Caracterización de los municipios de Colombia según el potencial de crecimiento del ingreso por impuesto predial utilizando Machine Learning

Daniel Delgado Vargas

Facultad de Ingeniería

Maestría en Analítica Aplicada

Director del Proyecto:

Andrés Felipe Cardona Ortegón

2023

Página de aceptación:

A continuación, se presentan las firmas de aceptación del proyecto titulado

“Caracterización de los municipios de Colombia según el potencial de crecimiento del ingreso por impuesto predial utilizando Machine Learning”

_____

Andrés Felipe Cardona Ortegón

Director del Proyecto

_____

Fidel Andrés Olarte Bustos

Jurado

_____

William Javier Guerrero Rueda

Jurado

_____

Manuel Alfredo Figueredo Medina

Jurado

Chía, 31 de Julio del 2023

Tabla de contenido:

Tabla de ilustraciones:

Tabla de Ecuaciones:

Lista de Tablas:

Caracterización de los municipios de Colombia según el potencial de crecimiento del ingreso por impuesto predial utilizando Machine Learning

1. Resumen

1.1. Abstract

This study explores the potential growth of towns and cities in Colombia. It uses Machine Learning to generate a recommendation to identify places with accelerated growth in the medium term, and to understand the characteristics that make these destinations interesting. These characteristics are difficult to identify due to the lack of integrated information.

The objective is to characterize towns based on annual growth of property tax using CRISP-DM methodology. To accomplish this, we defined a growth potential factor as a target variables and identified the variables that account for this behavior. In the analysis was used information from all towns in Colombia from 2008 to 2020. We included 1971 variables from official sources, such as DANE, DNP, Ministry of Education, and Health, among others

A novel ensemble of variable selection techniques is used, including PCA, XGBoost, Random Forest, RFE, and SelectKBest, to finally selects 600 variables that allowed efficient training for classification and regression models and identify the best performing model to create a growth potential factor. This factor was used to cluster the towns in 5 categories and identifying how variables related to health, education, population, security, poverty, agriculture, justice, and economy explain the success of certain municipalities where opportunities are worth exploring.

1.2. Resumen

El presente trabajo explora el potencial de crecimiento de los municipios de Colombia, utilizando aprendizaje automático para generar una recomendación que permita identificar lugares con crecimiento acelerado en el mediano plazo y entender las características que hacen estos destinos interesantes, ya que es complejo reconocerlos por la falta de información integrada.

Como objetivo se buscó realizar la caracterización de los municipios mediante metodología CRISP-DM, y para lograrlo, se definió un factor de potencial de crecimiento y se identificaron las variables que explican este comportamiento. Se dispuso de información de todos los municipios de Colombia de 2008 al 2020, teniendo como variable objetivo el crecimiento anual de ingresos por impuesto predial, complementado con 1971 variables provenientes de fuentes oficiales como el DANE, DNP, Ministerio de Educación y Salud entre otros.

Se utilizó un novedoso ensamble de técnicas de selección de variables donde se ejecutaron algoritmos de PCA, XGBOOST, Random Forest, RFE y SelectKBest, para seleccionar finalmente 600 variables que permitieron entrenar modelos de clasificación y regresión, identificando el de mejor desempeño para lograr crear un factor de potencial de crecimiento, que generó la posibilidad de agrupar los municipios en 5 categorías e identificar como variables de salud, educación, población, seguridad, pobreza, agricultura, justicia y economía explican el éxito de unos municipios donde es conveniente buscar oportunidades de inversión.

1.3. Resumen gráfico

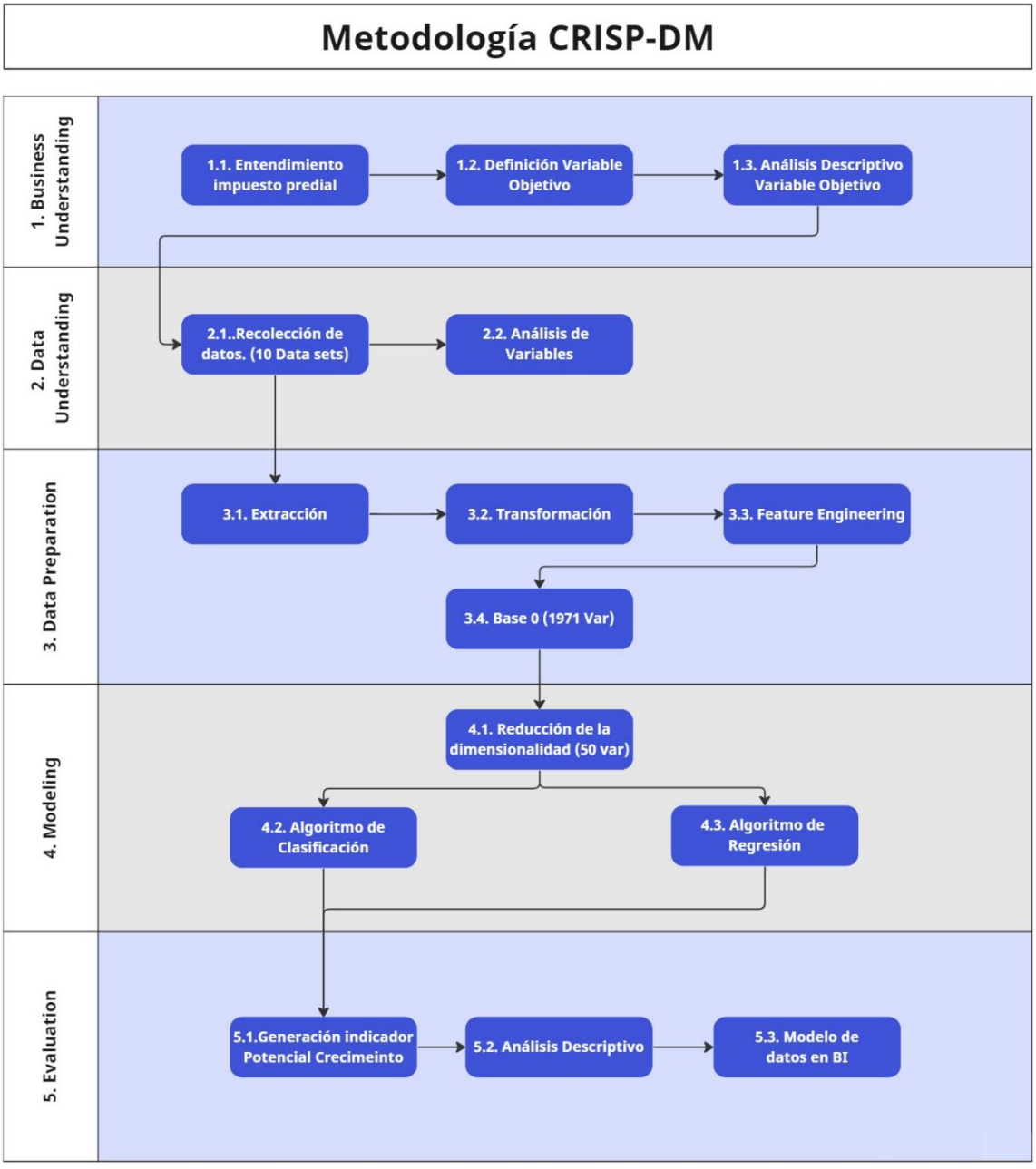

Ilustración 1. Resumen General del proyecto.

2. Introducción

El presente proyecto se enfoca en generar una solución que ayude a los inversionistas a la hora de elegir municipios en Colombia que tengan potencial de crecimiento, esto basado en el incremento de ingreso de impuesto predial y las variables relacionadas que explican su comportamiento. Si se desea explorar opciones que no estén en el contexto cercano del inversionista, elegir una opción se hace complejo al no tener información predictiva del crecimiento para la mayoría de los municipios.

La información histórica inmobiliaria y los registros públicos de avaluó catastral están protegidos por ley, dificultando aún más el análisis para inversionistas medianos y pequeños. A pesar de que existen estudios locales que exploran zonas y variables específicas, son pocos los que ofrecen una visión completa a nivel nacional. Además, la información abierta proporcionada por el gobierno se limita en su mayoría a la visualización, sin brindar recomendaciones específicas por municipios.

En este contexto, el objetivo principal de este proyecto es desarrollar una herramienta basada en aprendizaje automático que permitan visualizar el potencial de crecimiento de los municipios en Colombia. Los objetivos específicos del proyecto se centran en generar una medida de crecimiento futuro y caracterizar los municipios de Colombia utilizando más de 1000 variables demográficas, socioeconómicas y urbanísticas del período comprendido entre 2008 y 2020 tomadas de datos abiertos de diferentes entidades oficiales como son el DANE, DNP, Ministerio de Agricultura, Policía nacional, Contraloría general de la nación, Instituto Agustin Codazzi, Ministerio de Salud, Ministerio de turismo, Min TIC .

Esta información se procesa usando técnicas de regresión y clasificación, con las que se calcula el potencial futuro de crecimiento de los municipios de Colombia por ingresos de impuesto predial. Con esta medida se clasifican en 5 categorías y mediante técnicas de selección de variables, se identifican aquellas que afectan y se encuentran relacionadas con este crecimiento.

3. Justificación

3.1. Problema

Colombia es un país en vía de desarrollo, donde según estudios del DANE el 33% de la población para 2021 se encontraba en pobreza. Enfrenta grandes retos en temas de seguridad por delincuencia común, conflicto armado y narcotráfico. Al tener tres cordilleras, y estas ubicado en el trópico tiene barreras naturales que aíslan algunas zonas, como por ejemplo el Amazonas y el Vaupés. 5 ciudades concentran más de 15 millones de habitantes, lo que genera dinámicas comerciales cerca de las ciudades y en los lugares más seguros y accesibles. Pero hay zonas que, por nueva infraestructura, cambios en seguridad dados por ejemplo por el proceso de paz con las Farc cambiaron las condiciones de crecimiento. Con este dinamismo es importante utilizar datos para poder identificar que zonas están empezando su desarrollo, e identificar un buen momento para invertir.

Dadas las condiciones del contexto de Colombia y que la información disponible esta dispersa, es complejo el proceso de toma de decisión contemplando diferentes zonas del país. Poder estimar un potencial de crecimiento y analizar las variables que lo generan consume recursos que muchos no tienen a su disposición. Existen estudios locales que exploran zonas y variables específicas, pero pocos dan una visual completa del país. La información inmobiliaria histórica es de difícil acceso y los registros públicos de avaluó catastral por ley están protegidos. El gobierno habilita información abierta, que en la mayoría de los casos solo llega a la capa de visualización, pero no generan recomendaciones específicas por municipios, lo que dificulta poder tomar una decisión para seleccionar un lugar y realizar una inversión especifica, ya que los estudios se centran en zonas de evidente desarrollo.

Por lo anterior, surge la necesidad de abordar esta problemática y con la información disponible crear herramientas que permitan ver el potencial de crecimiento de los municipios en Colombia. Esto permitirá a los inversionistas y actores del mercado inmobiliario identificar oportunidades en el momento adecuado para invertir, maximizando el beneficio y reduciendo los riesgos.

3.2. Pregunta de investigación

El proyecto se centra en contestar la siguiente pregunta de investigación.

- ¿Cuáles municipios de Colombia tienen potencial de incrementar de forma acelerada la recaudación de impuesto predial y que variables generan este efecto?

4. Objetivos

Caracterizar los municipios de Colombia en función del potencial de crecimiento de ingresos generados por el impuesto predial, a través de un análisis cuantitativo y cualitativo de datos demográficos, socioeconómicos y urbanísticos del 2008 al 2020, utilizando aprendizaje automático para poder identificar oportunidades de inversión en municipios con alto desarrollo.

Los objetivos específicos del proyecto son los siguientes:

- Caracterizar y agrupar los municipios de Colombia con base en el cálculo del potencial de crecimiento de sus ingresos de impuesto predial, utilizando análisis de regresión y métodos de clasificación.
- Caracterizar los municipios de Colombia a partir de la identificación de las variables que tienen una mayor relación con el incremento de los ingresos por impuesto predial, utilizando datos de 2008 a 2020.
- Identificar la mejor configuración de los conjuntos de datos para obtener los mejores resultados en la predicción y optimizar el performance de los modelos de aprendizaje automático XGBOOST y Random Forest para obtener el mejor factor de potencial futuro.

5. Marco conceptual y Metodológico

5.1. Variable objetivo

El presente proyecto plantea clasificar los municipios de Colombia según la tasa de recaudación de impuesto predial, utilizando información desde el 2008 al 2020, y más de 1900 variables recolectadas de diferentes fuentes oficiales del gobierno colombiano. Uno de los retos iniciales es poder encontrar el valor de los predios en el tiempo. La entidad en Colombia que rige las reglas para el cálculo de avalúos catastral es el Instituto Geográfico Agustín Codazzi (IGAC). En los datos abiertos el IGAC comparte información descriptiva de todos los predios registrados del país, con ubicación área y limites, pero el avalúo es confidencial y no está permitido compartirlo como lo determina la ley colombiana[1]. La información de avalúo comercial o precios de venta en diferentes plataformas comerciales es de acceso privado, se podría realizar obtener mediante recolección de datos web (en ingles

1 https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=116241

web scraping), lo que requiere un desarrollo para leer información de diferentes portales y capturar información por varios años.

