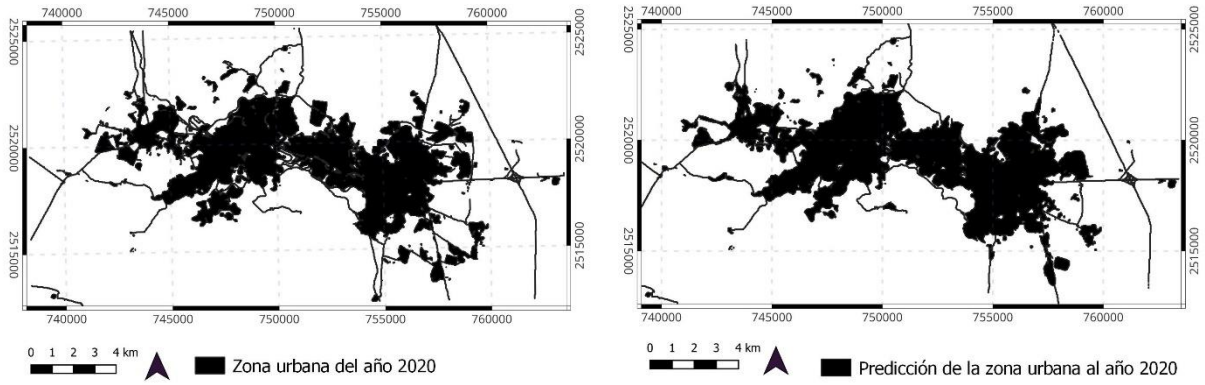

Figura 4.9. Comparación de la predicción realizado por el modelo SVM y la zona urbana real para el año 2020.

4.2.2 Bosques Aleatorios

El modelo desarrollado usando bosques aleatorios (RF) se desarrollaron usando diferentes parámetros del algoritmo. Los parámetros que se variaron en este trabajo se muestran en la Tabla 3.9. Las métricas de evaluación para la selección del mejor modelo fueron: exactitud (entrenamiento y validación), F1-Score y precisión. Estas métricas se calcularon usando la matriz de confusión de los píxeles construidos y no construidos.

Para poder visualizar de forma sencilla se realizaron dos graficas Figura 4.10. La Figura 4.10a se encuentran los resultados de los modelos generados con el número máximo de características de log2. En la Figura 4.10b son los modelos con el número máximo de características de none. Por

último, la Figura 4.10c contiene los modelos con el número máximo de características de sqrt.

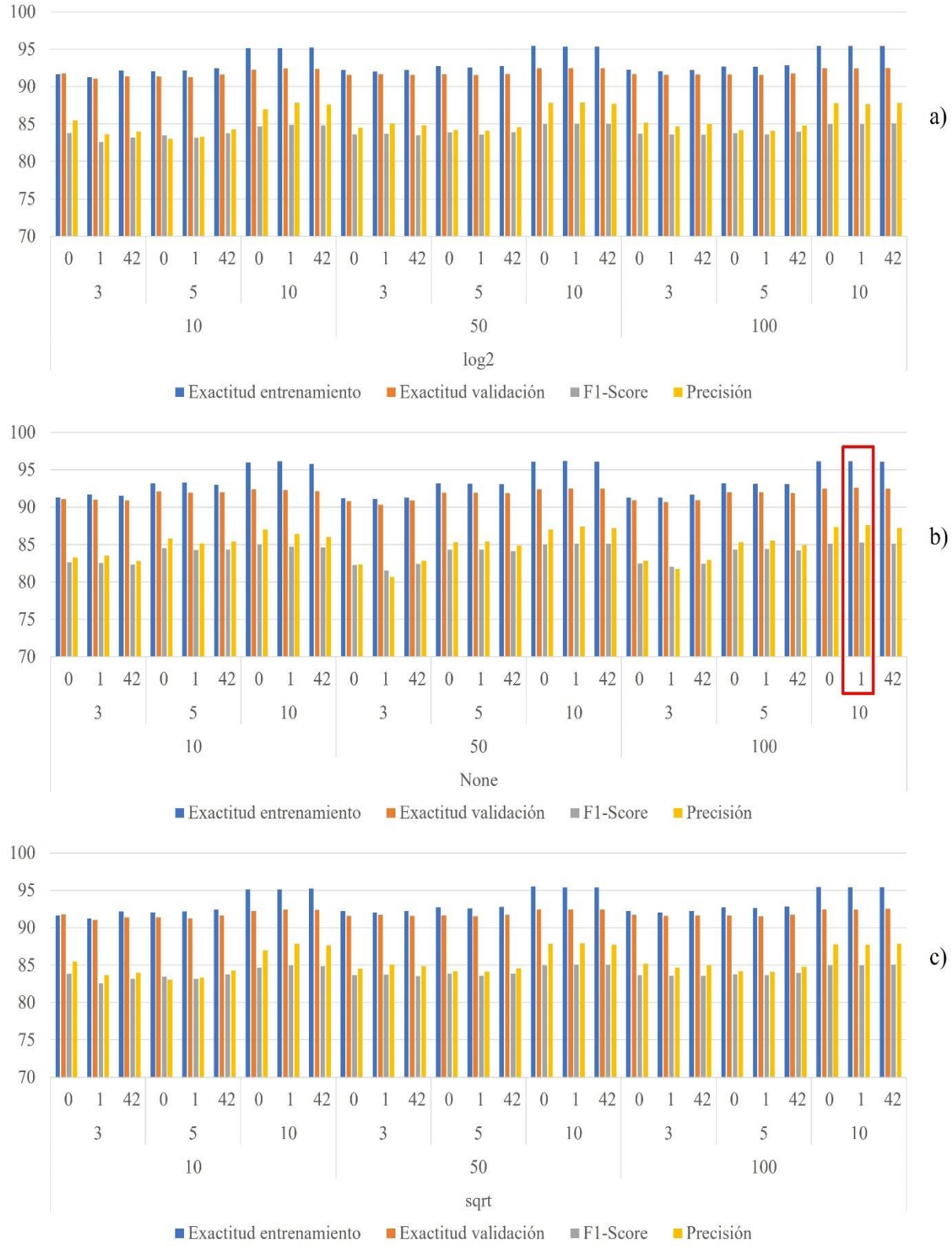

Figura 4.10. Resultados de la evaluación de los modelos de bosques aleatorios (RF) en el rectángulo rojo se encuentra señalado el modelo que se considera el mejor. a) resultados número máximo de características log2, b) none y c) sqrt

La exactitud de entrenamiento y validación en todos los modelos se encuentran arriba del 90%. Los resultados de la precisión los valores altos son alrededor del 80%, estos se encuentran en los valores de numero de árboles de 50 y 100, en la profundidad máxima de 10, en las tres semillas de aleatoriedad y en los tres números máximos de características seleccionados. Por otro lado, el F1-Score se concentra en valores entre 70% y 85%, distribuido en la mayoría de los modelos generados.

El modelo que se consideró el de mejor desempeño es el que se encuentra encerrado en la Figura 4.10b, este se seleccionó por tener el valor más alto en el F1-Score. Los parámetros de este modelo son el número de árboles con 100, profundidad máxima de 10, semilla de aleatoriedad de 1 y el número máximo de características con valor none. En las métricas se tiene un F1-Score de 85.3%, precisión de 87.6%, exactitud de entrenamiento de 96% y exactitud de validación de 92.6%. En la Figura 4.11 se encuentra la comparación del resultado del modelo respecto a los pixeles construidos y no construidos en el año 2020 con las reales.

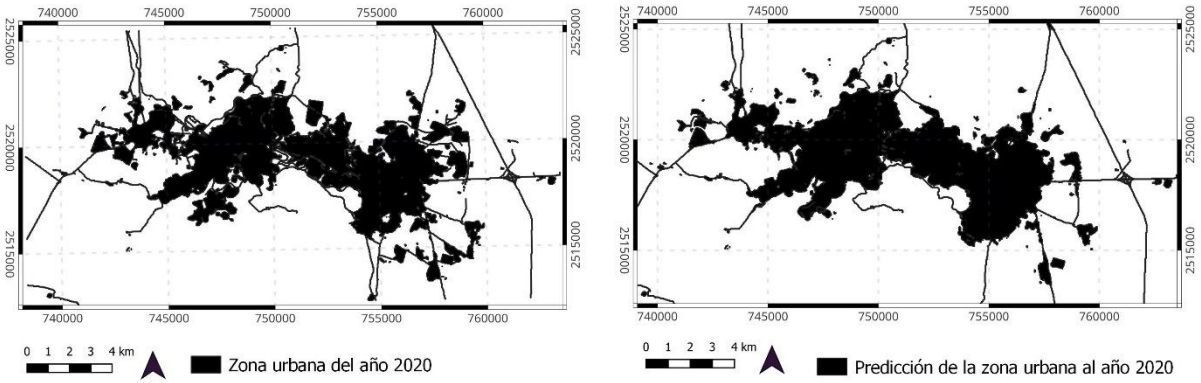

Figura 4.11. Comparación de la predicción realizado por el modelo RF y la zona urbana real para el año 2020.

4.2.3 Boosting Categórico

En la experimentación desarrollada para el algoritmo de Boosting Categórico (catBoost), se variaron los parámetros: iteraciones, tasa de aprendizaje, profundidad, iteraciones de estimación de hoja y regularización L2 de hoja. Los valores seleccionados en la experimentación se encuentran en la Tabla 3.10.

Las métricas de evaluación para la selección del mejor modelo fueron: exactitud (entrenamiento y validación), F1-Score y precisión. Estas métricas se calcularon usando la matriz de confusión de los píxeles construidos y no construidos.

En el caso de este algoritmo se agruparon los resultados por las iteraciones (100, 500 y 1000), debido al volumen de las mismas. En la Figura 4.12 se muestran los resultados de las iteraciones 500, la cual se consideraron con un mejor desempeño. La Figura 4.12a corresponde a la profundidad 3, Figura 4.12b profundidad 6 y por último Figura 4.12c la profundidad 9.

La exactitud de entrenamiento y validación se encuentran mayores a 90%. En las profundidades de 6 y 9 es donde se encuentran los valores más altos de exactitud de entrenamiento dentro del 97% y 98%. Mientras que en la exactitud de en validación se mantuvo estable en todas las profundades en un 94%. Por otro lado, los valores del F1-Score se encuentran mayor al 80%, los

valores más altos se encuentran en la profundidad 9. Por último, la precisión se encuentra dentro del 70% y 80%. Los valores más altos se encuentran en la profundidad de 9.

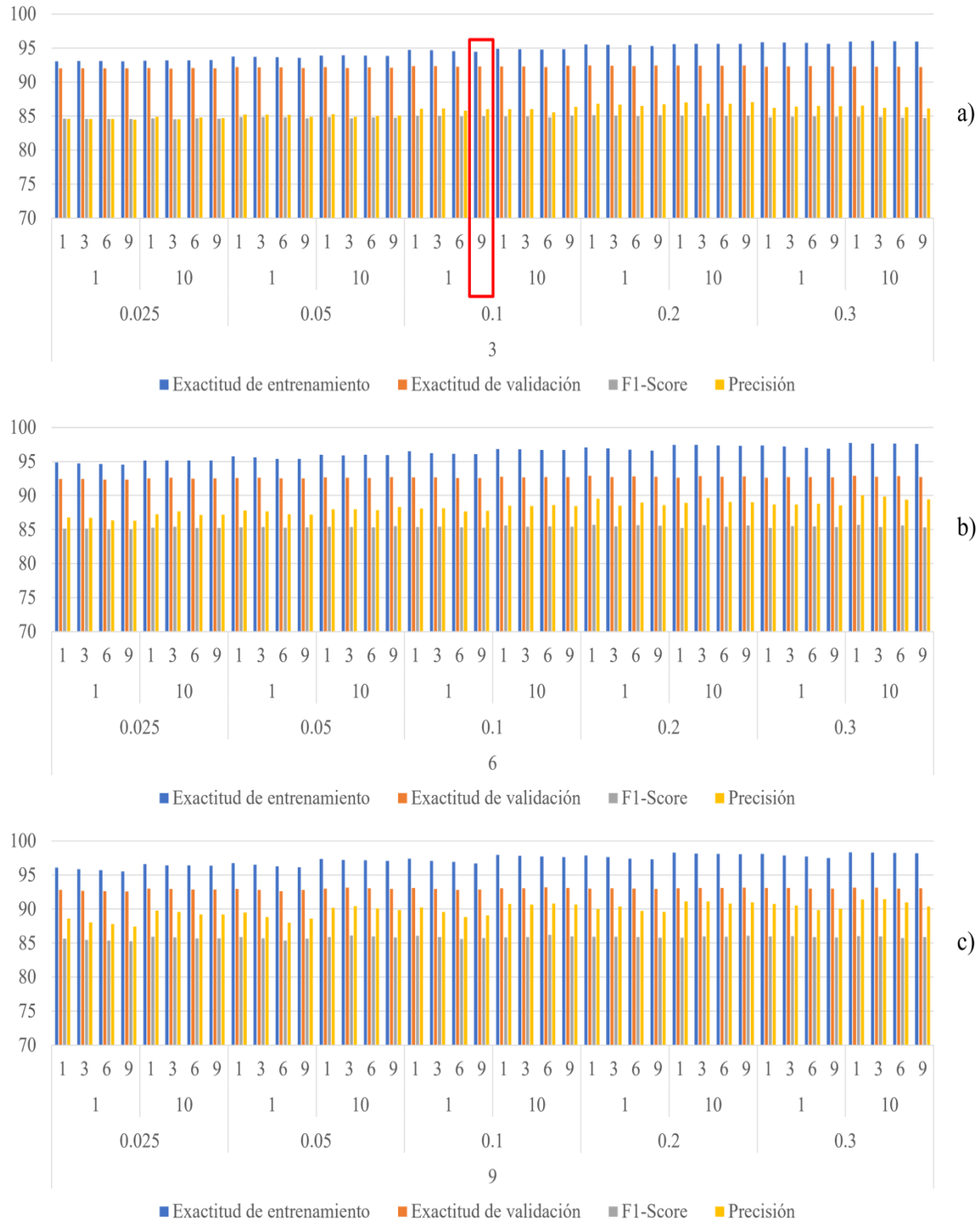

Figura 4.12. Resultados de la evaluación de los modelos de catBoost para la iteración 500, en el rectángulo rojo se encuentra señalado el modelo que se considera el mejor. a) resultados profundidad de 3, b) 6 y c) 9