Al no contar con información primaria disponible y para cumplir los objetivos, se necesita información de referencia del crecimiento del valor de los predios por municipio. Se validaron otras alternativas de información disponible que pudieran evidenciar la tasa de crecimiento de los municipios. Según (Andrews, 2010) una de las variables con alta relación con el incremento del valor es el crecimiento de la población, pero los datos disponibles son del censo, que solo se ha hecho en 2005 y 2018. En (Égert & Mihaljek, 2007) se presenta el PIB como variable explicativa del crecimiento del valor, pero al igual que el crecimiento de la población a nivel municipio no se encontró información suficiente disponible.

La variable de referencia debe mostrar el incremento del valor de las viviendas en el tiempo, pero como el objetivo no es calcular un avalúo sino encontrar comportamientos de crecimiento agregado por municipio, se eligió el monto de ingresos por concepto de impuesto prediale de cada municipio, ya que hay información detallada por año, tiene relación con crecimiento de valor, población y una alta correlación con los avalúos, como se puede evidenciar en la fórmula de cálculo del avalúo **(ej. Alcaldía de Bogotá[2] [3])** :

$$Impuesto\ predial = \frac{Avaluo\ catastral\ \times\ Factor\ tarifario}{1000}$$

Ecuación 1. Calculo impuesto predial

*"**Avaluó Catastral:** Es el valor de un predio, resultante de un ejercicio técnico que, en ningún caso, podrá ser inferior al 60% del valor comercial o superar el valor de este último. Para su determinación no será necesario calcular de manera separada el valor del suelo y el de la construcción."*[4]

Adicionalmente existe una relación con el avalúo comercial, como especifica el IGAC en el documento de definiciones del marco catastral en Colombia. Dado lo anterior y que cada municipio tiene factores tarifarios diferentes, el analisis comparativo de los ingresos prediales se debe manejar en una unidad de crecimiento y no con valores absolutos en pesos. El presente trabajo no pretende calcular precios ya que las relaciones de avalúos catastrar y predial en la práctica pueden variar, se busca identificar si un municipio está creciendo y si es probable que continue creciendo.

2 https://www.catastrobogota.gov.co/pregunta/que-relacion-tiene-el-avaluo-catastral-con-el-impuesto-predial

3 https://www.ambitojuridico.com/noticias/tributario/notariado-y-registro/como-se-calcula-el-impuesto-predial-de-bogota-para-el-2017

4 https://geoportal.igac.gov.co/sites/geoportal.igac.gov.co/files/geoportal/plantilla_objetos_para_publicacion_v_30_11_2022.pdf

Por lo anterior, la variable objetivo que tendrá el modelo es la tasa de crecimiento respecto a un punto de referencia (año 2008) y no el valor absoluto en pesos, adicionalmente, esta unidad de crecimiento debe ser calculada en un horizonte de tiempo, el cual, en esta investigación, es de tres años en el futuro.

$$Y\_Crecimeinto\_Predial_{\text{año } i} = \frac{Ingreso\_predial_{\text{año } i} - Ingreso\_predial_{\text{año } 2008}}{Ingreso\_predial_{\text{año } 2008}}$$

Ecuación 2. Ecuación variable objetivo

5.2. Elección algoritmos y metodología

Posterior a la definición de la variable objetivo a predecir, se realizó la investigación de diferentes fuentes que utilizan modelos de regresión de aprendizaje automático y modelos de agrupamiento por análisis no supervisado, que permiten identificar las variables más relevantes y que hayan sido utilizados en análisis para predecir variables económicas dentro de casos de uso de finca raíz o predicción de precios.

Se han desarrollado diferentes trabajos alrededor del mundo con el objetivo de estimar la valorización de predios urbanos y rurales utilizando aprendizaje automático (ML: Machine Learning), como Austing en USA (Alvarez et al., 2022) y Corea del Sur (Lee, 2021). Se evidencia como usar este tipo de técnicas genera una mejora en precisión, comparado con métodos tradicionales de cálculo de avalúos, por tanto, la estrategia definida en el proyecto es la construcción de un modelo de aprendizaje automático.

Para construir el modelo se siguió la metodología CRISP-DM®, como sugiere (Ma et al., 2020) en su proyecto de análisis de factores para valorizar los terrenos. En este caso solo se utilizaron las primeras 5 etapas de la metodología ya que la 6ta etapa es la implementación, y el presente proyecto finaliza con el resultado del modelo como se evidencia en la Ilustración 2.

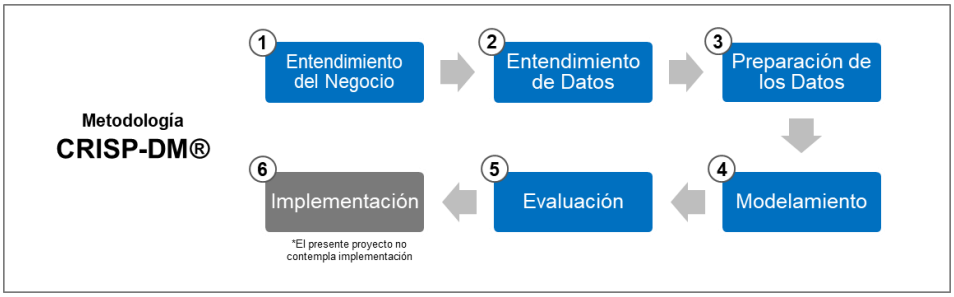

Ilustración 2. Etapas de la Metodología CRISP-DM®

La exactitud de los modelos depende en gran medida de la información disponible. Un análisis de terrenos en Nueva York (Ma et al., 2020) tiene disponibles cientos de variables, por lo que es muy importante recopilar la mayor cantidad de variables para mejorar la precisión del modelo y la descripción del resultado al asociarlo a diferentes variables. Para poder utilizar los modelos de ML se realizó una investigación de diferentes fuentes de información por Municipios de Colombia desde el 2008 al 2020. En el anexo A se presenta la lista de variables y fuentes de los datos.

Varios autores han abordado el análisis de regresión para predecir costos prediales. En el análisis presentado por (Bilgilioğlu & Yılmaz, 2023) se aplicaron cinco modelos de aprendizaje automático (Red Neuronal (ANN), Detección de Chi-cuadrado (CHAID), Support Vector Machine (SVM), Random Forest (RF)) para la tasación masiva de bienes raíces en 41,074 parcelas zonificadas ubicadas en el área municipal adyacente de Aksaray. En este contexto, se determinaron un total de 68 factores que afectan el valor del terreno. Se obtuvieron los valores de venta de 1982 parcelas. Las mediciones de error se realizaron utilizando las siguientes métricas MSE, Error Cuadrático Medio, medida que cuantifica el promedio de los errores al cuadrado; R2, Coeficiente de Determinación, bondad de ajuste del modelo, Un valor de R2 cercano a 1 indica que el modelo se ajusta muy bien a los datos; MAPE, Error Porcentual Absoluto Medio, mide el error porcentual promedio entre las predicciones. El autor comenta que los resultados obtenidos con los modelos de aprendizaje automático pueden ser utilizados por el sector privado, los gobiernos nacionales y locales en proyectos relacionados con la tierra, inversiones y recaudación de impuestos. En el presente proyecto se utilizaron los mismos indicadores para comparar la efectividad de los modelos de aprendizaje automático.

Para generar la clasificación se utilizó un ensamble de 2 modelos, uno de regresión que calcula el valor futuro de incremento del ingreso de impuesto predial y un segundo modelo de clasificación para determinar la similitud que tiene un municipio que crece con uno que no lo hace aún, pero que está mejorando en otras variables.

En cuanto a los modelos de regresión, luego de revisar varias fuentes, se evidencia que el performance de los algoritmos no depende del área en que se esté trabajando sino específicamente del set de datos y la relación que estos tienen, como se observa en el trabajo de (Bentéjac, Csörgő, & Martínez-Muñoz, 2019) donde se compara el rendimiento del modelo XGBoost y Random Forest en diferentes sets de datos con otros algoritmos. En este trabajo, se concluye que cada algoritmo tiene un resultado diferente según el set de datos.

En otro estudio realizado por (Chen et al., 2021) se planteó utilizar los siguientes modelos de aprendizaje automático, ARIMA, modelo estadístico utilizado para analizar y predecir series temporales; SVR, Support Vector Regression, versión de

regresión del modelo support vector machine; Prophet, herramienta de código abierto desarrollada por Facebook para el análisis de series temporales; XGBoost, algoritmo de aumento de gradiente que se utiliza para mejorar el rendimiento de modelos de aprendizaje; y LSTM, un tipo de red neuronal que se utiliza en el procesamiento de secuencias de datos. Estos modelos se utilizaron para predecir precios de productos agrícolas utilizando una gran cantidad de variables y datos históricos. Incluyeron en el análisis la forma tradicional de resolver problemas de predicción basados en métodos estadísticos lineales (como el modelo ARIMA) para ver el valor agregado que generan estas nuevas técnicas. En este estudio se ve como a medida que hay más variables, toma más relevancia utilizar un modelo de ML en lugar de métodos tradicionales como Arima.

Por otro lado, en un estudio realizado por (Zhu & He, 2022) aunque está enfocado a datos de Amazon, se utiliza para predecir precio y se comparan tres modelos: ARIMA, XGBoost y LSTM, usando como indicador de exactitud el error cuadrático medio (MSE). Los resultados indican que LSTM es el mejor modelo para predecir el precio, pero cuando se tiene un número limitado de registros, y gran cantidad de variables (como el presente proyecto) la red neuronal requiere muchos ajustes y es mejor utilizar modelos como XGBoost **y Random Forest** como indica (Ravi & Larochelle, 2017).

Para el modelo de clasificación (Li & Chen, 2020) realizaron la comparación de diferentes algoritmos para predecir la puntuación crediticia basado en una gran cantidad de variables y los algoritmos con un mejor performance fueron RF y XGBoost. Por tanto, en el presente trabajo fueron utilizados tanto para clasificación como para regresión los modelos de RF y XGBoost.

5.3. Conceptos de aprendizaje automático

El Aprendizaje Automático o Machine Learning en inglés, es una rama de la inteligencia artificial que busca replicar una de las características de la inteligencia, poder aprender. Esto lo hace cambiando el paradigma de la programación donde se genera una recomendación basada en reglas o restricciones, en aprendizaje automático se ingresan las respuestas, el modelo genera automáticamente las restricciones y a medida que cambian la respuesta puede automáticamente ajustarse para predecir con mayor exactitud.

"Machine Learning es el estudio de algoritmos de computación que mejoran automáticamente su rendimiento gracias a la experiencia."

Tom Mitchell autor de "Machine Learning" de 1997

5.3.1. Reducción de la dimensionalidad

Posterior a la selección de los modelos a utilizar y dado a que los conjuntos de datos de entrenamiento tienen más de mil variables, para poder optimizar el rendimiento de los algoritmos, es necesario reducir la dimensionalidad. Para este proceso se utilizó una combinación de técnicas que permiten de manera eficiente y sin perder información de las variables relevantes, elegir los datos de entrada óptimos.

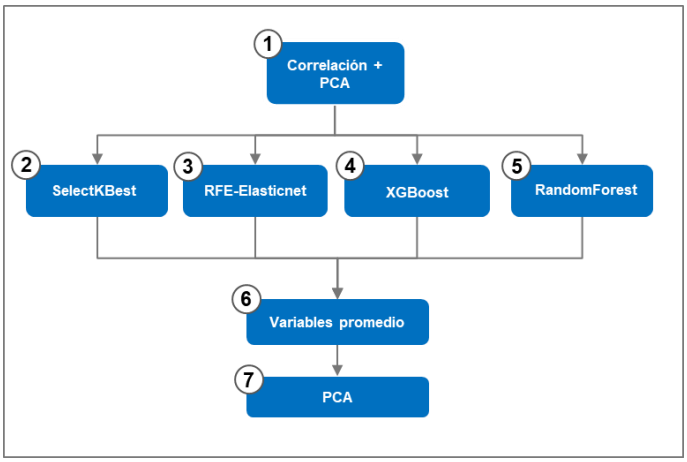

Ilustración 3. Proceso de reducción de la dimensionalidad

Se realizaron 7 pasos como muestra la Ilustración 3, integrando diferentes técnicas para recudir las dimensiones. La primera como plantean (Yu & Liu, 2003) es una técnica basada en la correlación de las variables, encontrando la relación que tienen las variables, para posteriormente unirlas en un componente principal utilizando PCA, análisis de componentes principales.

Este primer proceso genera la primera reducción para posteriormente aplicar 4 técnicas de manera independiente. En el artículo (Otchere et al., 2022) hacen una comparación de técnicas de selección de variables, el mejor resultado se obtiene utilizando todas las variables, pero el tiempo de computación es muy alto. En segundo lugar, se encuentra el modelo de python SelectKBest que en la mitad del tiempo de ejecución únicamente aumento el error en un 5%, esta técnica utiliza una función de puntuación, como la prueba estadística de chi-cuadrado, la ganancia de información o el coeficiente de correlación.

Las siguientes dos técnicas son la utilización de Random Forest y XGboost como generadores de puntaje de importancia para poder determinar la relación de las variables con un objetivo. En el estudio (Porkodi, 2014) se realiza la comparación de algoritmos de selección de variables, y el mejor resultado fue obtenido por estos dos algoritmos.

Finalmente se calcula un promedio de los puntajes para realizar la selección de variables que posteriormente ingresaron al modelo de manera de componentes principales con un análisis de PCA como muestra (Bro & Smilde, 2014)

PCA (Análisis de componentes principales): Es una técnica de análisis multivariante utilizada en estadísticas y aprendizaje automático para reducir la dimensionalidad generando combinaciones lineales de un conjunto de datos para crear nuevas variables que mantienen la mayor parte de la variabilidad original.

SelectKBest: Es una librería utilizada en Python que utiliza pruebas estadísticas para evaluar la importancia de cada característica en relación con la variable objetivo, utilizando como métrica la prueba de chi-cuadrado selecciona las mejores características que explican la variable objetivo.

5.3.2. Modelos XGBoost y Random forest

Los árboles de decisión son modelos de aprendizaje automático que representan decisiones basadas en reglas lógicas estructuradas en forma de un árbol, donde cada nodo interno representa una característica del conjunto de datos y cada rama indica una decisión o ramificación basada en esa característica, llevando finalmente a un nodo hoja que proporciona una predicción o resultado. Estos árboles se utilizan tanto en problemas de clasificación como de regresión y son conocidos por su facilidad de interpretación y versatilidad en el análisis de datos.

Para mejorar las predicciones de los árboles de decisiones se crean modelos de ensamble que toman el principio de los árboles de decisiones, pero corriéndolos múltiples veces para lograr capturar más variabilidad de los datos y tener mejores predicciones.

Bosque aleatorio (Random Forest):

Es un algoritmo que se basa en la construcción de múltiples árboles de decisión de manera independiente y luego combina sus resultados para lograr un modelo robusto y preciso. Cada árbol se entrena con una muestra aleatoria de los datos y utiliza una selección aleatoria de características para hacer divisiones. Para generar la recomendación final promedia los resultados de todos los árboles. En la Ilustración 4 se ve como se da el ensamble en este modelo, genera recomendaciones de árboles independientes y los une para la recomendación final.

XGBoost (Extreme Gradient Boosting):

Es un algoritmo basado en árboles que destaca por su eficacia en la predicción de valores numéricos y clasificación de datos. Se caracteriza por construir árboles de decisión secuencialmente, enfocándose en corregir los errores cometidos por árboles anteriores. En la Ilustración 4 se ve como se da el ensamble en este modelo, genera una recomendación y para hacer la siguiente aprende del modelo anterior.

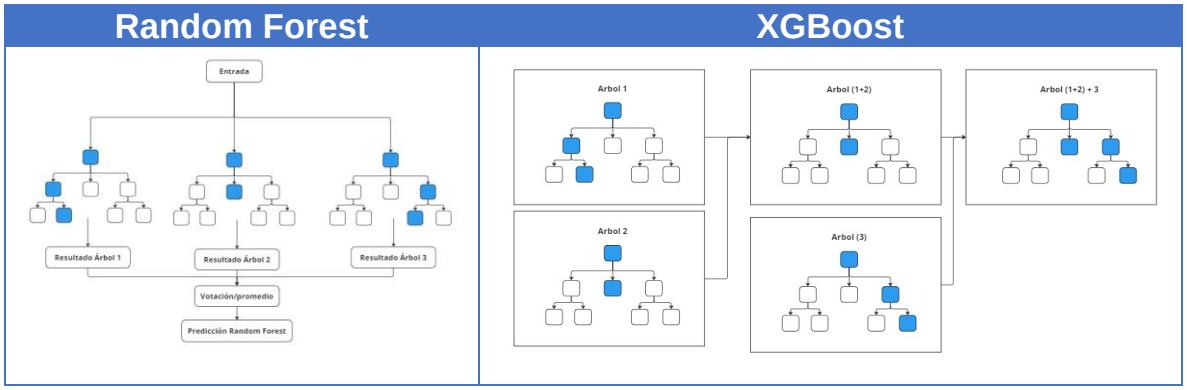

Ilustración 4. Explicación conceptual de Random Forest y XGBoost

5.3.3. Medias de desempeño de algoritmos

Para la medición del desempeño y precisión de los moldeos se utilizaron métricas asociadas al objetivo del moldeo. A continuación, se explica cada uno de los indicadores utilizados. Se utilizaron varios de estos indicadores como sugiere (Bilgilioğlu & Yılmaz, 2023)

Para los modelos de regresión se utilizaron los siguientes indicadores:

MAPE (Error Porcentual Absoluto Medio): Mide la diferencia entre el real y el pronosticado, dividido entre el valor real.

$$\frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}$$

n=Número de registros
y=Valor real
$\hat{y}_i =$ Valor pronosticado

MAE (Error Absoluto Medio): Mide la diferencia entre el valor real y el pronosticado, promedio por todas las observaciones.

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

n=Número de registros
y=Valor real
$\hat{y}_i =$ Valor pronosticado

R2 (coeficiente de determinación): Mide la proporción de la variabilidad en la variable dependiente que puede explicarse por el modelo de regresión.

$$1 - \frac{SSR}{SST}$$

SSR: Suma de los cuadrados de los residuos o errores de predicción.

SST: Suma total de los cuadrados o la variabilidad total de la variable dependiente.

Para los modelos de clasificación se utilizaron los siguientes indicadores:

Exactitud (En ingles Accuracy) Representa la proporción de predicciones correctas realizadas por un modelo en comparación con el número total de predicciones

$$\frac{Numero\ de\ predicciones\ Correctas}{Numero\ de\ predicciones\ Totales}$$

AUC-ROC (Área bajo la curva ROC) evalúa como un modelo de clasificación es capaz de distinguir entre las clases en un problema. La curva ROC es una representación gráfica que muestra la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR). Para considerar que el modelo tiene un mejor comportamiento que una variable aleatoria debe ser superior al 0.5

$$\int_0^1 sensibilidad(x) especificidad(x)$$

Sensibilidad: tasa de verdaderos positivos (TPR)
Especificidad: tasa de verdaderos negativos (TNR)

6. Metodología y resultados

El Proyecto se ejecutó utilizando metodología CRISP-DM®, desarrollando 5 de las 6 etapas que plantea la metodología, el proyecto no contempla dentro del alcance la sexta etapa, implementación (ver Ilustración 2).

6.1. Entendimiento de la situación

En esta sección se desarrolló la fase de entendimiento del proceso de cálculo del impuesto predial en Colombia, y de la variable objetivo de análisis.

Como medida indirecta del cálculo de la valorización se utilizaron los ingresos por impuesto predial, contando con información disponible del DNP y procuraduría general de la nación de 2008 al 2020 por cada municipio de Colombia. Como se muestra en el marco teórico en la Ecuación 1, es directa la relación entre la valorización y el impuesto predial. El incremento de los ingresos puede darse por 3 diferentes factores: valorización, incremento de la población y cambio de los factores tarifarios, por tanto, es mejor utilizar como variable objetivo un factor de incremento entre los años, además es más practico al utilizar modelos regresivos como Random forest y XGBoost tener un valor normalizado.

En la Ilustración 5, se observa la distribución de los Municipios de Colombia. En el gráfico izquierdo se muestran los municipios por nivel de ingresos (relación entre número de habitantes y valor de los predios), donde si tiene menos de mil millones de pesos se considera de ingresos bajos, por encima de $ 5 mil millones alto y más 50 mil millones muy alto. Solo 24 municipios hacen parte de grandes poblaciones

donde están las principales ciudades, por tanto, no es recomendable tratar la variable objetivo como el valor neto, se requiere una normalización para que puedan ser todos comparables. Los puntos de corte para las categorías se definieron donde hay una mayor variación del número de municipios y se da un cambio grande en el valor neto del ingreso. En el gráfico derecho se observa el crecimiento en los últimos 12 años en la recaudación, y se observa un comportamiento similar, donde existen municipios que han tenido un crecimiento muy superior al promedio, evidenciándose distribuciones similares si en lugar de 12 se toman otros periodos de tiempo (3 años, 6 años). En este proyecto se busca identificar con antelación esos crecimientos para identificar lugares potenciales para invertir.

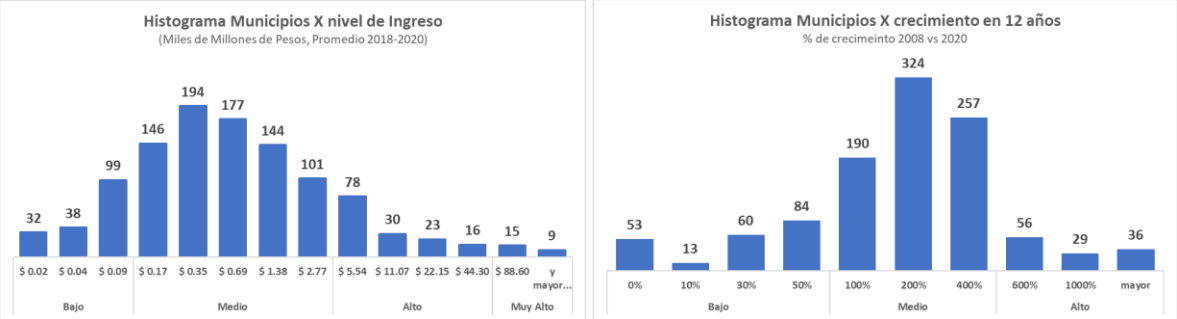

Ilustración 5. Histograma comportamiento de impuesto predial Municipios de Colombia

En la Tabla 1, se puede ver el número de municipios por categoría de ingreso y nivel de crecimiento, esto con el fin de ver que existe un comportamiento diferenciado de crecimiento según el nivel de ingresos de cada municipio. En la Tabla 1, adicionalmente se puede evidenciar que las ciudades dentro de la categoría de ingresos altos y muy altos presentan crecimientos más constantes. En municipios de ingreso medio se ven crecimientos altos.

Número de Municipios

Ingresos	Crecimiento Bajo	Crecimiento Medio	Crecimiento Alto
Bajo	198	108	9
Medio	196	449	49
Alto	6	71	7
Muy Alto		9	

% Municipios vs Total

Ingresos	Crecimiento Bajo	Crecimiento Medio	Crecimiento Alto
Bajo	18%	10%	1%
Medio	18%	41%	4%
Alto	1%	6%	1%
Muy Alto	0%	1%	0%

% Municipios vs Total de cat de ingresos

Ingresos	Crecimiento Bajo	Crecimiento Medio	Crecimiento Alto
Bajo	63%	34%	3%
Medio	28%	65%	7%
Alto	7%	85%	8%
Muy Alto	0%	100%	0%

Categorias de ingresos de impuesto predial (promedio 2018-2020) vs crecimiento de ingresos de 2008 a 2020

Tabla 1.Municipios por categoría de ingresos y tasa de crecimiento.

Para explorar los comportamientos se seleccionaron 5 municipios de diferentes tamaños que se conocen directamente, y en la Ilustración 6 se muestra el crecimiento normalizado de cada uno. Se observa Bogotá, una ciudad capital, donde su crecimiento es constante; segundo Villa de Leyva, un municipio en Boyacá que ha tenido un desarrollo importante por el incremento del turismo, este municipio ha tenido un crecimiento constante; Tercero Cajicá, donde desde 2012 ha tenido un incremento de población y valorización exponencial que hacia 2020 se ha venido estabilizando; Cuarto Macheta, donde su crecimiento ha sido estable y aún no se observa un claro comportamiento de crecimiento en el corto plazo y quinto Jenesano, un municipio que desde 2017 ha tenido un incremento exponencial, si se hubiera identificado en 2016 su potencial, antes de empezar su crecimiento, con una inversión pequeña se hubiera logrado una alta valorización.

Ilustración 6. Tasa de crecimiento, 5 Municipios.

6.2. Entendimiento de los datos

Con el objetivo de identificar cuales variables tienen relación con el potencial de crecimiento de los municipios, se identificaron varias fuentes de información oficiales del gobierno, DANE, DNP, Ministerio de Agricultura, Policía nacional, Contraloría general de la nación, Instituto Agustin Codazzi, Ministerio de Salud, Ministerio de turismo, Min TIC.

En la Tabla 2, se muestra el resumen de los sets de datos y las fuentes, y en el Anexo A se muestra la descripción detallada de las variables después de la preparación de los datos. En total fueron identificadas 1674 variables, sin embargo, debido a que no toda la información cubre los mismos años fue necesario un procesamiento detallado de limpieza de datos. En la Tabla 2 se muestran los datos resumidos de las 10 fuentes después del primer proceso de limpieza de datos. Los conjuntos de datos originales de seguridad de la policía, IGAC y DNP tenían aproximadamente 7 millones de registros inicialmente.