El modelo que se consideró el de mejor desempeño es el que se encuentra encerrado en la Figura

4.12a dentro del recuadro rojo, este se seleccionó por tener el valor más alto en el F1-Score. Los parámetros de este modelo son iteraciones 500, profundidad 3, tasa de aprendizaje de 0.1, iteraciones de estimación de hoja con un valor de 1 y la regulación L2 de hoja de 9. En las métricas se tiene un F1-Score de 85%, precisión de 86%, exactitud de entrenamiento de 94.5% y exactitud de validación de 92.3%. En la Figura 4.13 se encuentra la comparación del resultado del modelo respecto a los pixeles construidos y no construidos en el año 2020 con las reales.

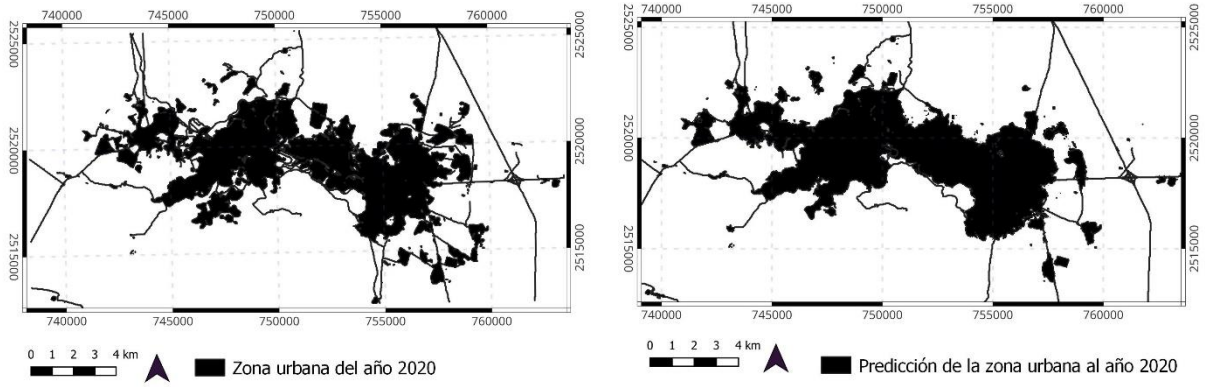

Figura 4.13. Comparación de la predicción realizado por el modelo catBoost y la zona urbana real para el año 2020.

4.3 Valores SHAP

Debido a que el cálculo de los valores SHAP se implica un gasto computacional elevado, se decidió seleccionar un subconjunto de los datos. La selección se realizó aplicando el algoritmo de k-means a los datos. Se determinó el número de grupos mediante el método del codo, al examinar la gráfica resultante, se decidió seleccionar cuatro grupos. El subconjunto creado es el 5 % del conjunto de datos. Los valores SHAP nos ayuda a indicar la importancia y dirección de las variables independientes que nos pueden ayudar a determinar si los píxeles se urbanizan o no.

4.3.1 Maquinas de Soporte Vectorial de Margen Suave (SVM)

En el caso del SVM el modelo del cual se calcularon los valores, se utilizaron los parámetros encontrados en la Tabla 4.10.

Los resultados de los valores SHAP se encuentran en la Figura 4.14. En el modelo de SVM, la variable que más impacta al modelo es el tipo de uso de suelo. Mientras más alta es el valor, es más probable que se urbanice. Esto quiere decir que si ya es urbanizado permanece urbanizados y si es suelo desnudo o vegetación cercana a zona urbana, el clasificador lo coloca como urbanizado.

Tabla 4.10. Parámetros de SVM del considerado el mejor modelo.

Parámetros	Valores
C	0.1
Kernel	Polinomial
Grado	5

La segunda variable que tiene mayor influencia es la probabilidad de urbanización. Esto indica

que el pixel se urbaniza en dependencia de cuantos pixeles dentro de su vecindario de 3x3 ya es urbanizado. Si el valor de esta probabilidad es alto el modelo lo indica como urbanizado. Esto va con la tercera variable, la distancia del centro. Si la distancia a la zona urbana es pequeña el modelo tiende a colocarlo como urbanizado. Por otro lado, la distancia a los centros de la ciudad es la cuarta variable que tiene importancia. Cuanto menor la distancia al centro, es más llamativa a la urbanización.

Por el contrario, las variables que no tuvieron un impacto fueron la elevación del terreno, los ingresos trimestrales, la pendiente y la dirección.

Figura 4.14. Valores SHAP de las diferentes variables del estudio en el modelo SVM.

4.3.2 Bosques Aleatorios (RF)

Los valores SHAP del modelo RF se calcularon en el modelo con los parámetros encontrados en la Tabla 4.11. En la Figura 4.15 se encuentran los resultados del método SHAP del modelo RF.

Tabla 4.11. Parámetros de RF que se consideraron del mejor modelo.

Parámetros	**Valores**
Números de árboles (n_estimators)	100
Profundidad máxima (max_depth)	10
Semilla de aleatoriedad (random_state)	1
Número máximo de características (max_features)	none

En el modelo RF seleccionado la variable que tiene más impacto en la decisión final del clasificador es la distancia a zonas urbanas, si la distancia es menor tiene más afectación a que este pixel sea clasificado como urbanizados. La segunda variable es la distancia al centro de la

ciudad, mientras más bajo el valor el modelo tiende a clasificar como urbanizado.

La tercera variable que tiene impacto en la decisión es el tipo de uso de suelo, si el pixel ya era urbanizado así permanece, mientras que si el pixel es suelo desnudo o vegetación el clasificador tiende a clasificarlo como urbano. O sea, si el valor es mayor influye en la decisión de urbanización.

Otra de las variables importantes es la distancia a cementerios, en este caso mientras la distancia es menor es más urbanizable. Esto se debe a que estas estructuras ya están rodeadas de zonas urbanas.

Por el contrario, las variables que no tienen tanta importancia en la decisión del clasificador es la densidad de total de viviendas, densidad de población ocupada, densidad de población y la densidad de población religiosa.

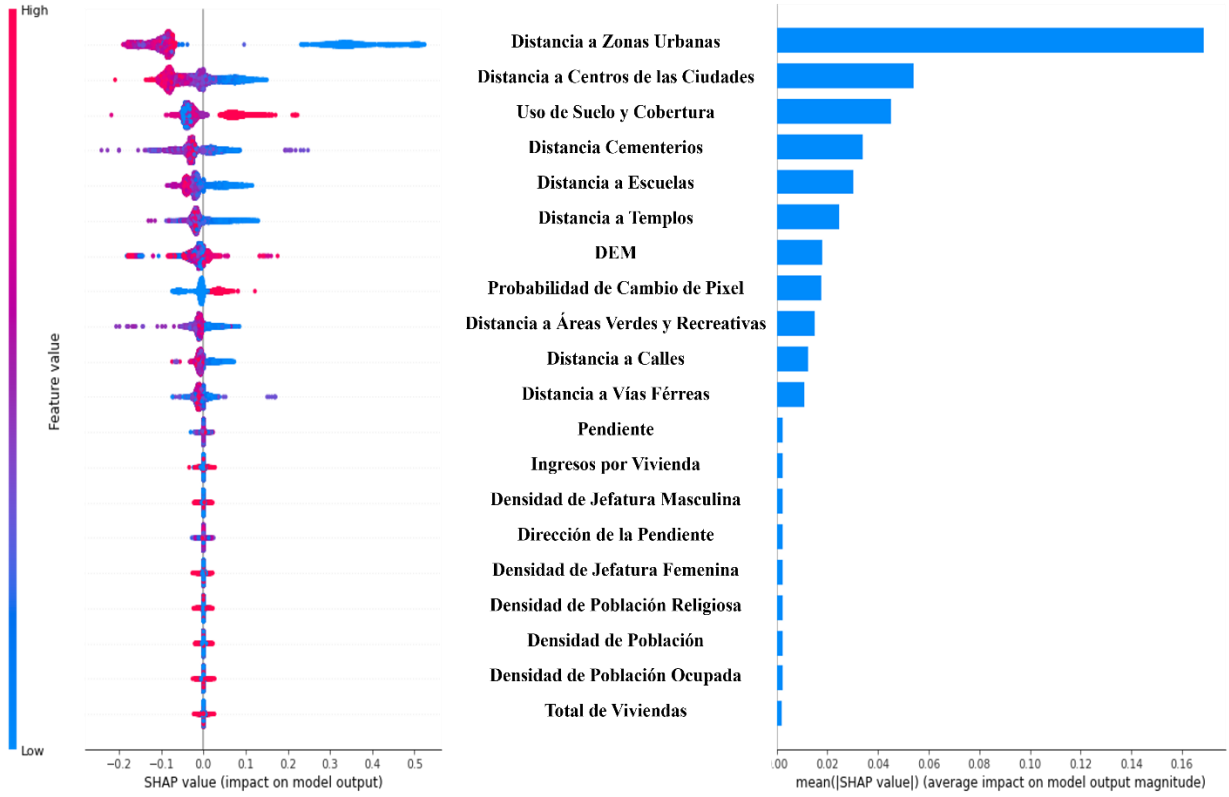

Figura 4.15. Valores SHAP de las diferentes variables del estudio en el modelo RF.

4.3.3 Boosting Categórico (catBoost)

Los valores SHAP nos ayuda a indicar la importancia y dirección de las variables independientes que nos pueden ayudar a determinar si los píxeles se urbanizan o no. En el caso del SVM el modelo del cual se calcularon los valores, se utilizaron los parámetros encontrados en la Tabla 4.12.

Los resultados de los valores SHAP se encuentran en la Figura 4.16. En el modelo de catBosst, la variable que más impacta al modelo es la distancia a zonas urbanas. En esta variable si el valor de la distancia es menor se clasifica como urbanizado, de caso contrario de que el valor de la distancia es mayor se clasificara como no urbanizado. La segunda variable es la distancia al centro de las ciudades, si esta distancia es pequeña el clasificador tiende a colocarlo como urbanizado.

Tabla 4.12. Parámetros del boosting categórico considerado el mejor.

Parámetros	**Valores**
Iteraciones (iterations)	500
Tasa de aprendizaje (learning_rate)	0.1
Profundidad (depth)	3
Iteraciones de estimación de hoja (leaf_estimation_iterations)	1
Regulación L2 de hoja (l2_leaf_reg)	9

La tercera variable es la distancia a escuelas, esta nos indica que si la distancia es menor es más probable su urbanización. Como cuarta variable importante es el tipo de uso de suelo. Mientras más alta es el valor, es más probable que se urbanice. Esto quiere decir que si ya es urbanizado permanece urbanizados y si es suelo desnudo o vegetación cercana a zona urbana el clasificador lo coloca como urbanizado.

Por último, las variables que no tienen un impacto significativo en la decisión del modelo catBoost son la densidad de jefatura masculina, densidad de población religiosa, dirección de la pendiente y la pendiente.