Fuente	Información	Registros	Variables	Años
Contraloría General	Ingresos impuesto predial; Banca Comercial; industria y comercio	14326	3	2008-2022
DANE	Municipio Oficial; Pobreza	1662	19	2005 y 2008
Datos Abiertos	# Hoteles; # Automotores	5569	19	2008-2020
DNP	Desempeño Fiscal; Economía; Población; Censo; Ambiente; Finanzas; Justicia; salud; Ordenamiento Territorial; Seguridad; vivienda; Desempeño; Educación; Pobreza; Mercado; seguridad Marítima; Área del municipio; Info departamento	16126	1147	2008-2022
IGAC	Características Predios	1101	159	2008-2022
Min. Agricultura	Cultivos agrícolas	13260	247	2011-2021
Min. Educación	Indicadores de educación	12343	34	2023
Min. Salud	IPS por municipio	865	15	2023
Mintic	Calidad del Internet	1118	11	2023
Policía Nacional	Tipos de delito diario	22280	20	2010-2022
		Variables	**1674**	

Tabla 2. Fuentes de Datos y Num. Variables

Se revisaron varios estudios para poder identificar que variables pueden tener relación con la valorización de los predios. En el artículo de Camacol[5] (Cámara Colombiana de la Construcción) resaltan 5 factores fundamentales a la hora de incrementar la valorización. El primero es la sostenibilidad, zonas donde el ambiente genere bienestar, por tanto, se incluyen 86 variables relacionadas con ecosistemas relevantes, deforestación y contaminación; segundo la ubicación y el entorno, relacionadas con este punto se incluye cercanía a centros urbanos, más de 211 variables de salud, educación y seguridad; Tercero y cuarto el tipo de construcción y tener proyectos por etapas, conectado a estos puntos se incluyen 241 variables de vivienda, ordenamiento territorial y características de los predios del IGAC y por último el estrato y condiciones socioeconómicas, incluidas en variables de vivienda, economía y desarrollo de la economía rural.

Complementando estas variables al revisar el trabajo de (Levantesi & Piscopo, 2020) sobre la importancia de las variables al predecir precios en Londres, se encuentra que la variable explicativa más fuerte es el incremento de la población explicado como un incremento de la oferta y la demanda de la región, y por tanto una mayor valorización. Para esto se incluyen 276 variables de censo y población, complementadas con características económicas.

[5] https://camacolantioquia.org.co/5-factores-que-inciden-en-la-valorizacion-de-un-inmueble/

En el estudio de (Yakub et al., 2020) se evidencia que las variables macroeconómicas tienen un efecto importante para la valuación de predios, por tanto, dentro de las variables económicas elegidas, se habla de la contribución al PIB de los municipios y la efectividad financiera. Los estudios de (Piergallini, 2020), (Égert & Mihaljek, 2007) y (Asal, 2018) complementan las variables con inflación, desempleo y competitividad como predictores del costo de vivienda.

En el artículo de (Andrews, 2010) sobre cálculo de factores de valuación en países de la OCDE (Organización para la Cooperación y el Desarrollo Económicos) del cual Colombia hace parte, resaltan el papel del incremento de la población que se traduce en mayor demanda y los ingresos que tienen los hogares.

En el estudio (Sandeep Kumar et al., 2019) recomiendan para realizar análisis de valor de vivienda utilizar todas las variables que se puedan incluir. Utilizaron modelos de aprendizaje automático y de las variables más importantes que identificaron destacan características de ubicación y de tamaño de la vivienda. Resaltan la importancia de variables como cercanía a calles específicas, tamaño de la vivienda, número de cuartos, baños y garajes. Estas variables fueron incluidas en el data set, dentro de la información del DNP.

6.3. Preparación de datos

La valorización de predios se da por el impacto de muchas variables, por tanto, se trabajó en el procesamiento de toda la información disponible. Los datos presentados en la Tabla 2 se encontraron en las fuentes de datos de manera detallada, como los casos de información de Lotes, Seguridad y Terridata, donde las tablas tenían millones de registros cada una y se requirió la creación de diferentes procesos ETL[6] para tener la información preparada para ser utilizada en el modelo, ver Tabla 3. Scripts Pre-Procesamiento.

Proceso	Descripción	Link
Raw Lotes	Entrada: 7,184,154 registros × 34 Variables Salida: 1,122 registros × 159 Variables	Script Lotes
Raw Seguridad	Entrada: 4,059,483 registros × 9 Variables Salida: 1104 registros × 243 Variables.	Script Seguridad
Raw Terridata	Entrada: 7,000,000 x 13 Salida: 17,010 registros × 1141 Variables.	Script Terridata
Base_0	Entrada: 12 conjuntos de datos. Salida Tabla Base_1102 registros × 1971 Variables	Script Base_0

Tabla 3. Scripts Pre-Procesamiento

[6] Los scripts en Python se encuentran disponibles en la cuenta de GitHub abierta https://github.com/Daniel1388/Towns_ML

Para la preparación de los datos se siguió el proceso que se muestra en la Ilustración 7. Iniciando con un proceso de extracción, que requirió un proceso de descarga de las páginas de las entidades oficiales. La información de datos abiertos no tiene conexiones automaticas actualizadas, en cada fuente hay un número diferente de años, variables con estructura diferentes y con más faltantes unas que otras. Por ejemplo, el conjunto de datos de la Policía Nacional requirió descargar más de 200 archivos. La información más consolidada que se obtuvo fue del portal Territata del DNP, no estaba en línea, eran archivos planos para ser descargados. Luego de tener la información en el entorno de trabajo se crearon las funciones que permitieron consolidar la información por cada uno de los tipos.

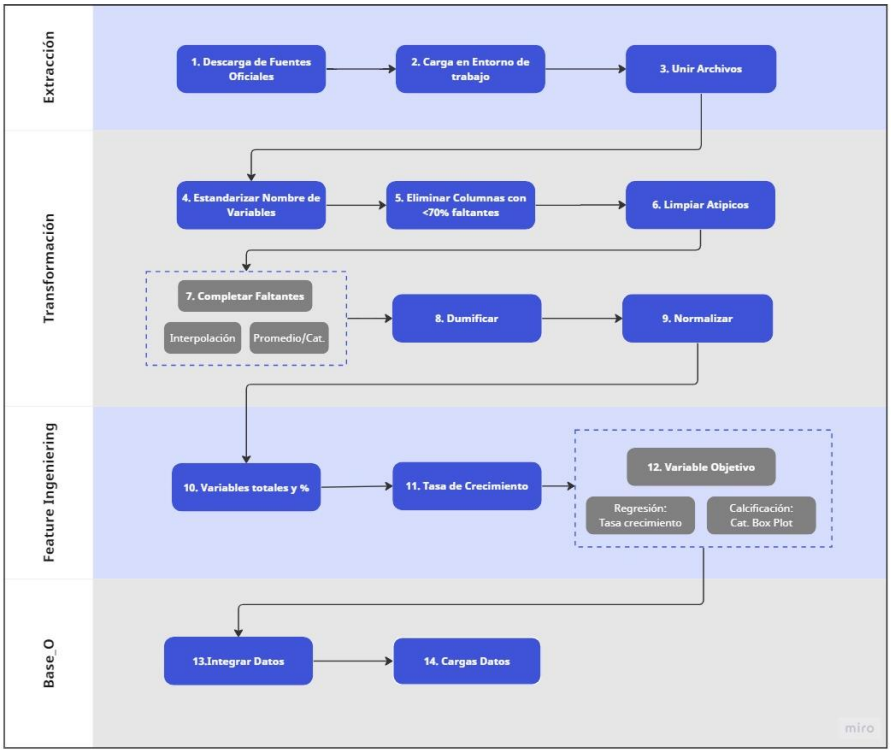

Ilustración 7. Proceso Preparación de Datos

En la segunda etapa, transformación, se requirió iniciar con una estandarización del nombre de las variables, para que no estuvieran repetidas, y fuera fácil su procesamiento y entendimiento, el nombre inicia con el tipo de variable seguido de una palabra de referencia, por ejemplo, Ambiente_Deforestacion. El ID de los municipios se definió como la llave de todas las tablas. Algunas fuentes no tenían el ID, solo el nombre por lo que se debió cruzar con el nombre del municipio para asociar un ID.

Luego se eliminaron las columnas con más de 70% de faltantes, y posteriormente se ajustaron atípicos utilizando el análisis de Boxplot. En el caso de variables que tenían cambios en el tiempo, se ejecutó el analisis por cada municipio ajustando el valor atípico al límite del bigote (Q3+1.5*Rango intercuartílico). Posteriormente para completar los faltantes, se realizó un proceso que elegía la función con mejor ajuste entre 7 seleccionadas (Lineal, exponencial, polinómica, logarítmica, cuadrática, cubica, logística.). Se utilizó esta función calculando cada punto de las series de tiempo y donde había faltantes se completaron, y donde había valores muy lejanos se trataron como atípicos y se utilizó el valor que generaba la función. En la Ilustración 8 se muestra un ejemplo del proceso, donde se inicia con la gráfica “Datos Reales” y al final del proceso se obtiene “Función de ajuste sin valores atípicos” que representa los valores que finalmente se usaran en el modelo. Para variables que no variaban en el tiempo y eran descriptivas se detectaron atípicos de toda la población y se completaron con los límites de Boxplot y los faltantes con el promedio. Todas las etapas tienen funciones en Python, los scripts se encuentran en Github (ver Tabla 3. Scripts Pre-Procesamiento). Con las variables resultantes se aplicó la función para generar variables Dummy para categorías, y normalización de las variables numéricas.

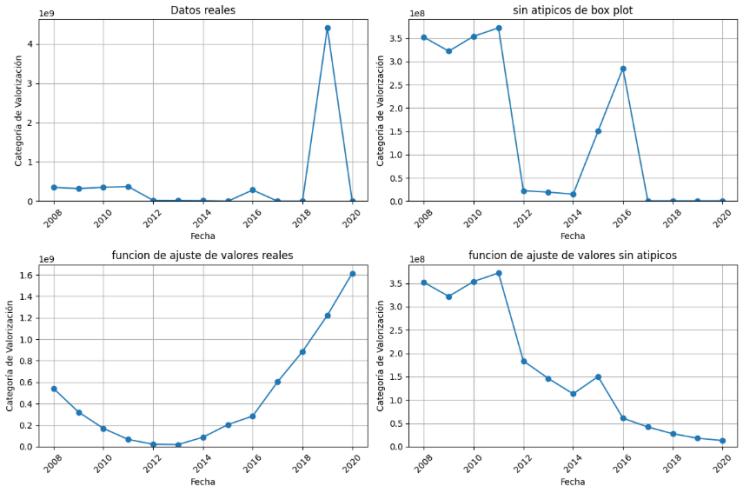

Ilustración 8. Limpieza de datos con ajuste de distribución.

En la Tercera Etapa, Feature Engeniring, se crearon variables compuestas para mejorar el performance de los modelos. Estas se agregaron como totales entre categorías y relaciones porcentuales entre datos, como por ejemplo el porcentaje de área construida sobre total del predio. Se creó una función que genera tasas de crecimiento entre los años para poder generar gradiente. Más adelante se puede observar cómo estos crecimientos ayudan considerablemente en los procesos de regresión y clasificación.

La variable objetivo para el proceso de clasificación es una categoría de tasa de crecimiento del municipio (Bajo, Medio, Alto, Muy alto), se definió el crecimiento con la Ecuación 2. Ecuación variable objetivo Utilizando n=12, y las categorías fueron generadas con los cuartiles, Q1=Bajo; Q2 y Q3=Medio, Q4=Alto, y los atípicos superiores Muy Altos. En el modelo de regresión, la variable objetivo es la tasa de crecimiento que se dará en los próximos años, utilizando n=3. El scrip tesis_Funciones tiene todas las funciones mencionadas en este apartado.

Finalmente, se integra la Base_0. Una tabla con toda la información lista para ser utilizada en el proceso de Modelado. En la Ilustración 9 se ve la estructura de la tabla con 1971 variables y el resultado del proceso, mostrando cuales variables quedaron definidas como tasas de crecimiento (TC), Valores con un solo dato de un año (FX), y para todos los casos la llave es el ID del municipio (PK).

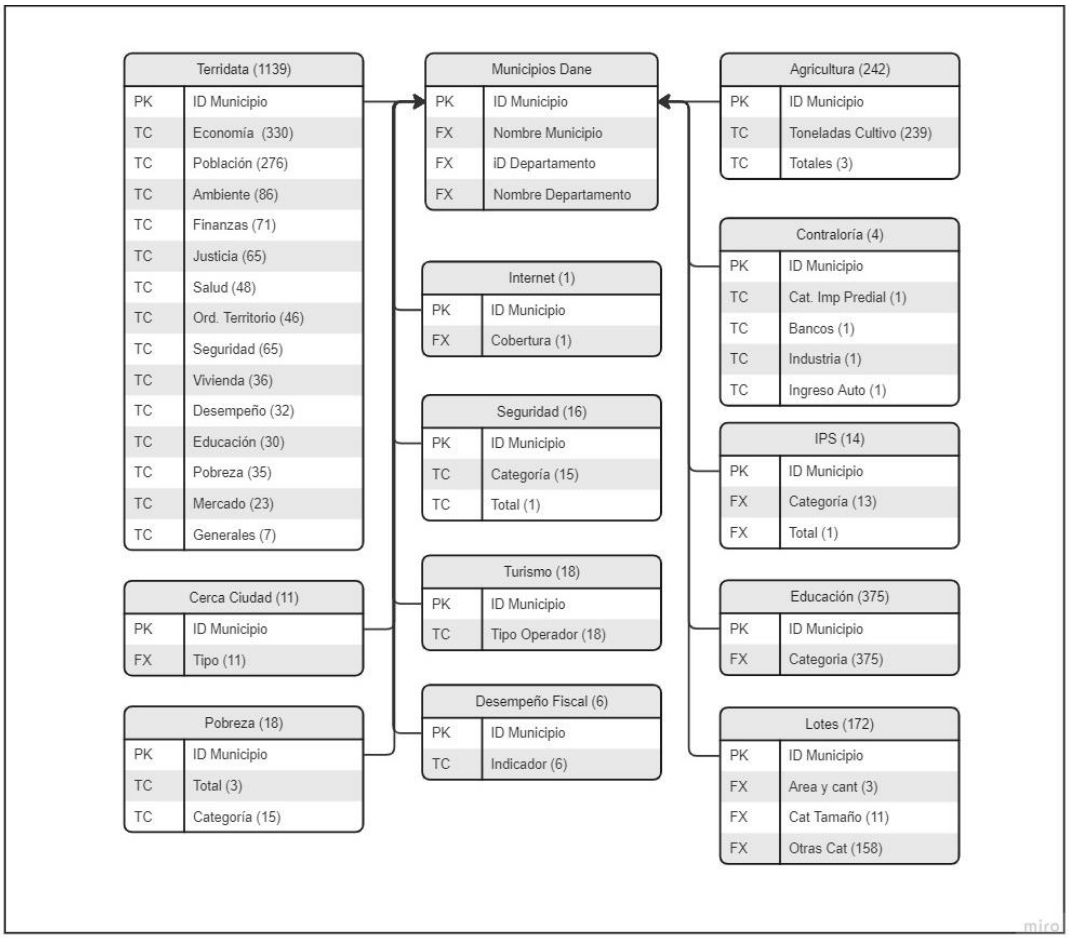

Ilustración 9. Diagrama entidad relación Base 0

6.4. Modelamiento

El proceso de modelamiento se divide en tres etapas, reducción de la dimensionalidad, donde de 1971 variables se seleccionaron aproximadamente 600 que logran explicar el 90% de la variabilidad, y un enfoque de 30 componentes principales; la segunda etapa generación predicción a partir de algoritmos, donde se ejecuta un proceso completo para clasificar y otro para regresión, generando cada uno una clasificación del potencial de crecimiento, para terminar en el proceso de la recomendación que integra los dos factores para generar la predicción final, y generar el conjunto de datos que se muestra en la herramienta de Power BI. En la Ilustración 10 se puede ver el flujo del proceso mencionado.

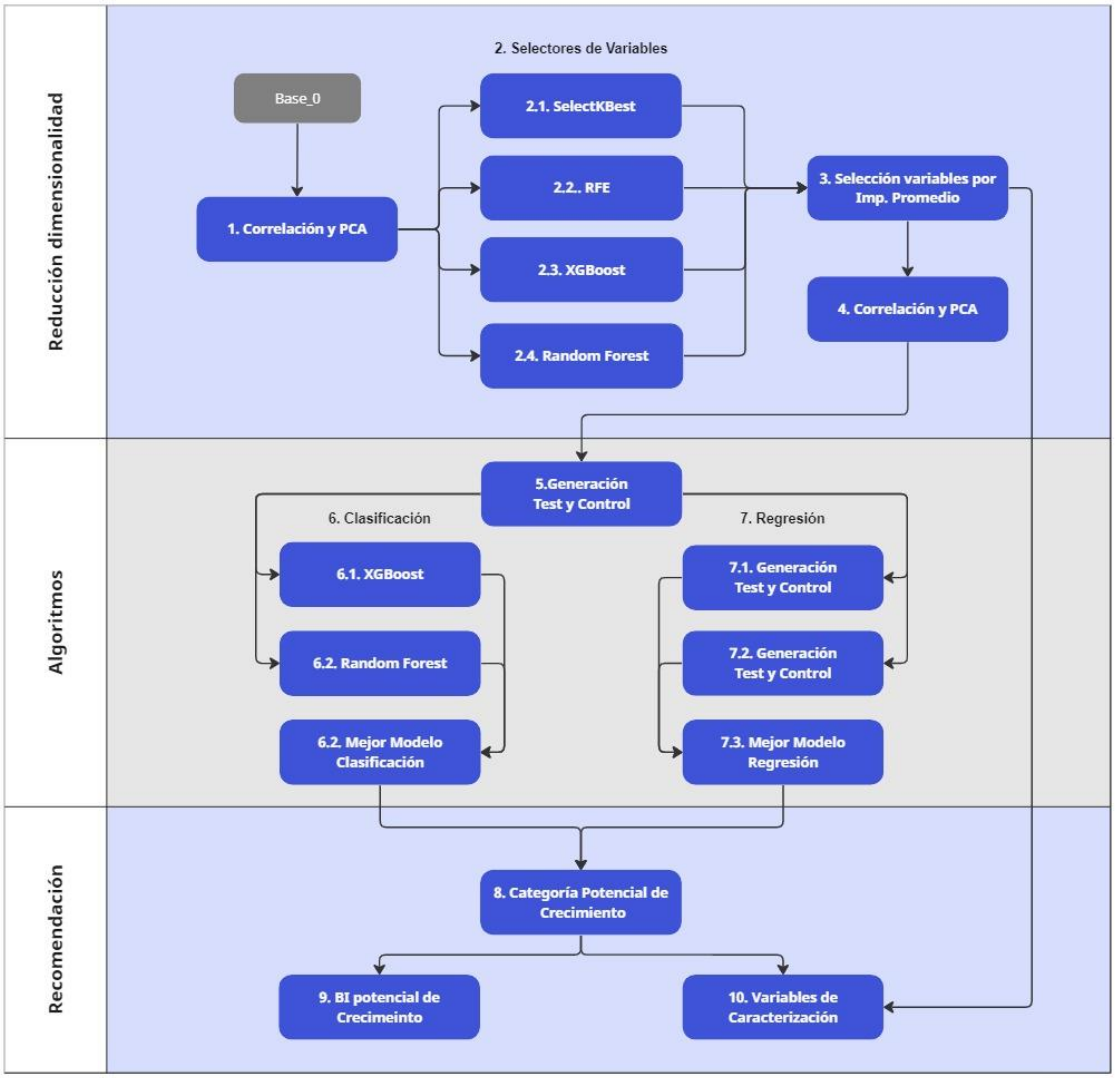

Ilustración 10. Diagrama proceso de Modelamiento

En la Tabla 4. Lista de Scripts de modelos de ML. Se listan los procesos en Python que cubren esta etapa de desarrollo.

Proceso	Descripción	Link
Clasificación	Script del proceso de clasificación, tiene como input tabla 0, y el output es las recomendaciones de clasificación y las variables relevantes.	Clasificación
Regresión	Script del proceso de regresión, tiene como input tabla 0, y el output es las recomendaciones de regresión del mejor modelo y las variables relevantes.	Regresión
Recomendación	Input, las recomendaciones de clasificación y regresión para generar el set de datos que utiliza el visualizador.	Integración
Funciones	Todas las funciones que se utilizan en el proyecto.	Funciones

Tabla 4. Lista de Scripts de modelos de ML.

6.4.1. Variable objetivo

Todo el proceso de modelamiento se corre en dos etapas, clasificación y regresión. Cada enfoque fue corrido en línea con el objetivo de calcular el potencial futuro de los Municipios. Del modelo de regresión y clasificación se obtiene una tasa de crecimiento que se utiliza para calcular el potencial final como muestra la Ecuación 3. Ecuación Potencial de Crecimiento donde él % de crecimiento futuro es el resultado del ejercicio de regresión, que evalúa directamente el crecimiento, por tanto, tiene un peso de 10, y la probabilidad de pertenecer a una categoría de crecimiento como resultado del algoritmo de clasificación tiene los pesos mostrados en la ecuación según sea Muy Alto, Alto, medio y bajo.

$$P\_Crecimiento = 10(\%\ Crecimiento\ Futuro) + 6(M\_Alta) + 4(Alta) + 2(Media) + (Baja)$$

Ecuación 3. Ecuación Potencial de Crecimiento

El modelo de regresión tiene como objetivo estimar que tanto va a crecer el ingreso por impuesto predial por municipio. Por tanto, la variable objetivo es el crecimiento a tres años como muestra la Ecuación 4, donde i, es el año actual que se compara con el valor 3 años en el futuro. Se elige trabajar a 3 años porque al tener 12 años de historia permite trabajar de forma más flexible grupos de entrenamiento, prueba y validación, también porque es cercano al mandato de un alcalde, donde se pueden ver cambios en el desarrollo del municipio. Los valores se escalan tomando como base el valor que cada municipio tenía en 2008, y posteriormente se normaliza para tener un valor entre 0 y 1. Se utiliza la información a nivel año municipio desde 2008 hasta 2017 para entrenamiento y validación, y los años 2018 al 2020 para predecir, generando la predicción del 2021 al 2023, datos que no están en los conjuntos de

datos. Este valor se distribuye normalmente como se puede apreciar en la Ilustración 2 en el grafico derecho.

$$Objetivo\_Regresión\ (\%\ Crecimiento_{i+3}) = \frac{(\ Imp.Predial_{i+3}) - (\ Imp.Predial_{2008})}{AVG(Ing.Imp.Predial_{2008})}$$

Ecuación 4. Variable Objetivo Regresión.

Para clasificación la variable objetivo es una categoría de que tanto se parece un municipio a los municipios que hoy en día están creciendo, la Base_0 está creada a nivel municipio, y todos los valores van comparando el promedio del 2008 al 2010 contra el promedio del 2018 al 2020. De esta manera se tiene por municipio la tasa de crecimiento, que posteriormente, realizando un analisis de cuartiles permite realizar categorías que se convertirán en la variable objetivo de los modelos de clasificación. Las categorías son mostradas en la Ilustración 11, donde la mayoría de los municipios presentan crecimiento porcentuales medios y bajos, el valor del analisis es ver que municipios están comportándose de manera muy similar a los que están mejorando, ya que pueden próximamente presentar crecimientos importantes.

$$Objetivo\ Clas.(Tasa\ Crecimiento) = \frac{AVG(\ Imp.Predial_{2018-2020}) - AVG(\ Imp.Predial_{2008-2010})}{AVG(Ing.Imp.Predial_{2008-2010})}$$

Ecuación 5. Variable Objetivo Clasificación

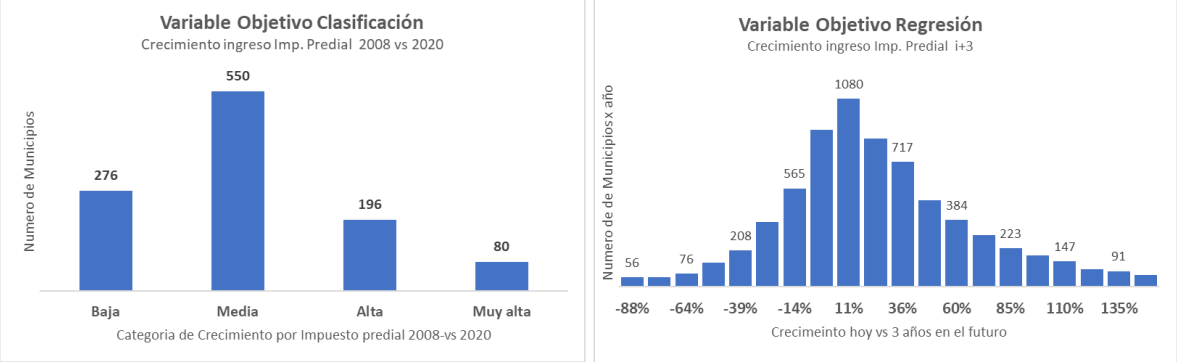

Ilustración 11. Distribución Variable Objetivo

6.4.2. Reducción de dimensionalidad

El proceso de reducción de la dimensionalidad busca seleccionar las variables más relevantes para entrenar los modelos de ML, usar las 1971 variables no es eficiente computacionalmente, además permite interpretar que tipo de variables explican

mejor las variables objetivo. Para desarrollar el proceso como se mostró en el marco teórico, se utilizaron varias técnicas recomendadas en distintos trabajos. El proceso se describe en la Ilustración 10. Se utilizó como Input la Tabla Base_0 de clasificación con registros por municipio, y la tabla de clasificación con información de municipio por año.

El primer paso es realizar un análisis de correlaciones entre todas las variables, y generar diferentes grupos entre las variables que comparten más de un factor de correlación superior al 80%, estas variables se agrupan y se genera un componente principal por grupo para de esta manera reducir la dimensionalidad. Esta primera etapa de correlaciones y PCA muestra la reducción presentada en la Tabla 5. Reducción de la dimensionalidad de VariablesSe calcularon las correlaciones entre todas las variables y las que tenían más del 80% se agruparon en una sola variable o componente principal.

Proceso	Clasificación	Regresión
Correlaciones y PCA	De 1971 a 1546 Variables (continua con 79% de las 1971 variables)	De 1024 a 616 Variables (continua con 60% de las 1024 variables)
Selectores de Variables	De 1546 a 623 Variables (continua con 31% de las 1971 variables)	De 616 a 106 Variables (continua con 10% de las 1024 variables)
PCA	De 623 a 50 Variables (continua con 3% de las 1971 variables)	De 106 a 50 Variables (continua con 5% de las 1024 variables)

Tabla 5. Reducción de la dimensionalidad de Variables

En la segunda etapa se utilizaron 4 modelos de selección de variables (SelectkBest, RFE, Random Forest y XGBoost), el proceso se muestra en Ilustración 10. Diagrama proceso de ModelamientoSe creó una función específica para procesar cada modelo, donde la entrada es la tabla con las variables y el output son todas las variables con un puntaje de 0 a 1 con la importancia de cada una. Posterior a tener el puntaje de cada uno de los modelos se generó un proceso que suma las variables y en una variable total, que se normaliza de 0 a 1, se definió el puntaje final de las variables. Posterior a esto se creó un proceso que selecciona las variables que representan el 90% de peso.