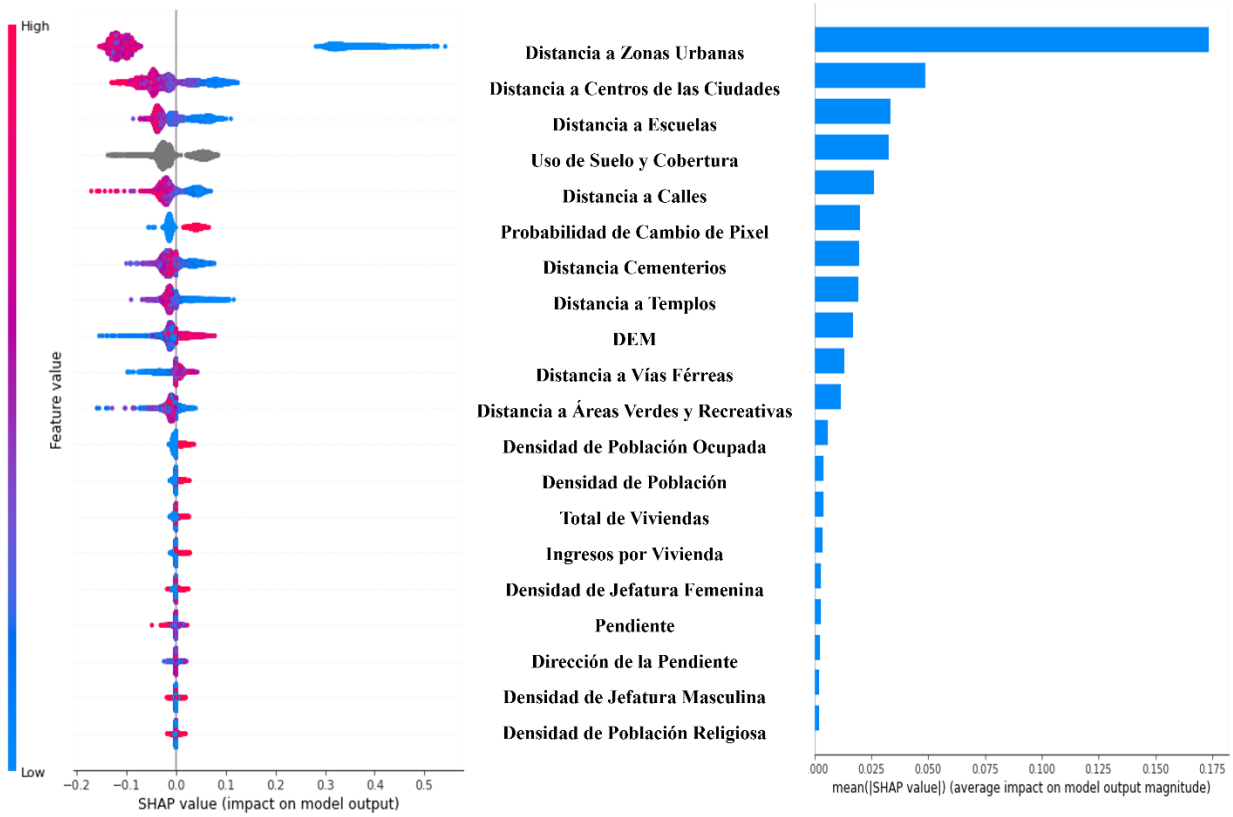

Figura 4.16. Valores SHAP de las diferentes variables del estudio en el modelo catBoost.

4.4 Escenario de expansión urbana al 2030

Con los modelos presentados en las Tabla 4.10, Tabla 4.11, Tabla 4.12 de los algoritmos

Máquinas de Soporte Vectorial, Bosques Aleatorios y Boosting Categórico respectivamente. Se realizó la predicción de las zonas urbanas para el año 2030, las cuales se muestran en la Figura 4.17.

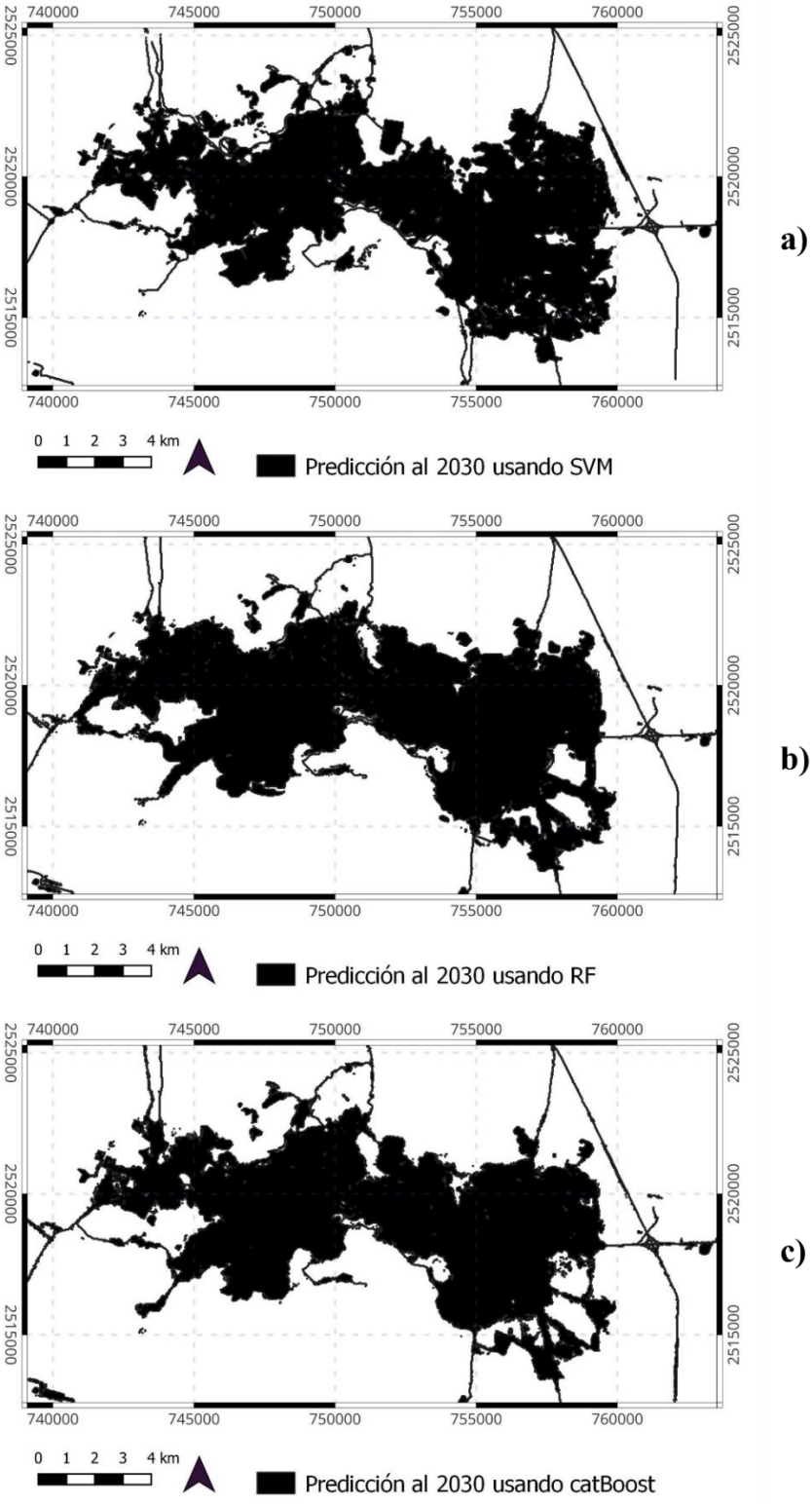

Figura 4.17. Predicción al 2030 de los diferentes modelos, a) resultado de las Maquinas de Soporte Vectorial, b) Bosques Aleatorios y c) Boosting Categórico.

Teniendo en cuenta nuestros contextos, se llevó a cabo una estimación de la tasa de cambio correspondiente al período comprendido entre 2020 y 2030, la cual se puede apreciar en la Tabla 4.13. En esta etapa se puede observar que el incremento en los escenarios es superior al observado en las otras fases analizadas. Se tiene una tasa de cambio del 8.11 % en el período de 2020 a 2030 predicho por SVM en el período de 2020 a 2030. La tasa de variación experimentada en el modelo de Bosques Aleatorios es del 9.57 %, siendo la mayor de los tres escenarios. El modelo de Boosting Categórico presenta una tasa de cambio menor, con un 7.51 %.

Tabla 4.13. Tasa de expansion del area de estudio en los periodos 2020 a 2030 en cada uno de los modelos.

	(2020-2030)		
	SVM	**RF**	**catBoost**
Cantidad de pixeles que cambiaron	29173	34408	27026
Cantidad de pixeles que no cambiaron	330329	325094	332476
Tasa de cambio	8.11%	9.57%	7.51%

Además de la tasa de expansión entre 2020 y 2030, se calculó una matriz de tabulación cruzada. Esta matriz incluye únicamente las zonas urbanas y no urbanas, ya que el resultado es una clasificación binaria. Con la tabulación cruzada se puede observar cuántos kilómetros cuadrados cambiaron de zonas no urbanas a urbanas en cada una de las predicciones. La matriz correspondiente se presenta en la Tabla 4.14, adicionalmente se creó el mapa de cambios en la Figura 4.18.

En los resultados obtenidos mediante el modelo de predicción SVM, se observa que 240.40 km^2 se mantuvieron como áreas no urbanizadas, mientras que 26.3 km^2 cambiaron de no urbanas a urbanas. Por otro lado, 56.7 km^2 permanecieron como zonas urbanas. Sin embargo, 0.17 km^2 de las zonas urbanas pasaron a ser zonas no urbanas.

Tabla 4.14. Matriz de cambio de uso de suelo y cobertura (superficie en km^2) de 2020 a 2030 de las predicciones de SVM, RF y catBoost.

2030 / **2020**	**SVM**		**RF**		**catBoost**	
	Zona No Urbana	**Zona Urbana**	**Zona No Urbana**	**Zona Urbana**	**Zona No Urbana**	**Zona Urbana**
Zona No Urbana	240.40	26.3	235.69	31.0	233.92	32.7
Zona Urbana	0.17	56.7	0.36	56.5	0.24	56.7

Asimismo, en la predicción realizada por el algoritmo de bosques aleatorios, 235.69 km^2 se conservaron como zonas no urbanas, mientras que 31.0 km^2 se transformaron de zonas no urbanas a urbanas. Por otro lado, 56.5 km^2 permanecieron como zonas urbanas y 0.36 km^2 indican una contracción, ya que esta superficie era urbana y se convirtió en no urbana.

En cuanto a los resultados obtenidos con el algoritmo CatBoost, se observa que 233.92 km^2 se conservaron como zonas no urbanas, mientras que 32.7 km^2 se transformaron de zonas no urbanas a urbanas. Por otro lado, 56.7 km^2 permanecieron como zonas urbanas y 0.24 km^2 pasaron de zonas urbanas a no urbanas.

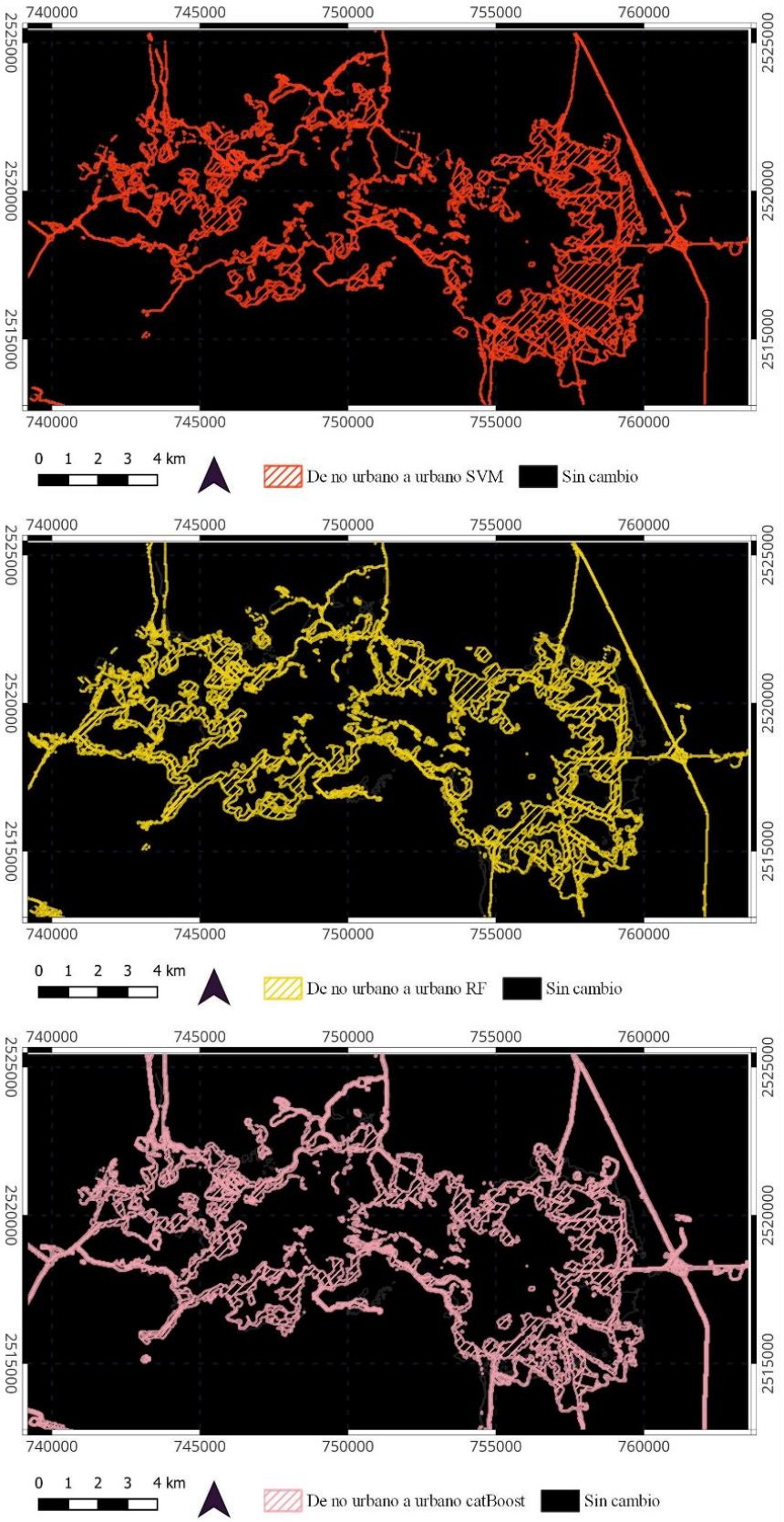

Figura 4.18. Mapas de cambio de uso de suelo de las predicciones a 2030, a) Maquinas de soporte Vectorial, b) Bosques Aleatorios y c) Boosting categórico.