La reducción de dimensiones de este proceso se muestra en la Tabla 5. Reducción de la dimensionalidad de Variables. Un ejemplo de los valores de las 10 variables que presentaron una mayor relación en el caso de clasificación se presenta en la Tabla 6, donde se muestran variables de seguridad, pobreza y algunos componentes principales presentado alta relación con la variable objetivo. Los procesos y las funciones utilizados se encuentran en GitHub según ver **Anexo B Archivos de Github**.

ID	Variable	Selectkbest	RFE	XGBoost	R.Forest	Total_imp
0	Componente_Principal_1140	0.4%	0.0%	0.9%	3.9%	1.3%
1	Seg_fluvial_Armas_Incautadas	1.2%	0.3%	1.1%	0.0%	0.7%
2	Pobreza_ excretas	1.0%	0.6%	0.2%	0.2%	0.5%
3	Componente_Principal_1360	1.4%	0.1%	0.0%	0.3%	0.5%
4	Censo_porc_pob_hombres_20_24	0.2%	0.1%	1.4%	0.1%	0.4%
5	Economia_Rendimiento_Name	0.5%	0.1%	1.0%	0.1%	0.4%

Tabla 6 Ejemplo importancia de variables.

Se toma la variable de Total Imp, se ordena de mayor a menor. Posteriormente se tomaron las variables que representan un 90% y se corrió una función donde a partir de estas variables se crean 50 componentes principales para tener un conjunto de datos con el cual se puede entrenar los modelos fácilmente, logrando una reducción del 95% de la complejidad, y en la sección de modelos se hicieron pruebas de que tan diferentes son los resultados generando estos procesos de reducción en comparación de utilizar todas las variables.

6.4.3. Algoritmos de aprendizaje automático

Utilizando como input la tabla de variables y los componentes principales, se realizó la comparación de los modelos XGBoost y Random forest para ver cuál de las dos tiene un mejor performance para clasificar la categoría de crecimiento de los municipios. Sigue el proceso presentado en la Ilustración 10.

Se inició con la generación de 3 grupos, entrenamiento (64%), prueba (16%) y evaluación (20%). El grupo de prueba se utilizó para poder generar ajuste de hiperparámetros, para el entrenamiento del modelo se balancearon los datos, porque como se puede ver en la Ilustración 11, la variable objetivo es desbalanceada en sus categorías. Posteriormente con la información de evaluación se hizo la predicción y se evaluó la precisión. El script de procesamiento de los algoritmos de clasificación se encuentra en GitHub, en el siguiente enlace. Modelo de clasificación[7].

Dentro del proceso se definió una función que evaluaba que los resultados del modelo no estuvieran presentando sobreajuste al tener tantas variables, se construyó el grafico de aprendizaje donde mediante validación cruzada, incrementando el tamaño de la muestra se veía el performance de clasificación del grupo de prueba, y se evidencia que entre mayor es la

7 https://github.com/Daniel1388/Towns_ML/blob/main/Scripts/Tesis_ML_Model_Clasificaci%C3%B3n.ipynb

predicción, por lo que se puede concluir que existe una buena generalización del modelo de predicción.

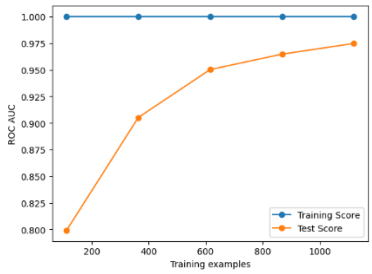

Ilustración 12. Gráfico de aprendizaje del modelo.

Se realizó la selección de los hiperparámetros utilizando validación cruzada con la función gridsearch, explorando la tabla que se muestra en la Tabla 7. Hiperparámetros XGBoost ClasificaciónY tomando los mejores parámetros, presentados tambien en la misma tabla.

Tabla de búsqueda de hiper parámetros	Mejores Hiper parámetros

`param_grid = {`
`'learning_rate': [0.1, 0.01, 0.001],`
`'n_estimators': [50, 100, 200, 500],`
`'max_depth': [3, 5, 7],`
`'min_child_weight': [1, 3, 5],`
`'subsample': [0.8, 0.9, 1.0],`
`'colsample_bytree': [0.8, 0.9, 1.0],`
`'gamma': [0, 0.1, 0.2, 0.3],`
`'reg_alpha': [0, 0.01, 0.1, 1],`
`'reg_lambda': [0, 0.01, 0.1, 1]}`
	`param_grid = {`
	`'learning_rate': [0.1],`
	`'n_estimators': [200],`
	`'max_depth': [7],`
	`'min_child_weight': [5],`
	`'subsample': [0.9],`
	`'colsample_bytree': [0.9],`
	`'gamma': [0.1],`
	`'reg_alpha': [0],`
	`'reg_lambda': [1]}`

Tabla 7. Hiperparámetros XGBoost Clasificación

Se realizaron diferentes corridas de los algoritmos para definir la mejor configuración de datos de entrada, hiperparámetros y tipos de modelos para encontrar una buena predicción. En la Tabla 8. Performance Algoritmos de clasificación. se presentan los resultados de exactitud, área bajo la curva ROC y tiempo de ejecución. En la columna número de variables se observa el input utilizado, por ejemplo 5 (PCA) significa que se procesó el algoritmo con 5 componentes principales que resumían todas las variables, en el caso de 600 hace referencia a las variables que representan el 90% del peso posterior al proceso de selección de variables, y 1500 son las variables después del proceso de integración de variables por correlación.

Número de Variables	Random Forest			XGBoost		
	Exactitud	Roc	Tiempo	Exactitud	Roc	Tiempo
5 (PCA)	43%	64%	45 seg	38%	61%	4 seg
15 (PCA)	53%	69%	56 seg	39%	65%	7 seg
50 (PCA)	51%	71%	1 min	37%	62%	18 seg
600	58%	81%	2 min	65%	83%	2 min
1500	55%	77%	3 min	64%	86%	3 min

Tabla 8. Performance Algoritmos de clasificación.

Como conclusión se observa que el modelo con mejor performance es el XGBOOST utilizando las 600 variables principales. Al ver los resultados es evidente como reducir las dimensiones genera que los modelos se ejecuten en un menor tiempo, estos valores se obtuvieron al procesar el modelo con un valor especifico de hiper parámetros, sin embargo, es de anotar que probando más combinaciones el tiempo crece exponencialmente.

El modelo de Random Forest presentó un resultado más consistente al reducir las variables, haciendo menos evidente la perdida de precisión con la generación de los componentes principales, en XGBOOST fue más claro cómo se pasa de un 65% a un 37% de exactitud al aplicar PCA, pero al procesar todas las variables presentó el mejor performance de los modelos generados.

Para el modelo de regresión se ejecutó el mismo proceso de evaluación de los modelos XGBoost y Random forest, en este caso los indicadores de comparación fueron el MAE (Error Absoluto Medio) que mide el promedio de las diferencias absolutas entre las predicciones del modelo y los valores reales, el R2 (coeficiente de determinación) que mide la proporción de la varianza total de la variable dependiente que el modelo explica y el MAPE (Error Porcentual Absoluto Medio), utilizado para medir la precisión de modelos. En la sección 5.3.3, se encuentran explicadas las medidas de clasificación y de regresión. Los hiperparámetros que mejor funcionaron en el modelo de regresión y las opciones de hiperparámetros utilizado están en la Tabla 9.

Tabla de búsqueda de hiper parámetros	Mejores Hiper parámetros
`param_grid = { 'max_depth': [10,20,30,40], 'learning_rate': [0.1,0.3,0.4], 'n_estimators': [100, 200, 300,600,800], 'colsample_bytree': [0.8, 1.0], 'gamma': [0, 0.1, 0.2], 'reg_alpha': [0, 0.1, 0.5,0.8], 'reg_lambda': [0, 0.1, 0.5,0.6], 'min_child_weight': [1, 3, 5], 'subsample': [0.2,0.5,0.8, 1.0] }`	`param_grid = { 'max_depth': [30], 'learning_rate': [0.3], 'n_estimators': [600], 'colsample_bytree': [0.1], 'gamma': [0], 'reg_alpha': [0.8], 'reg_lambda': [0.5], 'min_child_weight': [3], 'subsample': [1.0] }`

Tabla 9. Mejores hiperparámetros XGBoost de regresión

Luego de realizar múltiples corridas de los algoritmos en la Tabla 10, se observaron los resultados de los modelos de regresión probando diferentes configuraciones del conjunto de datos X, y comparando las medidas de desempeño generadas para los datos de validación (no fue utilizado para entrenar ni para la elección de hiper parámetros). Se observa que los dos modelos obtuvieron resultados muy similares, pero por poco se observa que el modelo XGBoost presentó un mejor performance con las 187 variables más importantes. Por tanto, este modelo fue utilizado para generar la recomendación. También se observa que el algoritmo Random Forest funciona mejor con la reducción de dimensionalidad generada por PCA.

Número de Variables	Random Forest				XGBoost			
	MAE	MAPE	R^2	Tiempo	MAE	MAPE	R^2	Tiempo
5 (PCA)	1.3	23%	67%	45 seg	1.8	28%	58%	10 seg
15 (PCA)	1.1	22%	70%	26 seg	1.8	28%	58%	10 seg
50 (PCA)	1.1	21%	71%	1 min	1.8	29%	53%	19 seg
187	1.2	21%	71%	4 min	1.1	20%	73%	34 seg
289	1.2	22%	71%	7 min	1.1	22%	69%	49 seg

Tabla 10. Performance Algoritmos de Regresión.

Las variables que más peso presentaron para cada uno de los modelos a la hora de clasificar y generar la regresión se encuentran en el **Anexo A Descripción de las variables**

6.4.4. Recomendación

Con los modelos de regresión y clasificación XGBoost seleccionados en el apartado anterior se realizó la recomendación y se calculó el factor de potencial de crecimiento de recaudación de impuesto predial de cada municipio de Colombia. Para el modelo de regresión se obtuvo un porcentaje de crecimiento esperado en los próximos tres años como variable objetivo, y para el modelo de clasificación se definió una categoría con la diferencia entre el 2008 y 2020. En la Tabla 11 se observa un ejemplo de la recomendación de cada uno de los algoritmos y del factor combinado generado. La columna **Potencial** es el valor generado por el algoritmo de regresión que indica en tres años como será el crecimiento respecto al 2008. Las columnas desde Bajo a Muy Alto son los puntajes de similitud que tiene cada municipio con la categoría de crecimiento. Finalmente, se calcula el factor de crecimiento de la siguiente manera: 10(potencial) + 6(Muy_Alto) + 4(Alto) + 2 (medio) + (Bajo). Los factores se definieron para hacer más evidente el potencial de crecimiento de municipios acelerados.

Id	Municipio	Potencial	Bajo	Medio	Alto	Muy Alto	Factor de crecimiento
25126	Cajicá	79,2%	0,1%	2,1%	96,9%	0,9%	1298,6%
68079	Barichara	72,9%	0,4%	1,9%	96,0%	1,7%	1239,2%
15367	Jenesano	53,8%	5,1%	9,9%	83,6%	1,4%	1055,4%
25426	Macheta	46,0%	1,6%	97,5%	0,7%	0,2%	1058,1%
76109	Buenaventura	31,1%	98,2%	1,3%	0,5%	0,1%	812,5%

Tabla 11. Cálculo del factor de potencial de crecimiento

Este factor se utiliza para catalogar los municipios en 5 grupos diferentes teniendo en cuenta dos ejes. En el eje Y, está el tamaño de recaudación actual del municipio y en el eje X, el factor calculado mediante la integración de los modelos de aprendizaje automático de clasificación y regresión (ver Tabla 11). Las categorías generadas son:

1. **Grandes Creciendo:** Alta valorización actual con alta aceleración de crecimiento, municipios grandes donde la inversión es alta, pero se siguen valorizando a una tendencia constante.
2. **Desacelerados:** Municipios de alta inversión que están creciendo, pero con tasas más bajas de aceleración.
3. **Emergentes:** Municipios donde actualmente no es tan costoso invertir, y están creciendo muy rápido.
4. **Esperanza Futura:** Municipios pequeños que tienen las condiciones para crecer, pero aún no ha empezado su aceleración.
5. **Estáticos:** Municipios que con la información disponible no muestran comportamientos similares a los municipios que están creciendo.

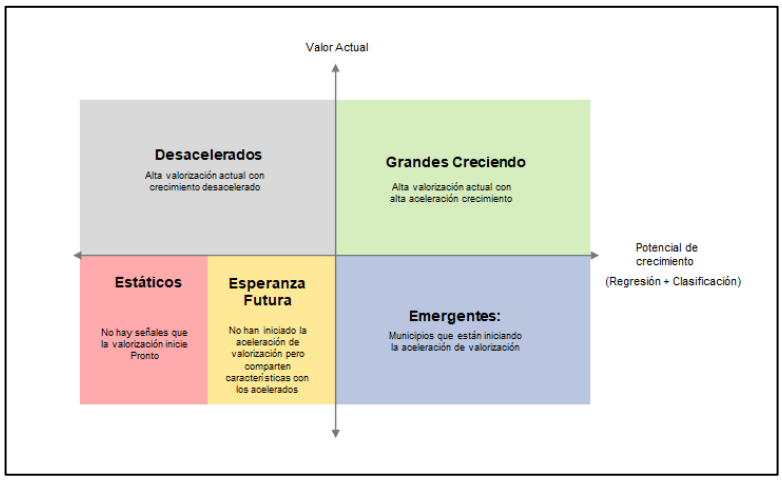

Ilustración 13. Categorías de clasificación de municipios según valorización actual vs potencial

En la Ilustración 13, se observan los diferentes cuadrantes que ocupan los municipios según las dos variables generadas para catalogarlos. En la Tabla 12 se observa la cantidad de municipios que quedaron en las diferentes categorías y que serán analizados en la sección de evaluación y resultados.