Al comparar los resultados obtenidos con los tres algoritmos de predicción, se observan algunas diferencias clave en la clasificación de las zonas urbanas y no urbanas. El algoritmo SVM identificó 0.17 km^2 como zonas que pasaron de urbanas a no urbanas, mientras que el algoritmo de bosques aleatorios (RF) reportó 0.36 km^2 en esta categoría, y CatBoost identificó 0.24 km^2. Estos valores pueden interpretarse como predicciones de contracción urbana o como errores de clasificación del modelo. Sin embargo, dado que el objetivo principal del análisis es la expansión, se sugiere que son errores de clasificación.

En términos de expansión, CatBoost predijo que 32.7 km^2 de zonas no urbanas se transformaron en urbanas, el valor más alto entre los modelos evaluados. Esto recomienda que CatBoost tiene una mayor sensibilidad para detectar nuevas áreas urbanizadas. Por otro lado, SVM predijo 26.3 km^2 de expansión, el valor más bajo, lo que podría indicar que es más conservador al identificar cambios, pero también es menos propenso a errores en zonas de transición. El modelo de bosques aleatorios (RF) predijo 31.0 km^2 de expansión, mostrando un comportamiento intermedio entre CatBoost y SVM.

Para profundizar en el análisis, se dividió el área de estudio en cuatro regiones: noreste (NE), noroeste (NW), suroeste (SW) y sureste (SE) (Figura 4.6), con el propósito de llevar a cabo un análisis detallado. En la Figura 4.19a se muestran las áreas, en kilómetros cuadrados, de los diferentes cuadrantes según las predicciones realizadas por nuestros modelos. En este contexto, se puede constatar que, en las proyecciones realizadas por el SVM, el sector SE contiene la mayor área edificada, seguido por la zona NW. Por el contrario, las predicciones de RF y CatBoost indican un incremento más significativo de áreas en la sección NW, seguida de la SE. Además, el cuadrante con menor área construida en los tres modelos fue el SW.

Para determinar la magnitud del incremento en las áreas entre el año 2020 y las predicciones de nuestros modelos (Figura 4.19b), en los tres modelos se observa que el crecimiento mayor se dio en el cuadrante SE, lo cual es consistente con el patrón de crecimiento de años anteriores. El cuadrante que experimenta mayor crecimiento en el modelo SVM y RF fue el NE, mientras que en CatBoost fue el NW. Finalmente, se aprecia que el modelo que presenta un crecimiento más homogéneo entre las regiones es CatBoost.

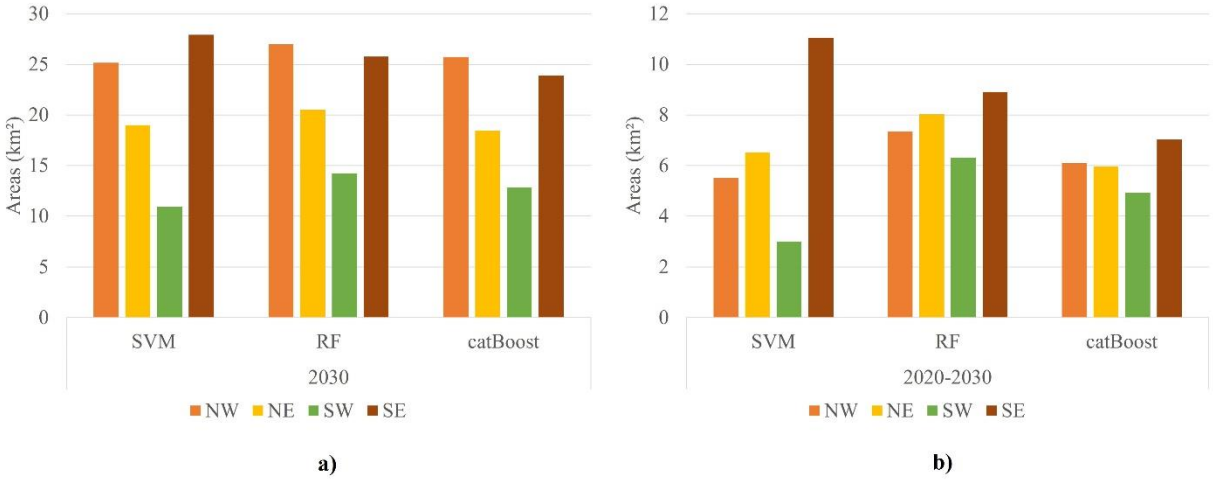

Figura 4.19. a) Area en kilometro cuadrados de construccion en los diferentes excenarios predichos por los modelos SVM, RF y catBoost. b) Diferencia de las ares en los sectores del 2020 a 2030 de los modelos.

Tras examinar las proyecciones de expansión urbana en diversas regiones, fue imperativo examinar la intensidad de la expansión en cada área con el fin de obtener una perspectiva más

precisa del proceso de urbanización. Para ello, se calculó el Índice de Intensidad de Expansión Urbana (IIEU), que permite cuantificar el ritmo de la expansión de las zonas urbanas en cada cuadrante y comparar los resultados entre los modelos evaluados. Este índice aporta una medida adicional para interpretar no solo la extensión del área urbanizada, sino también la velocidad y concentración de dicha expansión. Estos resultados se encuentran en la Tabla 4.15 de las predicciones de SVM, RF y catBoost, respectivamente.

Los resultados del índice de la predicción realizada por el modelo SVM (Tabla 4.15), indican en el cuadrante SW una velocidad de desarrollo baja. Los cuadrantes NW y NE tienen un desarrollo de velocidad media y, por último, SE con velocidad rápida. Por otro lado, los índices calculados en la predicción de los bosques aleatorios (Tabla 4.15) tienen desarrollo en velocidad media y rápida. Al igual que en el modelo anterior, el sector SE es el que tiene el índice mayor, categorizado como un desarrollo de velocidad rápida. Los otros sectores (NW, NE y SW) se categorizaron en un desarrollo de velocidad media. De estos cuadrantes el que tuvo un índice mayor fue el NE, seguido de NW y por último el SW.

Tabla 4.15. Valores del IIEU en el periodo 2020 a 2030 de los resultados de los algoritmos SVM, RF y catBoost.

Sector	SVM		RF		catBoost	
	IIEU	**Velocidad**	**IIEU**	**Velocidad**	**IIEU**	**Velocidad**
NW	0.68	Media	0.91	Media	1.07	Rápida
NE	0.81	Media	1.00	Media	1.00	Media
SW	0.37	Baja	0.78	Media	0.76	Media
SE	1.36	Rápida	1.10	Rápida	1.19	Rápida

Por último, se tienen los resultados de los índices calculados del algoritmo catBoost. El IIEU se calculó para cada uno de los cuadrantes. Los resultados muestran que el sector SE y NW, tuvieron un desarrollo de velocidad rápida. De estos cuadrantes SE fue el de índice mayor seguido del sector NW. Mientras tanto, NE y SW, tuvieron un desarrollo de velocidad media. Siendo el de mayor índice NE seguido del SW.

Comparando estos resultados con los obtenidos contra el periodo 2010 a 2020. En el caso de los algoritmos SVM podemos observar que los sectores que cambiaron fueron NW y SE. NW cambió de velocidad baja a velocidad media y SE de velocidad baja a rápida Por otro lado, en el caso de los bosques aleatorios, NE continúa siendo de desarrollo medio. Los cuadrantes modificados fueron NW, SW y SE. NW cambio de velocidad baja a velocidad media. SW de baja a media y por último SE de baja a rápida. La comparación del catBoost muestran que, al igual que los otros dos algoritmos, el sector NE continua con una velocidad de desarrollo media. NW cambio de baja a rápida. Mientras que SW se movió de baja a media y SE de baja a rápida.

Revisando los resultados se puede observar que los tres algoritmos concuerdan en una velocidad de desarrollo rápido en el sector SE. Al igual que la continuidad del desarrollo medio en el sector NE. En el aspecto del sector NW la predicción del SVM y RF concuerdan en una velocidad media, mientras que catBoost lo coloca en velocidad rápida. Por otro lado, en el cuadrante SW la predicción del RF y catBoost concuerdan en una velocidad media, mientras que SVM indica una velocidad baja.

A partir de los resultados obtenidos, la velocidad de expansión en los cuadrantes varía en función del modelo empleado. El modelo SVM muestra una combinación de velocidades, lo que sugiere una predicción más conservadora en áreas con expansión moderada. Por otro lado, el modelo de

bosques aleatorios (RF) predice una velocidad de expansión, ofreciendo un enfoque balanceado y consistente. Finalmente, CatBoost muestra velocidades rápidas, indicando una mayor sensibilidad para detectar áreas con crecimiento acelerado.

Cápitulo 5. Discusión y conclusión

5.1 Discusión

La rápida urbanización es la causa principal del cambio de uso de suelo de áreas vegetales, agricultura y suelo desnudo a urbanización [108]. En este estudio, hemos explorado los patrones y tendencias de la expansión urbana en las ciudades Zacatecas y Guadalupe. El periodo analizado fue el comprendido del 2000 al 2020. Se creó un conjunto de datos de veintiuna variables independientes, para cada uno de los años (2000, 2010 y 2020). Los datos del año 2000 se usaron para el entrenamiento, 2010 para la fase de prueba y los datos de 2020 para realizar la predicción al año 2030.

Se utilizaron los algoritmos de SVM de margen suave, bosques aleatorios y Boosting categórico. Posteriormente, se llevó a cabo un análisis entre el rendimiento de los algoritmos y un análisis de la predicción de la expansión urbana. La capacidad de SVM para manejar espacio de alta dimensión [26], la robustez e interpretabilidad de los bosques aleatorios [84] y la capacidad de CatBoost para gestionar características categóricas de manera efectiva [87] , son unas de las razones para la elección de estos algoritmos.

Posterior al entrenamiento, la evaluación de los modelos se realizó usando la exactitud, precisión, F1-Score y Recall. Apoyándose principalmente del F1-Score para la configuración del conjunto de datos. Seleccionado el mejor modelo de cada uno de los algoritmos de aprendizaje automático, se calcularon los valores SHAP.

En el análisis de los resultados llevados a cabo destaca el desempeño del SVM con margen suave, debido a que logró una precisión de un 87.9 %, F1-Score de un 86.3 % y recall de un 84.7 %. El modelo SVM tuvo un buen desempeño en todas las métricas de evaluación en comparación con RF y catBoost. Sin embargo, RF y catBoost tienen una predicción homogénea en los sectores del área de estudio, esto puede ser beneficioso para la ciudad y futuras decisiones.

De acuerdo con los resultados de los valores de SHAP las variables clave que comparten los modelos fueron: mapas de uso de suelo y cobertura terrestre, distancia a zonas urbanas y la distancia a los centros de las ciudades. Los resultados de los mapas de uso de suelo son consistentes con los encontrados en el análisis de cambio de uso de suelo. Las distancias más cercanas a las zonas urbanas son algo que otros autores señalan en sus hallazgos [26], [28]. A pesar de estos hallazgos, hay diferentes autores [6] que consideran los aspectos DEM y la pendiente como una variable con un alto impacto en la predicción de la expansión urbana. Sin embargo, en los modelos resultantes, estas variables tienen un pequeño impacto en la predicción, al igual que lo encontrado por los trabajos de [108], [109]. Adicionalmente, las variables con menos influencia son: densidad de población religiosa, dirección de la pendiente y los ingresos por vivienda.

En la predicción realizada por los modelos al año 2030, se observó que la distancia a las calles tiene una influencia. Esto también se aprecia en los valores SHAP de catBoost y en el trabajo de [26]. Adicionalmente, la predicción se concentra en la ciudad de Guadalupe; la razón puede ser la conexión vial. En el área de Zacatecas, hay una menor densidad de carreteras y calles, lo que puede provocar una menor expansión en esa dirección. En cuanto a las variables culturales, los resultados indican que la distancia a los templos y la densidad de población religiosa no exhibieron un efecto significativo, por lo que se consideraría utilizar otras variables que sean más

llamativas para habitar un lugar.

Finalmente, consideramos que los urbanistas deben considerar la importancia de la construcción de calles, ya que estas incrementan la probabilidad de construcción y la proximidad de nuevos asentamientos al centro de la ciudad.

5.2 Conclusión

En este estudio, se realizó la segmentación de imágenes satelitales de la zona de análisis utilizando el algoritmo FRFCM para los años 2000, 2010 y 2020. Los resultados de la segmentación se evaluaron mediante índices reconocidos como el índice de Jaccard, Kappa de Cohen y el MCC (Coeficiente de correlación de Matthews), obteniendo valores satisfactorios por encima de 0.7. A partir de esta segmentación, se categorizaron los resultados y se generaron los mapas de uso de suelo.

Tras la creación de los mapas, se llevó a cabo un análisis de los cambios en los usos de suelo. Los mapas se dividieron en cinco clases: agricultura, cuerpo de agua, vegetación, suelo desnudo y zona urbana. Inicialmente, se calculó la tasa de cambio dinámico para cada categoría en los dos periodos analizados. Los resultados muestran que la zona urbana presentó la mayor tasa de cambio en ambos periodos. Los cuerpos de agua también experimentaron un aumento, mientras que el suelo desnudo disminuyó. La agricultura disminuyó en el primer periodo y aumentó en el segundo, mientras que la vegetación mostró un crecimiento en el segundo periodo.

Para la detección de cambios de uso de suelo, se utilizaron matrices de cambio mediante tabulación cruzada de las categorías, expresadas en kilómetros cuadrados. En el primer periodo, de 2000 a 2010, los cambios más significativos fueron de vegetación a suelo desnudo, seguidos de agricultura a suelo desnudo, y, en tercer lugar, el cambio de suelo desnudo a vegetación. En el segundo periodo, de 2010 a 2020, los cambios más marcados fueron de suelo desnudo a agricultura, de vegetación a suelo desnudo, y finalmente, de suelo desnudo a vegetación.

El modelado y predicción de la expansión urbana de las ciudades de Zacatecas y Guadalupe se llevó a cabo, con los algoritmos SVM con margen suave, RF y CatBoost. El mejor algoritmo para este problema fue SVM con margen suave. Adicionalmente, se calcularon los valores SHAP para interpretar los resultados de los modelos, revelando que las variables clave en común fueron los mapas de uso y cobertura del suelo, las distancias a las zonas urbanas y las distancias a los centros de las ciudades. De forma contraria, la dirección de la pendiente, densidad de población religiosa y los ingresos por vivienda tuvieron una influencia menor en la predicción.

Los hallazgos resaltan la importancia de la proximidad a las áreas urbanas existentes y los centros de las ciudades para impulsar la expansión urbana. El uso de suelo y la cobertura terrestre son herramientas invaluables para monitorear el cambio ambiental, ayudándonos a proteger la flora y fauna local, al tiempo que promovemos la inclusión de espacios verdes y la expansión vertical. Asimismo, la comprensión de la proximidad de los centros urbanos propicia una planificación urbana más estratégica, asegurando que la expansión urbana pueda ser equilibrada con la preservación de las áreas naturales.

Comprender el impacto de las calles y carreteras puede ayudar a prevenir la congestión del tráfico. A pesar de que el modelo refleja la interacción de las variables y la forma en la que se expande la ciudad. Consideramos que existen variables que no se exploraron. Lo que podría aumentar la precisión de los modelos. Aunque nuestro estudio se centra en Zacatecas y Guadalupe, los conocimientos adquiridos sobre el impacto de factores cruciales en la expansión urbana sirven

como una base valiosa para futuras investigaciones en otras regiones. Sin embargo, para confirmar la generalización de estos hallazgos, es esencial una validación en diferentes áreas geográficas. Por otra parte, sería beneficioso para futuros trabajos incluir variables como proyecciones de población, valores de propiedades y otros factores culturales y sociales que demuestren una relación más sólida.

5.3 Objetivos alcanzados

Al comienzo de esta investigación, se establecieron los objetivos descritos en la sección 1.6 y, los cuales fueron establecidos y cumplidos a lo largo del proyecto. A continuación, se detallan los principales logros alcanzados:

- El primer objetivo consistió en realizar la segmentación. En este paso se utilizaron algoritmos difusos debido a la resolución del píxel; el utilizado fue FRFCM. Posteriormente, se crearon las máscaras que nos ayudaron a evaluar la segmentación de las zonas urbanas. Las métricas de evaluación seleccionadas fueron el índice de Jaccard, el coeficiente Cohen Kappa y el coeficiente de correlación de Matthew.
- Como segundo objetivo se planteó la elaboración de los mapas de uso de suelo del área de estudio, para cada uno de los años 2000, 2010 y 2020. Estos mapas se realizaron combinando datos vectoriales y los resultados de la segmentación. Este procesamiento se realizó utilizando la herramienta de uso libre llamada QGIS.
- Posteriormente, se planteó la detección y análisis del cambio de uso de suelo entre los mapas. Los periodos analizados fueron de 2000 a 2010 y de 2010 a 2020. El análisis se realizó creando tablas de tabulación cruzada, tasas de cambio e índices de intensidad de expansión urbana. A partir de las tablas de tabulación cruzada se realizaron las matrices de cambio expreso en superficies en áreas en kilómetros cuadrados. Aquí se detectó que la superficie que tiene más cambio es la transformación a suelo desnudo, ya sea de agricultura y vegetación. O viceversa, de suelo desnudo a agricultura o vegetación. Asimismo, el uso que cambió más la zona urbana fue el suelo desnudo. En el caso del índice de expansión urbana, se tiene que en los dos periodos se tuvo en su mayoría la velocidad de desarrollo baja y un solo cuadrante en desarrollo medio.
- Para lograr el siguiente objetivo se creó un conjunto de datos que incluían datos demográficos, socioeconómicos y ambientales. Se usaron datos de los censos en los años 2000, 2010 y 2020. Teniendo el conjunto de datos se planteó la implementación de los algoritmos de aprendizaje automático. Para lograr este objetivo se creó una búsqueda exhaustiva en cuadrícula de los parámetros de cada uno de los algoritmos. Los algoritmos seleccionados fueron Máquinas de Soporte Vectorial, Bosques Aleatorios y Boosting Categórico. La evaluación de estos algoritmos se realizó con las métricas supervisadas de F1-Score, Exactitud, Precisión y Recall. Seleccionada el mejor modelo de la búsqueda exhaustiva, se llevó a cabo la predicción al año 2030. Por último, se efectuó un análisis al mapa de expansión al año 2030 de cada uno de los algoritmos y llevar a cabo un análisis de la expansión urbana. En este se calcularon las áreas que cambiaron relación al año 2020, el índice de intensidad de expansión urbana, tasas de cambio y la tabulación cruzada del tipo de uso de suelo (Zona Urbana y Zona no Urbana). El análisis se realizó dividiendo la zona de estudio en cuatro áreas menores. Adicionalmente, se utilizó la técnica de SHAP para conocer como los algoritmos tomaban una decisión a partir de los datos.

Se puede apreciar que todos los objetivos específicos planteados se lograron, llevando así al cumplimiento de nuestro objetivo general el cual era la propuesta de una metodología basada en aprendizaje automático para predecir la expansión urbana empleando imágenes satelitales segmentadas, datos demográficos y ambientales.

5.4 Hipótesis demostradas

En el presente estudio, se planteó la hipótesis descrita en la sección 1.7. Propone la posibilidad de predecir la expansión urbana de las ciudades de Zacatecas y Guadalupe mediante un modelo basado en técnicas de aprendiza automático (ML). Este modelo utilizaría datos demográficos, ambientales y mapas de uso de suelo obtenidos a través de la segmentación de imágenes satelitales. El rendimiento de los modelos de ML debe tener un rendimiento medido por un F1-Score superior al 80 %

Los resultados obtenidos a partir de la experimentación realizada con los algoritmos Máquinas de Soporte Vectorial, Bosques Aleatorios y Boosting Categórico, demostró la hipótesis en su totalidad. Los modelos entrenados con los datos demográficos, ambientales y mapas de uso de suelo lograron predecir la expansión urbana con un F1-Score bueno, mayor al 80 %.

En el resultado se resalta el desempeño del SVM con un F1-Score de un 86.3 %, RF con un 85.3 % y catBoost de un 85 %. El rendimiento de los modelos demuestra que las técnicas de aprendizaje automático pueden ser herramientas precisas para la predicción de la expansión urbana en las ciudades de Zacatecas y Guadalupe.

5.5 Contribuciones de la investigación

Esta investigación aporta una serie de contribuciones significativas en el ámbito académico y en el práctico. Se abordó la predicción de la expansión urbana mediante el uso de técnicas de aprendizaje automático. Se desarrolló y válido un modelo basado en aprendizaje automático con datos demográficos, ambientales y mapas de uso de suelo derivados de imágenes satelitales segmentadas. Este modelo combina datos multivariables nos permiten capturar los factores complejos que influyen en la expansión urbana.

Los hallazgos ofrecen una base sólida para la toma de decisiones en la planificación urbana de las ciudades de Zacatecas y Guadalupe. Al encontrar las áreas con mayor probabilidad de expansión, los urbanistas pueden diseñar estrategias más efectivas y sostenibles.

Adicionalmente, la metodología propuesta puede ser adaptada y aplicada en otras regiones con características urbanas similares. Esto permite la creación de futuros estudios que busquen comprender y predecir la expansión urbana en diferentes regiones geográficas, usando técnicas de aprendizaje automático. Sin embargo, es esencial realizar una validación en nuevas áreas de estudio.

Por último, este trabajo refuerza el uso de modelos de aprendizaje automático para abordar problemas complejos en el ámbito de la planificación urbana, ofreciendo una solución eficiente para la predicción de la expansión urbana a mediano y largo plazo.

5.6 Limitaciones

A pesar de que el trabajo presentado ofrece hallazgos interesantes y aplicables, tiene diferentes limitaciones. Dentro de estas limitaciones se encuentra la calidad y resolución de las imágenes

satelitales. Esto pudo haber afectado la precisión de la segmentación del uso de suelo y, por ende, afectar el rendimiento del modelo predictivo.

Por otra parte, el estudio se centró exclusivamente en las ciudades de Zacatecas y Guadalupe. Como se mencionó anteriormente, sería necesaria una validación para la aplicación de los modelos en otras áreas de estudio.

Por último, el modelo no incluye algunas variables que se puedan considerar importantes para la expansión urbana. Algunas de estas pueden ser factores económicos detallados, aspectos políticos y aspectos sociales. La falta de estas características puede influir en la capacidad del modelo para capturar ciertas dinámicas de expansión.

5.7 Trabajos Futuro

A pesar de que los hallazgos de esta investigación han demostrado la capacidad de las técnicas de aprendizaje automático para anticipar la expansión urbana de las ciudades de Zacatecas y Guadalupe. Sin embargo, existen diferentes áreas de mejora o de exploración en estudios futuros.

Mejorar la calidad de los datos satelitales es una vía de mejora en nuestro estudio. Utilizando imágenes satelitales con una resolución espacial menor, mejoraría la segmentación y la calidad de los mapas de uso de suelo.

Asimismo, incluir datos adicionales que representen las interacciones sociales y políticas. Adicionalmente, utilizar la zonificación de áreas descritas por INEGI, ya que cada una de ellas tiene interacción diferente. Esta zonificación enriquecería la calidad del análisis al capturar factores de expansión vigentes en la actualidad que no han sido considerados.

Por último, la exploración de métodos más avanzados para la optimización de hiperparámetros y otras técnicas de aprendizaje automático podrían crear modelo con mejor rendimiento.

Referencias

[1] N. United, D. of Economic, S. Affairs, and P. Division, "World Urbanization Prospects The 2018 Revision," 2018.

[2] W. S. Ahmad, S. Jamal, A. Sharma, and I. H. Malik, "Spatiotemporal analysis of urban expansion in Srinagar city, Kashmir," *Discover Cities*, vol. 1, no. 1, Jun. 2024, doi: 10.1007/s44327-024-00009-3.

[3] Mexico. Secretaría de Desarrollo Social., Mexico. Secretaría de Gobernación., and Consejo Nacional de Población (Mexico), *Catálogo : sistema urbano nacional 2012*, 2012th ed.

[4] Y. Deng, B. Fu, and C. Sun, "Effects of urban planning in guiding urban growth: Evidence from Shenzhen, China," *Cities*, vol. 83, pp. 118–128, Dec. 2018, doi: 10.1016/j.cities.2018.06.014.

[5] C. G. Rangel, T. C. Soto, V. Mata, and E. Jiménez López, "Un modelo de expansión urbana no estacionario en el espacio: Autómatas Celulares y Regresión Geográficamente Ponderada," in *La Situación Demográfica en México*, México: CONAPO, 2022, pp. 95–113.