Categoría	Número de Municipios
Grandes Creciendo	162
Desacelerado	49
Emergentes	333
Esperanza Futura	274
Estáticos	282
Total, general	**1100**

Tabla 12. Cantidad de municipios por categoría de potencial de crecimiento.

Las categorías sirvieron para analizar como cada grupo tiene características específicas si se comparan con los datos de la población. En el siguiente apartado se describen las variables principales que tienen significancia estadística para cada grupo.

6.5. Evaluación y resultados

La generación de las variables combinadas y la definición del tamaño de los municipios permitió generar la asignación en el cuadrante de crecimiento como se observa en la Ilustración 14, donde el gráfico superior izquierdo muestra todos los municipios. Destaca el caso de Buenaventura, un municipio muy grande pero que presenta una desaceleración en su crecimiento, variables de seguridad limitan el potencial que tiene. En el gráfico de Cundinamarca se puede evidenciar dos tipos

de municipios, primero Chía y Funza, dos municipios muy grandes que siguen creciendo, tienen mucho potencial, pero invertir es costoso. Por otro lado, esta Viotá, un municipio creciendo de forma acelerada, pero al no ser muy grande aún, se encuentran buenas oportunidades. Este tipo de comportamiento también se evidencian en departamentos como Antioquia y Santander, con municipios que siguen creciendo como Sabaneta, Piedecuesta, y emergentes como Necoclí, Vélez y Barichara.

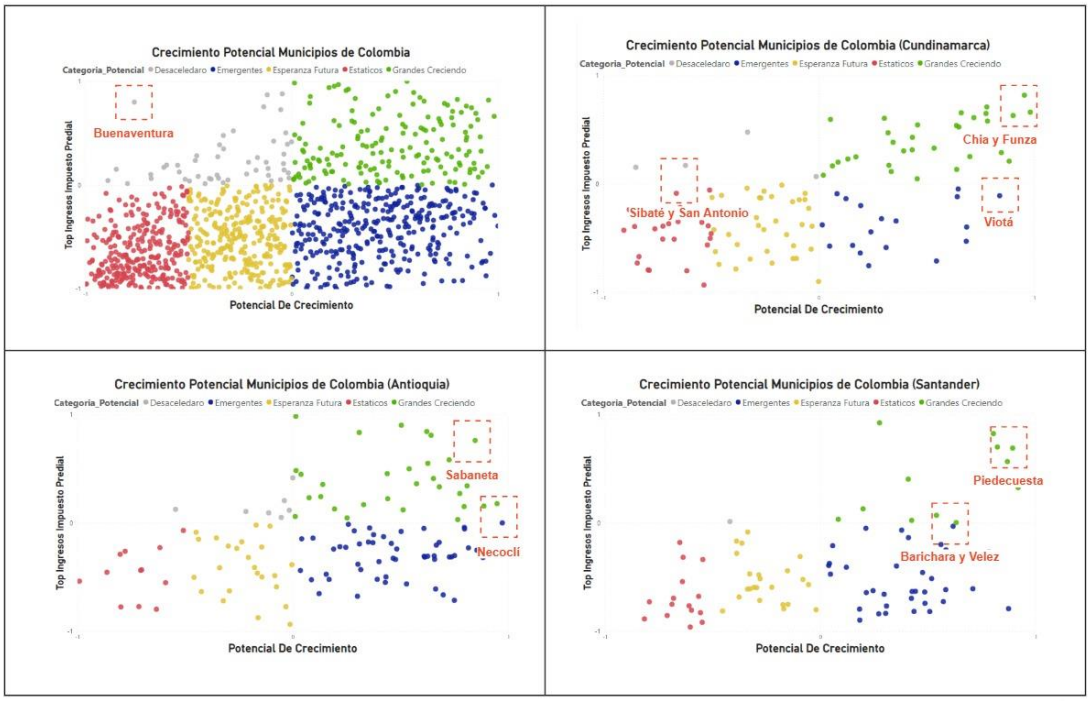

Ilustración 14. Gráficos de potencial de crecimiento

El comportamiento histórico de algunos municipios comentados anteriormente se observa en la Ilustración 15, donde se presenta el crecimiento de ingresos, que hace relación con la categoría antes mencionada, Buenaventura - Desacelerado. Municipios como Chía, Viotá y Necoclí muestran un evidente crecimiento en los últimos años, presentando un alto potencial, haciendo sentido con su categoría emergente. Sibaté y San Antonio del Tequendama, tienen crecimiento a velocidad menor por tanto sus categorías son desacelerado y estático respectivamente. En el caso de Buenaventura, se observa una desaceleración evidente, es un caso muy particular por el potencial que tiene, este requiere una importante intervención para modificar la curva de crecimiento.

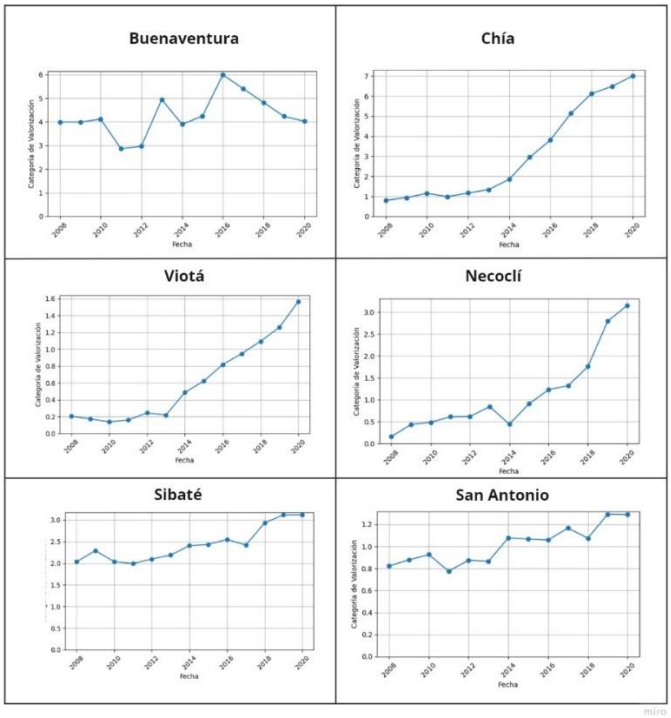

Ilustración 15. Ejemplo evolución Municipios

En la Ilustración 16 se observa en el mapa de Colombia[8] la asignación de las diferentes categorías, se ve como el sur del país que es en su mayoría selva aún no muestra un crecimiento considerable, y se ve rojo (estático); Los llanos orientales por otro lado, aunque no son municipios aún muy grandes en población han presentado aceleración en su crecimiento, y presentan en su mayoría municipios emergentes y esperanza futura; En el centro del país en ciudades como Bogotá y Medellín es evidente como los municipios que rodean estas ciudades se desarrollan (verdes: Grandes creciendo); en Bogotá los municipios que rodean al norte crecen, pero al sur no se presenta el mismo comportamiento, en la zona del Sumapaz donde hay bosques protegidos el crecimiento ha sido menos acelerado; En Medellín se presenta un comportamiento similar, pero destaca que muchos municipios del departamento de Antioquia está emergiendo al mismo tiempo (Azul); En la costa caribe destaca la Guajira, donde Uribia es grande y sigue creciendo; Cerca de Cali hay un gran número de municipios grandes, pero destaca como a sus alrededores varios de estos municipios están desacelerados.

[8] Mapa interactivo: https://github.com/Daniel1388/Towns_ML/blob/main/Resumen/mapa_interactivo.html

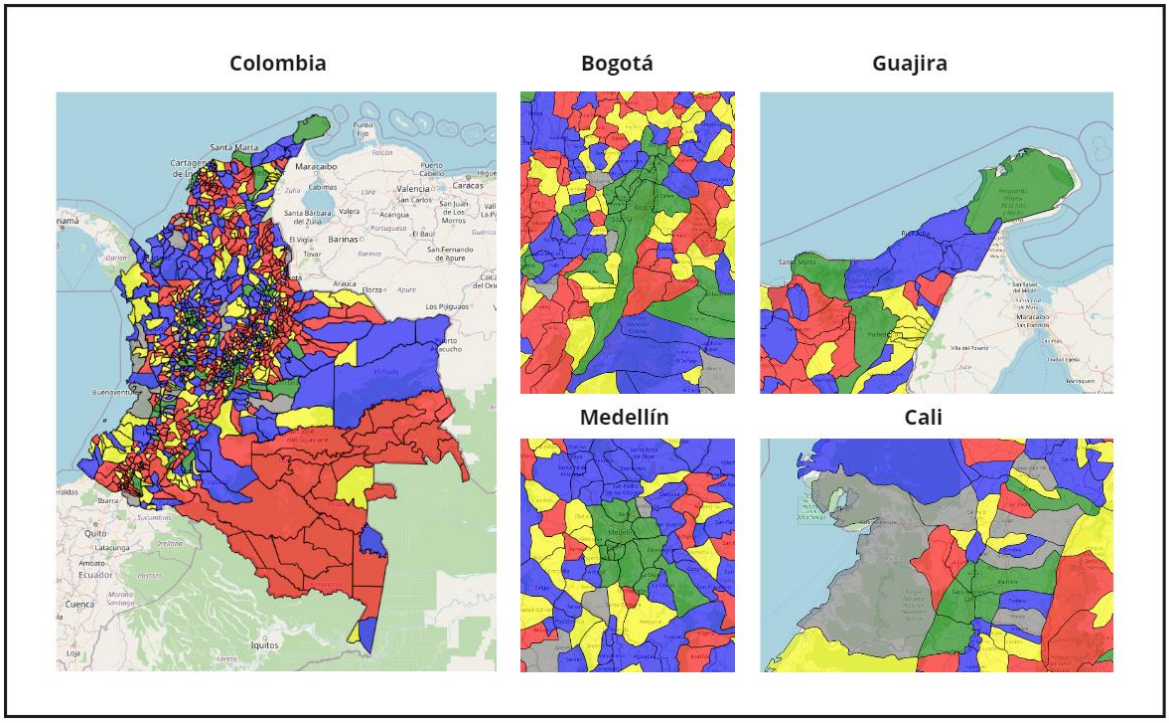

Ilustración 16. Mapa Municipios por categoría.

Para el análisis de las variables, se realizaron pruebas de significancia entre el promedio de la población de los municipios de Colombia, y cada una de las categorías. Clasificando las diferentes variables según la dirección de la variación que presentan respecto a la población (Positivas o Negativas) y una categoría de magnitud, según sean un cambio grande o pequeño. En la Tabla 13, se evidencian las características más relevantes que presentan significancia. Todo este proceso se puede observar en el Script de generación de la recomendación Final (**Ver Anexo B Archivos de Github**).

Categoría	Negativas (Menos Que la población)	Positivas (Más que la población)
Grandes Creciendo	Áreas en estratos bajos; Pobreza no atendida; Población sobre 80 años; incendios Forestales, inundaciones; Repitencia en educación; Delitos sexuales en Proporción	Inversión en Medio Ambiente; Crecimiento poblacional; Economia_PIB_municipal_actividades_terciarias_Efecto; Cobertura del sistema de salud; Cobertura de Educación; Mejor manejo de recursos para educación; Riesgos ambientales; Investigación ambiental; Cobertura educación en primera infancia; Conciliación en procesos judiciales; Proporción de zonas urbanas
Desacelerados	inversión Nacional; ingresos per cápita; Censo_pob_rural; Suministro gas agua; Alimentacion_escolar; seguridad_TasaHurtoComun	RegistrosHogares; IPS_Total; Conflicto_obras_civiles_urbanas; Gastos_corrientes
Emergentes	Formalidad laboral; Avaluos Bajos (Oportunidad de Invertir); Menos predios por población (Oportunidades de lotear); Poblaciones; Instituciones ambientales en desarrollo; Baja industria manufacturera; Casos judiciales	Agricultura_Cacao_Efecto; Ambiente_area_no_amenazada; Ambiente_porcentaje_area_humedales; Uso adecuado de lotes; Pobreza; Áreas de Ecosistemas; Conservación Ambiental; Rendimiento agrícola; Producción Plátano y yuca; Deserción escolar urbana; Analfabetismo urbano; Proyectos sin contratar; proyectos nuevos de gran inversión; Tasa de extorción; Hurto a motos; Cobertura en primaria; Cobertura de servicios públicos; Agua potable; Economia_PIB_Agricola; Educacion_Deserción; Alimentación escolar
Esperanza Futura	IPS_Total; Mercado laboral; Predios urbanos; Pobreza; Crecimiento poblacional; Desarrollo Urbano; Mercado laboral no formalizado; Ambiente_indice_vulnerabilidad	Buena Gestión; Gobierno_transparencia; Recaudo; Seguridad_grupo_DI_Efecto; Desempeno_Salud_Efecto; Desempeno_Seguridad_Efecto; Finanzas_Cultura; Finanzas_num_proyectos_terminados; Finanzas_regalias_per_capita; Mercado_Laboral; Estado_catastro_rural_Efecto; salud_CobRegSubsidiado; vivienda_servicios_publicos_CoberturaAlcantarilladoCenso_Efecto; Rendimiento agricola; Educacion_Aprobación; Finanzas_ingresos_recursos; Agricultura_Cacao,Aguacate,Frijol y Yuca; Lejos de ciudades; Recursos_asignados_per_capita_SGP
Estáticos	Numero de predios urbanos; Logros Educativos; Generación de Valor agregado; Crecimiento de población; Afiliación sistema de salud; Infraestructura de salud; Ambiente_conocimiento; Censo_pob_urbana; Fiscal_Apoyo_Nación; Mercado laboral pequeño; Formalidad laboral; Nivel de pobreza; Nivel de hacinamiento; Índice de turismo; Nivel de educación	Ambiente_area_ecosistemas; Tamaño Bosques; Mineria; Deserción escolar y repitencia; Lotes grandes; violencia interpersonal; Tasa de mortalidad en neonatos; Deficit de vivienda; Ambiente_area_amenazada; Cultivos de café y maíz; Alta afiliación al SISBEN

Tabla 13. Caracterización Categoría de Potencial de Crecimiento

La caracterización muestra diferentes comportamientos de las categorías. Los municipios **Grandes Creciendo**, donde invertir en finca raíz es más costoso por la etapa de madurez en que se encuentran, tienen en común que han logrado mejorar en seguridad, salud y reducción de la pobreza multidimensional. Avanzando en el cuidado por el medio ambiente, generando más cobertura de salud y educación. En estos municipios la educación de primera infancia se ha hecho más fuerte para complementar las etapas del sistema educativo. Tienen procesos judiciales más robustos, y las tasas de conciliación per cápita son mayores que en otros lugares, destacando por ser centros urbanos desarrollados, donde hay más lotes, ya que se ha permitido dividirlos para poder urbanizar, lo que trae una mayor valorización de predios.