[6] M. Kim and G. Kim, "Modeling and Predicting Urban Expansion in South Korea Using Explainable Artificial Intelligence (XAI) Model," *Applied Sciences (Switzerland)*, vol. 12, no. 18, Sep. 2022, doi: 10.3390/app12189169.

[7] Z. Zhang *et al.*, "Urban Expansion in China Based on Remote Sensing Technology: A Review," Oct. 01, 2018, *Science Press*. doi: 10.1007/s11769-018-0988-9.

[8] J. González-Madrigal, H. Solano-Lamphar, and M. Ramírez, "La contaminación lumínica como aproximación a la planeación urbana de ciudades mexicanas," 2020.

[9] N. Khanal, K. Uddin, M. A. Matin, and K. Tenneson, "Automatic detection of spatiotemporal urban expansion patterns by fusing OSM and Landsat data in Kathmandu," *Remote Sens (Basel)*, vol. 11, no. 19, Oct. 2019, doi: 10.3390/rs11192296.

[10] S. Navarro, M. Gabriel Orozco-del-Castillo, and J. Carlos Valdiviezo-N, "Análisis del crecimiento urbano y su relación con el incremento de temperaturas en la ciudad de Mérida utilizando imágenes satelitales Aerospace science and technology View project boundary element method View project," 2018. [Online]. Available: https://www.researchgate.net/publication/332686540

[11] Y. Norouzi, "Measuring Land Use Changes and Quantifying Urban Expansion Using Remote Sensing and GIS Techniques - A Case Study of Qom," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Copernicus Publications, Jan. 2023, pp. 609–615. doi: 10.5194/isprs-annals-X-4-W1-2022-609-2023.

[12] I. Amri, I. W. Prabaswara, and S. R. Giyarsih, "Spatio-Temporal Analysis of Post-Disaster Built-up Expansion in Banda Aceh City, Indonesia," in *Proceedings - 2020 6th International Conference on Science and Technology, ICST 2020*, Institute of Electrical and Electronics Engineers Inc., 2020. doi: 10.1109/ICST50505.2020.9732823.

[13] Y. Hou, W. Kuang, and Y. Dou, "Observing the compact trend of urban expansion patterns in global 33 megacities during 2000–2020," *Journal of Geographical Sciences*, vol. 33, no. 12, pp. 2359–2376, Dec. 2023, doi: 10.1007/s11442-023-2180-0.

[14] V. Chaturvedi and W. T. de Vries, "Machine Learning Algorithms for Urban Land Use Planning: A Review," Sep. 01, 2021, *MDPI*. doi: 10.3390/urbansci5030068.

[15] F. Nugroho and O. I. Al-Sanjary, "A review of simulation urban growth model," *International Journal of Engineering and Technology(UAE)*, vol. 7, no. 4, pp. 17–23, 2018, doi: 10.14419/ijet.v7i4.11.20681.

[16] S. W. Wang, L. Munkhnasan, and W. K. Lee, "Land use and land cover change detection and prediction in Bhutan's high altitude city of Thimphu, using cellular automata and Markov chain," *Environmental Challenges*, vol. 2, Jan. 2021, doi: 10.1016/j.envc.2020.100017.

[17] C. Kara and N. Doratlı, "Predict and simulate sustainable urban growth by using gis and mce based ca. Case of famagusta in northern cyprus," *Sustainability (Switzerland)*, vol. 13, no. 8, Apr. 2021, doi: 10.3390/su13084446.

[18] X. Liang, X. Liu, X. Li, Y. Chen, H. Tian, and Y. Yao, "Delineating multi-scenario urban growth

boundaries with a CA-based FLUS model and morphological method," *Landsc Urban Plan*, vol. 177, pp. 47–63, Sep. 2018, doi: 10.1016/j.landurbplan.2018.04.016.
[19] H. Wang, J. Guo, B. Zhang, and H. Zeng, "Simulating urban land growth by incorporating historical information into a cellular automata model," *Landsc Urban Plan*, vol. 214, Oct. 2021, doi: 10.1016/j.landurbplan.2021.104168.
[20] E. López, G. Bocco, M. Mendoza, and E. Duhau, "Predicting land-cover and land-use change in the urban fringe A case in Morelia city, Mexico," *Lanscape and Urban Planning*, vol. 55, pp. 271–285, 2001.
[21] C. Garrocho, E. Jiménez, and T. Chávez-Soto, "Expansión de la ciudad: un instrumento de simulación de escenarios para los sectores público y privado," in *La Situación Demográfica en México*, México: CONAPO, 2020, pp. 195–219. [Online]. Available: https://medium.com/espanol/
[22] C. Garrocho, T. C. Soto, and E. Jiménez López, "Autómata Celular Metro-NASZ: laboratorio experimental de expansión urbana," in *La Situación Demográfica en México*, México: CONAPO, 2021, pp. 149–175. [Online]. Available: https://medium.
[23] L. A. Lopez-Rivera, M. Romero-Huertas, and D. Valle-Cruz, "A neural network-cellular automata based model for predicting vertical urban growth," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jun. 2021, pp. 548–550. doi: 10.1145/3463677.3463751.
[24] X. Li and P. Gong, "Urban growth models: progress and perspective," *Sci Bull (Beijing)*, vol. 61, no. 21, pp. 1637–1650, Nov. 2016, doi: 10.1007/s11434-016-1111-1.
[25] M. Kim, D. Kim, D. Jin, and G. Kim, "Application of Explainable Artificial Intelligence (XAI) in Urban Growth Modeling: A Case Study of Seoul Metropolitan Area, Korea," *Land (Basel)*, vol. 12, no. 2, Feb. 2023, doi: 10.3390/land12020420.
[26] F. Karimi, S. Sultana, A. Shirzadi Babakan, and S. Suthaharan, "An enhanced support vector machine model for urban expansion prediction," *Comput Environ Urban Syst*, vol. 75, pp. 61–75, May 2019, doi: 10.1016/j.compenvurbsys.2019.01.001.
[27] B. Mirbagheri and A. Alimohammadi, "Integration of local and global support vector machines to improve urban growth modelling," *Canadian Historical Review*, vol. 7, no. 9, Sep. 2018, doi: 10.3390/ijgi7090347.
[28] F. Karimi, S. Sultana, A. S. Babakan, and S. Suthaharan, "Urban expansion modeling using an enhanced decision tree algorithm," *Geoinformatica*, vol. 25, no. 4, pp. 715–731, Oct. 2021, doi: 10.1007/s10707-019-00377-8.
[29] M. Asif *et al.*, "Modelling of land use and land cover changes and prediction using CA-Markov and Random Forest," *Geocarto Int*, vol. 38, no. 1, 2023, doi: 10.1080/10106049.2023.2210532.
[30] O. R. Pérez, M. A. L. Sánchez, X. C. Camacho, and M. A. Robledo, "Methodology for mining economic assimilation in Zacatecas, Mexico," *Economia, Sociedad y Territorio*, vol. 20, no. 62, pp. 241–272, May 2020, doi: 10.22136/est20201415.
[31] INEGI, "Comunicado de prensa num 57/21," 2021.
[32] O. P. Garbutt, *Metrópolis de México 2020 Primera edición*. 2024. [Online]. Available: https://www.gob.mx/sedatuhttps://www.inegi.org.mx
[33] T. y U. Secretaría de Desarrollo Agrario, "Delimitación de las zonas metropolitanas de México 2015," 2018. [Online]. Available: https://www.gob.mx/sedatu
[34] R. Almejor Hernández, J. García Galeana, and I. Benítez Villegas, "La urbanización en México 2010-2030: un esbozo de los retos y oportunidades asociados al crecimiento urbano y regional," *La situación demográfica de México 2014*, pp. 139–163, 2014.
[35] Y. Al-Darwish, H. Ayad, D. Taha, and D. Saadallah, "Predicting the future urban growth and it's impacts on the surrounding environment using urban simulation models: Case study of Ibb city – Yemen," *Alexandria Engineering Journal*, vol. 57, no. 4, pp. 2887–2895, Dec. 2018, doi: 10.1016/j.aej.2017.10.009.
[36] J. P. Schuster-Olbrich, O. Marquet, C. Miralles-Guasch, and L. Fuentes Arce, "Spatial patterns and drivers of urban expansion: An exploratory spatial analysis of the Metropolitan Region of Santiago, Chile, from 1997 to 2013," *Cities*, vol. 153, Oct. 2024, doi: 10.1016/j.cities.2024.105305.

[37] K. C. Seto, M. Fragkias, B. Güneralp, and M. K. Reilly, “A meta-analysis of global urban land expansion,” *PLoS One*, vol. 6, no. 8, 2011, doi: 10.1371/journal.pone.0023777.
[38] M. N. Rahman, K. S. Akter, and M. I. Faridatul, “Assessing the impact of urban expansion on carbon emission,” *Environmental and Sustainability Indicators*, vol. 23, Sep. 2024, doi: 10.1016/j.indic.2024.100416.
[39] R. G. Meyer Falcon *et al.*, “Programa Nacional de Ordenamiento Territorial y Desarrollo Urbano, 2021-2024,” 2019.
[40] R. G. Meyer Falcón *et al.*, “Estrategia Nacional de Ordenamiento Territorial,” 2021.
[41] H. Yang *et al.*, “Evaluating the effects of future urban expansion on ecosystem services in the Yangtze River Delta urban agglomeration under the shared socioeconomic pathways,” *Ecol Indic*, vol. 160, Mar. 2024, doi: 10.1016/j.ecolind.2024.111831.
[42] M. N. Rahman, K. S. Akter, and M. I. Faridatul, “Assessing the impact of urban expansion on carbon emission,” *Environmental and Sustainability Indicators*, vol. 23, Sep. 2024, doi: 10.1016/j.indic.2024.100416.
[43] C. Hyandye, “GIS and Logit Regression Model Applications in Land Use/Land Cover Change and Distribution in Usangu Catchment,” *American Journal of Remote Sensing*, vol. 3, no. 1, p. 6, 2015, doi: 10.11648/j.ajrs.20150301.12.
[44] C. Liping, S. Yujun, and S. Saeed, “Monitoring and predicting land use and land cover changes using remote sensing and GIS techniques—A case study of a hilly area, Jiangle, China,” *PLoS One*, vol. 13, no. 7, Jul. 2018, doi: 10.1371/journal.pone.0200493.
[45] E. Zumaya Ávila and J. Luis Hernández Suárez, “El Comercio Informal en el Centro Histórico de la Ciudad de Zacatecas, Zacatecas, México.,” 2020. [Online]. Available: http://hdl.handle.net/20.500.11763/turydes28comercio-informal-mexico
[46] C. Ünsalan and K. L. Boyer, *Multispectral Satellite Image Understanding*. Springer London, 2011. doi: 10.1007/978-0-85729-667-2.
[47] E. Omia *et al.*, “Remote Sensing in Field Crop Monitoring: A Comprehensive Review of Sensor Systems, Data Analyses and Recent Advances,” Jan. 01, 2023, *MDPI*. doi: 10.3390/rs15020354.
[48] U.S. Geological Survey, “Landsat—Earth Observation Satellites,” 2015, doi: 10.3133/fs20153081.
[49] M. Hemati, M. Hasanlou, M. Mahdianpari, and F. Mohammadimanesh, “A systematic review of landsat data for change detection applications: 50 years of monitoring the earth,” Aug. 01, 2021, *MDPI AG*. doi: 10.3390/rs13152869.
[50] (INEGI) Instituto Nacional de Estadística y Geografía, “Producción y publicación de la Geomediana Nacional a partir de las imágenes del Cubo de Datos Geoespaciales,” 2020. [Online]. Available: www.inegi.org.mx
[51] D. Roberts, N. Mueller, and A. McIntyre, “High-Dimensional Pixel Composites from Earth Observation Time Series,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6254–6264, Nov. 2017, doi: 10.1109/TGRS.2017.2723896.
[52] N. Bagwari, S. Kumar, and V. S. Verma, “A Comprehensive Review on Segmentation Techniques for Satellite Images,” Sep. 01, 2023, *Springer Science and Business Media B.V.* doi: 10.1007/s11831-023-09939-4.
[53] Y.-J. Zhang, “Handbook of Image Engineering,” 2021.
[54] L. Tao and N. Asoke K., “Image Segmentation: Principles, Techniques and Applications,” 2023.
[55] H. Mittal, A. C. Pandey, M. Saraswat, S. Kumar, R. Pal, and G. Modwel, “A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets,” *Multimed Tools Appl*, vol. 81, no. 24, pp. 35001–35026, Oct. 2022, doi: 10.1007/s11042-021-10594-9.
[56] K. G. Dhal, A. Das, B. Sasmal, S. Ray, R. Rai, and A. Garai, “Fuzzy C-Means for image segmentation: challenges and solutions,” *Multimed Tools Appl*, 2023, doi: 10.1007/s11042-023-16569-2.
[57] A. Taufik, S. S. S. Ahmad, and N. F. E. Khairuddin, “Classification of Landsat 8 satellite data using fuzzy cmeans,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jan. 2017, pp. 58–62. doi: 10.1145/3036290.3036330.

[58] T. Lei, X. Jia, Y. Zhang, L. He, H. Meng, and A. K. Nandi, "Significantly Fast and Robust Fuzzy C-Means Clustering Algorithm Based on Morphological Reconstruction and Membership Filtering," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 5, pp. 3027–3041, Oct. 2018, doi: 10.1109/TFUZZ.2018.2796074.

[59] W. Burger and M. J. Burge, "Digital Image Processing," 2016. [Online]. Available: http://www.springer.com/series/3191

[60] Q. Qin *et al.*, "Morphological Reconstruction-Based Image-Guided Fuzzy Clustering with a Novel Impact Factor," *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/6747371.

[61] Z. Wang, E. Wang, and Y. Zhu, "Image segmentation evaluation: a survey of methods," *Artif Intell Rev*, vol. 53, no. 8, pp. 5637–5674, Dec. 2020, doi: 10.1007/s10462-020-09830-9.

[62] P. A, S. Sebastian, G. Rohith, and L. S. Kumar, "Significant full reference image segmentation evaluation: a survey in remote sensing field," *Multimed Tools Appl*, vol. 81, no. 13, pp. 17959–17987, May 2022, doi: 10.1007/s11042-022-12769-4.

[63] D. Chicco, M. J. Warrens, and G. Jurman, "The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment," *IEEE Access*, vol. 9, pp. 78368–78381, 2021, doi: 10.1109/ACCESS.2021.3084050.

[64] J. Pont-Tuset and F. Marques, "Supervised Evaluation of Image Segmentation and Object Proposal Techniques," *IEEE Trans Pattern Anal Mach Intell*, vol. 38, no. 7, pp. 1465–1478, Jul. 2016, doi: 10.1109/TPAMI.2015.2481406.

[65] R. Shi, K. N. Ngan, and S. Li, *Jaccard Index Compensation for Object Segmentation Evaluation*. 2014.

[66] K. Dhanaraj and D. P. Angadi, "Land use land cover mapping and monitoring urban growth using remote sensing and GIS techniques in Mangaluru, India," *GeoJournal*, vol. 87, no. 2, pp. 1133–1159, Apr. 2022, doi: 10.1007/s10708-020-10302-4.

[67] L. David T., *Land Use Planning and Remote Sensing*. M. Nijhoff Publishers, 1984.

[68] D. Ramírez-Mejía, G. Cuevas, P. Meli, and E. Mendoza, "Land use and cover change scenarios in the Mesoamerican Biological Corridor-Chiapas, México," *Bot Sci*, vol. 95, no. 2, pp. 221–234, Apr. 2017, doi: 10.17129/botsci.838.

[69] A. Y. Yesuph and A. B. Dagnew, "Land use/cover spatiotemporal dynamics, driving forces and implications at the Beshillo catchment of the Blue Nile Basin, North Eastern Highlands of Ethiopia," *Environmental Systems Research*, vol. 8, no. 1, Dec. 2019, doi: 10.1186/s40068-019-0148-y.

[70] A. Butt, R. Shabbir, S. S. Ahmad, and N. Aziz, "Land use change mapping and analysis using Remote Sensing and GIS: A case study of Simly watershed, Islamabad, Pakistan," *Egyptian Journal of Remote Sensing and Space Science*, vol. 18, no. 2, pp. 251–259, Dec. 2015, doi: 10.1016/j.ejrs.2015.07.003.

[71] D. Camacho Olmedo María Teresa and García-Álvarez, "Basic and Multiple-Resolution Cross-Tabulation to Validate Land Use Cover Maps," in *Land Use Cover Datasets and Validation Tools: Validation Practices with QGIS*, M. T. and P. M. and M. J. F. García-Álvarez David and Camacho Olmedo, Ed., Cham: Springer International Publishing, 2022, pp. 99–125. doi: 10.1007/978-3-030-90998-7_7.

[72] D. and P. M. and D.-V. R. and C.-S. M. Á. Mas Jean-François and García-Álvarez, "Metrics Based on a Cross-Tabulation Matrix to Validate Land Use Cover Maps," in *Land Use Cover Datasets and Validation Tools: Validation Practices with QGIS*, M. T. and P. M. and M. J. F. García-Álvarez David and Camacho Olmedo, Ed., Cham: Springer International Publishing, 2022, pp. 127–151. doi: 10.1007/978-3-030-90998-7_8.

[73] P. Ren, S. Gan, X. Yuan, H. Zong, and X. Xie, "Spatial Expansion and Sprawl Quantitative Analysis of Mountain City Built-Up Area," 2013.

[74] Z.-L. Hu, P.-J. Du, and G. Da-Zhi, "Analysis of Urban Expansion and Driving Forces in Xuzhou City Based on Remote Sensing," 2007.

[75] H. Wang *et al.*, "Urban expansion patterns and their driving forces based on the center of gravity-GTWR model: A case study of the Beijing-Tianjin-Hebei urban agglomeration," *Journal of*

Geographical Sciences, vol. 30, no. 2, pp. 297–318, Feb. 2020, doi: 10.1007/s11442-020-1729-4.
[76] M. Tanveer, S. Hassan, and A. Bhaumik, "Academic policy regarding sustainability and artificial intelligence (Ai)," *Sustainability (Switzerland)*, vol. 12, no. 22, pp. 1–13, Nov. 2020, doi: 10.3390/su12229435.
[77] H. Sheikh, C. Prins, and E. Schrijvers, "Mission AI Research for Policy," 2023.
[78] R. T. Kneusel, *How AI Works: From Sorcery to Science*. 2023.
[79] M. Jordan, J. Kleinberg, and B. Schölkopf, "Pattern Recognition and Machine Learning."
[80] K. P. . Murphy, *Machine learning : a probabilistic perspective*. MIT Press, 2012.
[81] R. González, A. Barrientos, M. Toapanta, and J. Del Cerro, "Aplicación de las Máquinas de Soporte Vectorial (SVM) al diagnóstico clínico de la Enfermedad de Párkinson y el Temblor Esencial," *RIAI - Revista Iberoamericana de Automatica e Informatica Industrial*, vol. 14, no. 4, pp. 394–405, Oct. 2017, doi: 10.1016/j.riai.2017.07.005.
[82] Awad Mariette and Khanna Ragul, "Efficient Learning Machines," 2015.
[83] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif Intell Rev*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.
[84] J. J. Espinosa Zúñiga, "Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito," *Ingeniería Investigación y Tecnología*, vol. 21, no. 3, pp. 1–16, Jul. 2020, doi: 10.22201/fi.25940732e.2020.21.3.022.
[85] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, Second. 2017. [Online]. Available: http://www.springer.com/series/692
[86] A. V. Dorogush, V. Ershov, and A. G. Yandex, "CatBoost: gradient boosting with categorical features support." [Online]. Available: https://github.com/Microsoft/LightGBM
[87] S. Salehi, M. Arashpour, E. Mohammadi Golafshani, and J. Kodikara, "Prediction of rheological properties and ageing performance of recycled plastic modified bitumen using Machine learning models," *Constr Build Mater*, vol. 401, Oct. 2023, doi: 10.1016/j.conbuildmat.2023.132728.
[88] A. Panesar, *Machine learning and AI for healthcare: Big data for improved health outcomes*. Apress Media LLC, 2020. doi: 10.1007/978-1-4842-6537-6.
[89] M. Miró-Nicolau, A. Jaume-i-Capó, and G. Moyà-Alcover, "A comprehensive study on fidelity metrics for XAI," *Inf Process Manag*, vol. 62, no. 1, p. 103900, Jan. 2024, doi: 10.1016/j.ipm.2024.103900.
[90] C. O. Retzlaff *et al.*, "Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists," *Cogn Syst Res*, vol. 86, Aug. 2024, doi: 10.1016/j.cogsys.2024.101243.
[91] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," 2017. [Online]. Available: https://github.com/slundberg/shap
[92] R. Hernández Sampieri, S. Méndez Valencia, C. P. Mendoza Torres, and A. Cuevas Romo, *Fundamentos de Investigación*, Primera Edición. McGraw Hill/Intertamericana Editores, S.A. de C.V., 2017.
[93] Instituto Nacional de Estadistica y Geografia, "XII Censo General de Población y Vivienda 2000." Accessed: Mar. 21, 2024. [Online]. Available: https://www.inegi.org.mx/programas/ccpv/2000/
[94] Instituto Nacional de Estadistica y Geografia, "Censo de Población y Vivienda 2010." Accessed: Mar. 21, 2024. [Online]. Available: https://www.inegi.org.mx/programas/ccpv/2010/
[95] Instituto Nacional de Estadistica y Geografia, "Censo de Población y Vivienda 2020." Accessed: Mar. 21, 2024. [Online]. Available: https://www.inegi.org.mx/programas/ccpv/2020/
[96] INEGI, "Biblioteca digital de Mapas." Accessed: Sep. 05, 2024. [Online]. Available: https://www.inegi.org.mx/app/mapas/
[97] INEGI, "Conjunto de datos vectoriales de información topográfica F13B58 Zacatecas escala 1:50 000 serie II." Accessed: Sep. 05, 2024. [Online]. Available: https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825001060
[98] INEGI, "Carta topográfica. F13B58d." Accessed: Sep. 05, 2024. [Online]. Available: https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825724191
[99] INEGI, "Carta topográfica. F13B58e." Accessed: Sep. 05, 2024. [Online]. Available:

https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825724207
[100] INEGI, “Carta topográfica. F13B58f.” Accessed: Sep. 05, 2024. [Online]. Available: https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825724214
[101] INEGI, “Conjunto de Datos Vectoriales de Información Topográfica F13B58 Zacatecas escala 1:50 000, 2019.” Accessed: Sep. 05, 2024. [Online]. Available: https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=889463832034
[102] INEGI, “Conjunto de datos vectoriales de información topográfica F13B68 Guadalupe escala 1:50 000 serie II.” Accessed: Sep. 05, 2024. [Online]. Available: https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825003716
[103] INEGI, “Carta topográfica. F13B68a.” Accessed: Sep. 05, 2024. [Online]. Available: https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825724221
[104] INEGI, “Carta topográfica. F13B68b.” Accessed: Sep. 05, 2024. [Online]. Available: https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825724238
[105] INEGI, “Conjunto de Datos Vectoriales de Información Topográfica F13B68 Guadalupe escala 1:50 000, 2019.” Accessed: Sep. 05, 2024. [Online]. Available: https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=889463832980
[106] Abrams Michael and Crippen Robert, “ASTER GDEM V3 (ASTER Global DEM) User Guide,” 2019.
[107] “Diversidad Zacatecas.” Accessed: May 10, 2024. [Online]. Available: https://cuentame.inegi.org.mx/monografias/informacion/zac/poblacion/diversidad.aspx
[108] M. S. Rana and S. Sarkar, “Prediction of urban expansion by using land cover change detection approach,” *Heliyon*, vol. 7, no. 11, Nov. 2021, doi: 10.1016/j.heliyon.2021.e08437.
[109] B. F. Frimpong and F. Molkenthin, “Tracking urban expansion using random forests for the classification of landsat imagery (1986–2015) and predicting urban/built-up areas for 2025: A study of the kumasi metropolis, ghana,” *Land (Basel)*, vol. 10, no. 1, pp. 1–21, Jan. 2021, doi: 10.3390/land10010044.

Anexos

Durante el XVI Congreso Mexicano de Inteligencia Artificial - COMIA 2024, llevado a cabo del 3 al 7 de junio de 2024 en Celaya, Guanajuato, México. Se presento en modalidad de ponencia el artículo titulado “Predicción de Expansión Urbana de las Ciudades Zacatecas-Guadalupe Usando Maquinas de Soporte Vectorial.”

COMIA 2024

TECNOLÓGICO NACIONAL DE MÉXICO

La Sociedad Mexicana de Inteligencia Artificial (SMIA) y el Tecnológico Nacional de México – Instituto Tecnológico de Celaya

OTORGAN ESTE CERTIFICADO A

A. Carmina Llamas-Valenzuela, José I De la Rosa, G Moreno-Chávez, Efrén Gonzalez-Ramirez, Jesús Villa, Jose Celaya-Padilla

por la presentación del artículo titulado

Predicción de Expansión Urbana de las Ciudades Zacatecas-Guadalupe Usando Maquinas de Soporte Vectorial

en el XVI Congreso Mexicano de Inteligencia Artificial - COMIA 2024

Celaya, Guanajuato, México, del 3 al 7 de junio de 2024

Sociedad Mexicana de Inteligencia Artificial www.smia.org.mx RFC SMI871231FB7

Dra. Lourdes Martínez Villaseñor
Presidenta SMIA

Dra. Iris I. Méndez Gurrola
Presidenta Comité de Programa

Dr. Roberto A. Vázquez Espinosa
Presidente Comité de Programa

Dr. Alejandro I. Barranco Gutiérrez
Comité Local COMIA

Predicción de Crecimiento Urbano de la Ciudad Zacatecas-Guadalupe Usando Maquinas de Soporte Vectorial

Abstract. Las ciudades son espacios físicos en los cuales se establece la población. En México, tres de cada cuatro personas viven en una ciudad. El crecimiento rápido y sin planeación trae consigo consecuencias indeseables para el desarrollo social y económico. El objetivo de este estudio es modelar y predecir la expansión urbana en la zona metropolitana Zacatecas-Guadalupe usando máquinas de soporte vectorial. Para lograr este objetivo se utilizaron mapas de uso de suelo y cobertura terrestre correspondientes al periodo 2000 a 2020, así como la inclusión de variables socioeconómicas, topográficas y atributos culturales. Se desarrolló un modelo de SVM con una exactitud de entrenamiento de un 94 %, exactitud de validación 93 % y F1-Score de un 84 %. En los resultados obtenidos, observar que la cercanía a áreas ya urbanizadas y el tipo de uso de suelo tienen una alta influencia en la urbanización. Adicionalmente, la pendiente del terreno tiene muy poca influencia en las decisiones de urbanización. Comparado con otros estudios, este incorpora variables culturales y e integra los valores SHAP, con el objetivo de conocer la influencia de las variables en el modelo final.

Keywords: expansión urbana, máquinas de soporte vectorial, predicción.

1 Introducción

Las ciudades son espacios físicos en los cuales se establece la población. En México, tres de cada cuatro personas viven en una ciudad [1] Este incremento ocasiona el crecimiento urbano. El crecimiento es un proceso socioeconómico complejo que involucra la transformación de espacios destinados a actividades primarias a actividades terciarias [2].

El crecimiento rápido y sin planeación trae consigo consecuencias indeseables para el desarrollo social y económico. Algunas de las consecuencias son: el daño ecológico [3], insuficientes viviendas para la demanda, sistemas de transporte deficientes, aumento en la segregación social [4] y el incremento del tráfico [5]. Estas consecuencias se pueden aminorar, realizando una planeación a largo plazo, apoyada de diferentes instrumentos [6].

A lo largo de los años, expertos en el área de planeación urbana han realizado diferentes métodos para el modelado del crecimiento urbano. Algunos de estos modelos se apoyan de datos obtenidos de la percepción remota, en los cuales se utilizan imágenes satelitales. Estas se emplean para conocer el uso de suelo y cobertura de la tierra. Adicionalmente, estos modelos se apoyan ampliamente en el uso de Sistemas de Información Geográfica (GIS, por sus siglas en inglés) [7]. Es importante aclarar que estos esfuerzos de prever el crecimiento urbano no sustituyen el conocimiento de los

16 Congreso Mexicano de Inteligencia Artificial – COMIA 2024

Cuernavaca, Morelos, a 26 de junio de 2024

A. Carmina Llamas-Valenzuela

Universidad Autónoma de Zacatecas

Asunto: Invitación a publicar en la revista IJCOPI

A nombre del comité del programa del Congreso Mexicano de Inteligencia Artificial COMIA 2024. Queremos agradecer su participación y notificarte que su artículo "*Predicción de Crecimiento Urbano de la Ciudad Zacatecas-Guadalupe Usando Maquinas de Soporte Vectorial*" ha sido seleccionado para ser publicado en un número especial en la revista International Journal of Combinatorial Optimization Problems and Informatics (IJCOPI), indexada en JCR ESCI y reconocida como revista científica del CONAHCYT, https://ijcopi.org/ojs/about .

Si es de tu interés la publicación, te pedimos nos lo hagas saber a la brevedad y nos envíes una versión extendida del artículo conforme al formato de la revista IJCOPI para el **22 de julio del 2024**.

La calidad de los artículos que se presentaron en el evento es apreciada razón por la cual la Sociedad Mexicana de Inteligencia Artificial (SMIA) desea editar un número especial con los mejores artículos del COMIA 2024.

Quedamos en espera de tu amable respuesta.

Atentamente

Dra. Iris I. Méndez Gurrola
Presidenta Comité de Programa
COMIA-SMIA

Dr. Roberto A. Vázquez Espinosa
Presidente Comité de Programa
COMIA-SMIA

Dr. Gustavo Arroyo Figueroa
Editor invitado IJCOPI

International Journal of Combinatorial Optimization Problems and Informatics, 10(2), May-Aug 2019, 17-25. ISSN: 2007-1558.

www.editada.org

Predicting Urban Expansion in Zacatecas-Guadalupe cities: A Support Vector Machine Approach Enhanced by SHAP Values

A. Carmina Llamas-Valenzuela[,1], José I. de la Rosa[*,1], G. Moreno-Chávez[1], Efrén Gonzales-Ramírez[1], Jesús Villa[1], and José M. Celaya-Padilla[1]*

[1] Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Zacatecas, México
allamas@uaz.edu.mx, ismaelrv@ieee.org

Abstract. Cities are centers that generate employment, innovation, and improve the quality of life, basic services, and housing. Rapid and unplanned expansion brings with it undesirable consequences for social and economic development. In Mexico, three out of four people live in a city. For this reason. the aim in this paper is to model and predict urban expansion in Zacatecas-Guadalupe cities (Mexico) using support vector machines, and thus carry out a better planning. In order to achieve this objective, land use and land cover maps corresponding to the period 2000–2020 were used, as well as the inclusion of socioeconomic, topographic, and cultural attribute variables. A soft SVM model was developed with a training accuracy of 92.4%, a validation accuracy of 93% and an F1-Score of 86.3%. In the obtained results, it can be observed that the proximity to already urbanized areas and the type of land use have a high influence on urbanization

Keywords: Urban expansion, Support vector machine, Urban expansion prediction, SHAP values, Satellite images,

Article Info
Received Jul 20, 2024
Accepted Dec, 2024

1 Introduction

Cities are centers that generate employment, innovation, and improve the quality of life, basic services, and housing (United et al., 2018). These opportunities generate a urban expansion, this means that this expansion changes the landscape around the city (Roy, 2021). This is called land use land cover change, where the natural cover adapts or changes to something useful for humans. Nonetheless, this transformation may possess either a favorable or unfavorable impact on the environment or society (Amri et al., 2020). Environmental damage (Deng et al., 2018), insufficient housing for demand, deficient transportation systems, increased social segregation (Rangel et al., 2022) and increased traffic (Kim & Kim, 2022) are some of the negative aspects.

In order to address these unfavorable impacts, it is important to have long-term planning supported by various instruments (González-Madrigal et al., 2020). Numerous approaches to modeling expansion have been devised, some of which rely on data obtained from remote sensing, utilizing satellite imagery to ascertain land use and land cover. Moreover, these models are extensively supported by Geographic Information Systems (GIS) (Nugroho & Al-Sanjary, 2018). Several mathematical models have been utilized for urban expansion, including Cellular Automata (CA) (Kara & Doratlı, 2021; Liang et al., 2018; Rangel et al., 2022; S. W. Wang et al., 2021), a model commonly employed to simulate urban expansion (H. Wang et al., 2021). The CA is a simplification of reality, constructed with key elements (Rangel et al., 2022). However, it has been assumed that urban areas are spatially homogeneous, making it challenging to accurately represent individual decisions and socioeconomic interactions.

Thus, in order to address this limitation, researchers have employed not only statistical methods, such as different types of regressions (Hyandye, 2015; Rangel et al., 2022), but also various machine learning techniques, such as XGBoost-SHAP (Kim et al., 2023; Kim & Kim, 2022), support vector machines (SVM) (Karimi et al., 2019; Mirbagheri & Alimohammadi, 2018), decision trees (Karimi et al., 2021), random forests (Frimpong & Molkenthin, 2021) and others. However, these studies were conducted outside of Mexico.

Durante el 2024 IEEE Generation, Transmission & Distribution LA/International Autumn Metting on Power, Electronics and Computing, llevado a cabo del 11 al 13 de junio de 2024 en Ixtapa, Zihuatanejo, México. Se presento en modalidad de ponencia el artículo titulado "Predicting Urban Expansion in the cities of Zacatecas and Guadalupe, Mexico: A Comparative Analysis."

4/71

IEEE

IEEE Centro Occidente Section

IEEE PES Power & Energy Society®

THE ORGANIZING COMMITTEE OF THE 2024 IEEE GENERATION, TRANSMISSION & DISTRIBUTION LA/ INTERNATIONAL AUTUMN MEETING ON POWER, ELECTRONICS AND COMPUTING

GT&D LA 2024 ROPEC

GRANTS THIS

CERTIFICATE

TO

A. Carmina Llamas-Valenzuela, José I de la Rosa, Gamaliel Moreno-Chávez, Efrén Gonzales-Ramirez, Jesús Villa and Daniel Alaniz-Lumbreras

for the presentation of the paper

Predicting Urban Expansion in the Cities of Zacatecas and Guadalupe, Mexico: A Comparative Analysis

Dr. Claudio Rubén Fuerte Esquivel
GENERAL CHAIR
IEEE GTDLA/ROPEC 2024

IEEE SECCION CENTRO OCCIDENTE

Dr. José Ortiz Béjar
CHAIR
IEEE CENTRO OCCIDENTE SECTION

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC
IXTAPA, ZIHUATANEJO, GRO. MÉXICO; NOVEMBER 11-13, 2024

Predicting Urban Expansion in the Cities of Zacatecas and Guadalupe, Mexico: A Comparative Analysis

A. Carmina Llamas-Valenzuela[1], José I. de la Rosa[2],
G. Moreno-Chávez[3], Efrén Gonzales-Ramírez[4], Jesús Villa[5], Daniel Alaniz-Lumbreras[6]
Unidad Academica De Ingeniería Eléctrica
Universidad Autónoma de Zacatecas
Zacatecas, México
[[1]allamas, [3]gamalielmch, [4]gonzalezefren, [5]jvillah, [6]dalaniz]@uaz.edu.mx,
[2]ismaelrv@ieee.org

***Abstract*—The present paper proposes a study which aims the modeling and prediction of the urban expansion of the Zacatecas-Guadalupe cities (Mexico), using techniques such as Support Vector Machines (SVM), Random forests (RF), and the catBoost algorithm. According to the implementation of the previous techniques, the land use and the land cover maps corresponding to the period 2000–2020 were used, as well as the inclusion of socioeconomic, topographic and cultural attribute variables. The obtained results reveals that the proposed models make reasonable predictions of the urban expansion in the year 2030. Among the used techniques, the best algorithm for this problem was Soft SVM.**

Keywords— Urban Expansion Prediction, Support Vector Machines; Random forests; catBoost Algorithm.

I. INTRODUCTION

Cities face different challenges and opportunities, depending on their size, time, or location. They confront issues such as the coverage of basic public services, security, employment, quality of life, and equity, among others [1]. Rapid and unplanned urban expansion can amplify these problems. However, these consequences can be mitigated through long-term planning supported by different instruments [2].

To enhance urban expansion benefits, models using satellite data and Geographic Information Systems (GIS) [3] have been developed to analyze land use changes. It's important to note that these predictive efforts complement, not replace, the expertise of city planners, serving as decision-making aids.

Widely used models for urban expansion are cellular automata [4], [5], [6], [7] and agent-based models (ABM) [8]. Have proven effective in various studies, but they often assume urban areas are spatially uniform. This limitation can reduce their accuracy in representing individual behaviors and complex socioeconomic interactions [9].

To overcome these limitations, machine learning techniques like XGBoost-SHAP [10], support vector machines (SVM) [11], [12], decision trees [13], and random forests [14], among others. However, most of these studies have been conducted outside of Mexico, despite the pressing need for solutions in a country where, by 2020, nearly one hundred million people were living in unplanned urban settlements [4]. This need has motivated the proposed study.

The study presented in this paper aims to modeling and predict the urban expansion of the Zacatecas-Guadalupe cities (Mexico), exploring techniques such as Support Vector Machines, Random forests, and the catBoost algorithm and drive a comprehensive analysis of the results. By analyzing land use and the land cover maps corresponding to the period 2000 to 2020, as well as the inclusion of socioeconomic, topographic and cultural attribute variables. The results reveals that the proposed models make reasonable predictions of the urban expansion in the year 2030.

The rest of the paper is structured as follows: Section II outlines the materials and methods used, Section III details the experimental setup, Section IV presents the results obtained from the three techniques, Section V offers a brief discussion, and Section VI concludes with key remarks.

II. MATERIALS AND METHODS

A. Area under study

The cities of Zacatecas-Guadalupe have been chosen as the study area. These cities are situated within the Zacatecas-Guadalupe metropolitan region. It is one of the 48 metropolitan areas of Mexico. This makes it a metropolitan region that experiences rapid urban expansion, cultural diversity, and a significant demand for natural resources and services [15]. It is essential to gather and analyze comprehensive data to understand these dynamics.

B. Data

Following the definition of the area under study, extensive data collection was initiated to capture the development of the region. Raster, vector data set of topographic maps, and census results data were collected from the National Institute of Statistics, Geography, and Informatics (INEGI) for the period from 2000 to 2020. Furthermore, Landsat Geomedian images and Digital Elevation Model (DEM) were collected in the years 2000, 2010 and 2020. All the data used in this paper were projected in Universal Transverse Mercator (UTM) system to ensure consistency in analysis.

979-8-3315-3052-5/24/$31.00 ©2024 IEEE