Los municipios **Desacelerados** fueron los que tenían menos características con diferencias significativas en comparación de la población. Lo que destacó en este grupo es la falta de inversión nacional y la reducción de la población rural. Siguen trabajando para mejorar su sistema de salud, pero en conjunto no se destacan por sobresalir en muchas variables específicas.

Los Municipios **Emergentes** presentan variables muy interesantes, empezando tienen avalúos bajos, y menos predios, lo que los convierten en una gran oportunidad para invertir. Tienen actualmente poblaciones en crecimiento e industrias manufactureras pequeñas. El componente agrícola es desarrollado y cuentan con un potencial ambiental grande que puede convertirlos en referentes de turismo. Enfrentan problemas al estar creciendo sus centros urbanos, su deserción estudiantil y analfabetismo urbano es alto, ya que hasta ahora se están consolidando. Pero se evidencia como tienen más proyectos de desarrollo que la media y el gobierno invierte en esos municipios.

En el caso de la categoría **Esperanza Futura,** municipios que aún no han empezado a crecer de forma acelerada, pero que tienen otras variables que los hacen destacar. Se encontró una alta diferenciación por el manejo de los recursos públicos, así sean pequeños se destacan por buena gestión y transparencia. Están mejorando el recaudo para poder ejecutar más proyectos. Sus mercados laborales están creciendo, así como la cobertura de servicios públicos, tienen un sector agrícola fuerte y han encontrado la forma, según su tamaño, de captar recursos del gobierno central. Tienen que seguir trabajando, hay que desarrollar el sistema de salud e incrementar la formalización laboral. Tienen centros urbanos pequeños que deben seguir creciendo, sin embargo, con buena gestión pueden hacerlo de forma ordenada.

Finalmente, la categoría **Estáticos,** donde resaltan municipios con centros urbanos pequeños, que tienen oportunidad de incrementar la generación de valor. Deben fortalecer su sistema de salud y educación. Son municipios en zonas apartadas, lejos de centros urbanos grandes y con muchos bosques. El mercado laboral es pequeño, hay problemas de pobreza, pero no de hacinamiento y tienen un costo de vida más bajo. Son zonas donde la minería se destaca, no existe mucha manufactura, se concentran en materias primas. Los terrenos son grandes y aún no divididos en parcelas, se evidencia altos cultivos de café y maíz. El sistema de salud tiene un apoyo muy alto del SISBEN.

En la Tabla 14, se observa el top 10 de municipios de cada categoría, son aquellos que mejor desempeño tienen dentro de cada uno de los grupos.

Categoría	Top 10
Grandes Creciendo	Cundinamarca: Chía; Cundinamarca: Funza; Santander: Floridablanca; Cesar: Valledupar; Antioquia: Sabaneta; Santander: Piedecuesta; Bogotá: Bogotá; Cundinamarca: Tocancipá; Boyacá: Tunja; Nariño: Pasto
Desacelerado	Norte de Santander: Cúcuta; Caldas: Manizales; Córdoba: Montería; Quindío: Armenia; Valle del Cauca: Guadalajara de Buga; Antioquia: Girardota; Nariño: Ipiales; Tolima: Melgar; Valle del Cauca: Florida; Valle del Cauca: Cartago
Emergentes	Antioquia: Necoclí; Chocó: El Litoral del San Juan; Bolívar: Turbaná; Antioquia: Segovia; Cesar: Chiriguaná; Antioquia: San Vicente; Cesar: La Gloria; Cundinamarca: Viotá; Chocó: Juradó; Sucre: San Benito Abad
Esperanza Futura	Risaralda: La Virginia; Boyacá: Samacá; Valle del Cauca: Andalucía; Tolima: Chaparral; Antioquia: Támesis; Cundinamarca: Sasaima; Valle del Cauca: Obando; Caldas: Salamina; Cundinamarca: Gachancipá; Tolima: Alvarado
Estáticos	Quindío: Circasia; Cundinamarca: Nemocón; Antioquia: Peñol; Guaviare: San José del Guaviare; Córdoba: Pueblo Nuevo; Valle del Cauca: La Victoria; Córdoba: San Antero; Casanare: Villanueva; Chocó: Istmina; Cundinamarca: San Antonio del Tequendama

Tabla 14. Top 10 municipios y su departamento por categoría

7. Conclusiones

Una vez agrupados los municipios de Colombia en 5 categorías en función del potencial de crecimiento de ingresos de impuesto predial, utilizando modelos de aprendizaje automático, se concluye que es un método adecuado para identificar las variables que explican la variable objetivo.

Se logró calcular el potencial futuro de crecimiento utilizando un modelo (XGBoost) que procesa un gran número de variables, y permite apoyar la caracterización. Se consolidó un número alto de variables (1971) después de generar varios procesos robustos de consolidación y limpieza de datos. Este fue un paso fundamental, porque se debía garantizar la coherencia de la información con la variable objetivo, y la información base estaba dispersa en datos abiertos.

Los resultados de la categorización de municipios permitieron evidenciar comportamientos comunes entre los municipios de cada una de las categorías (Grandes Creciendo, Desacelerados, Emergentes, Esperanza futura y Estáticos). Se confirma la importancia del PIB como medida que destaca en los municipios con mayor potencial de crecimiento, en línea con la reducción de la pobreza. Municipios con sistemas de salud y educación fuerte crecen más rápido. Los municipios emergentes están en un proceso de desarrollo, presentan crecimiento acelerado de sus centros urbanos, y empieza a enfrentarse con problemáticas como deserción escolar urbana, incremento de hurtos y necesidad de formalización laboral. Estas zonas emergentes presentan avalúos bajos y predios mas grandes que tienen posibilidad de ser divididos para su venta, factores que hacen las zonas atractivas para invertir, porque con menos dinero se puede obtener predios con potencial de crecimiento.

Otros municipios que destacan en su oportunidad son los de esperanza futura, se puede concluir con los resultados que al ser lugares que aún no están creciendo de forma acelerada y destacan variables como buena gestión gubernamental, sistema de salud, educación y servicios públicos, presentan un buen potencial a mediano y largo plazo. Los municipios estáticos aún presentan condiciones complejas y no hay evidencia en los datos de que cambien al corto plazo. Se concluye que son zonas apartadas con muchas barreras naturales como selva y presentan conflictos de seguridad, falta de inversión y baja presencia del estado, por lo que se que efectivamente aún presentan un alto riesgo de inversión.

Se genera una contribución al probar un caso de uso en Colombia de técnicas de ensamble para lograr reducción de la dimensionalidad, cumpliendo el objetivo de identificar la mejor configuración de las variables para optimizar el rendimiento de los modelos. Estas técnicas lograron producir buenos resultados, permitiendo entrenar los modelos de una forma rápida, sin perder un buen resultado. Es importante probar diferentes modelos, ya que el desempeño varía según el número de variables y el modelo entrenado, por ejemplo XGBoost perdió exactitud utilizando componentes principales, mientras que no afectaba al modelo Random Forest, por lo que, con base en los resultados de la Tabla 8 y Tabla 10, se concluye que el mejor modelo para procesar la información disponible fue un XGBoost reduciendo la dimensionalidad de 1971 variables a 600 para el caso de clasificación y llegando a 289 para el caso de regresión.

Utilizando las recomendaciones generadas de los modelos entrenados, se logró producir el potencial futuro de crecimiento. Se genera un valor ensamblado entre el crecimiento futuro (pronóstico 3 años en el futuro) y la similitud entre variables de unos municipios con aquellos que actualmente crecen (Modelo de clasificación). Se contribuye al conocimiento en Colombia al definir este factor para cada uno de los municipios del país.

Se realizó la caracterización en 2 etapas, primero, encontrar variables relacionadas al crecimiento. Estas se obtuvieron al ser seleccionas en los procesos de aprendizaje de regresión y clasificación. Se concluye que las variables más relevantes al pronosticar el modelo de regresión fueron la tendencia de crecimiento de meses anteriores, incrementos en población, evolución en servicios públicos, evolución mercado laboral e indicadores de seguridad. Al momento de clasificar se mantienen constantes, complementando con desarrollo del sistema de salud, educación, rendimiento económico y pobreza. (las variables detalladas de importancia se pueden ver en el **Anexo A Descripción de las variables**).

Finalmente se concluye que los métodos de aprendizaje automático son altamente efectivos en procesos de caracterización masiva, logrando generar categorías que describen efectivamente el comportamiento de los municipios e identifican las

variables relacionadas. Se evidencia como variables como seguridad, educación, sistema de salud y desempeño financiero están altamente relacionadas en los comportamientos de los municipios.

8. Trabajo futuro

El presente trabajo consolidó un gran número de datos de diferentes fuentes oficiales de Colombia, lo cual permite que para próximos estudios se puedan evaluar otras funciones objetivos diferentes, por ejemplo, agricultura, seguridad, comercio, turismo etc. Es importante en proyectos futuros lograr conexiones de los datos a fuentes que se actualicen en tiempo real para de esta manera poder tener las recomendaciones actualizadas para mejorar el proceso de toma de decisiones.

En próximos estudios se puede ahondar en la exploración de las variables utilizando procesos para demostrar causalidad y tener recomendaciones automáticas para priorizar acciones a tomar de los diferentes municipios. Si en estudios posteriores se logra adquirir información detallada de valor de los diferentes predios, puede complementarse el estudio para saber en qué zonas se puede buscar una mejor inversión.

9. Referencias bibliográficas

Alvarez, F., Roman-Rangel, E., & Montiel, L. V. (2022). Incremental learning for property price estimation using location-based services and open data. *Engineering Applications of Artificial Intelligence*, *107*. https://doi.org/10.1016/j.engappai.2021.104513

Andrews, D. (2010). Real House Prices in OECD Countries: The Role of Demand Shocks and Structural and Policy Factors. *OECD Economics Department Working Papers*, *831*.

Asal, M. (2018). Long-run drivers and short-term dynamics of Swedish real house prices. *International Journal of Housing Markets and Analysis*, *11*(1). https://doi.org/10.1108/IJHMA-08-2017-0070

Bilgilioğlu, S. S., & Yılmaz, H. M. (2023). Comparison of different machine learning models for mass appraisal of real estate. *Survey Review*, *55*(388), 32–43. https://doi.org/10.1080/00396265.2021.1996799

Bro, R., & Smilde, A. K. (2014). Principal component analysis. In *Analytical Methods* (Vol. 6, Issue 9). https://doi.org/10.1039/c3ay41907j

Chen, Z., Goh, H. S., Sin, K. L., Lim, K., Chung, N. K. H., & Liew, X. Y. (2021). Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques. *Advances in Science, Technology and Engineering Systems Journal*, *6*(4). https://doi.org/10.25046/aj060442

Égert, B., & Mihaljek, D. (2007). Determinants of House Prices in Central and Eastern Europe. *Comparative Economic Studies*, *49*(3). https://doi.org/10.1057/palgrave.ces.8100221

Lee, C. (2021). Exploring uncertainty in land price through learning models. *International Journal of Advances in Soft Computing and Its Applications*, *13*(2), 78–94.

Levantesi, S., & Piscopo, G. (2020). The importance of economic variables on London real estate market: A random forest approach. *Risks*, *8*(4). https://doi.org/10.3390/risks8040112

Li, Y., & Chen, W. (2020). A comparative performance assessment of ensemble learning for credit scoring. *Mathematics*, *8*(10). https://doi.org/10.3390/math8101756

Ma, J., Cheng, J. C. P., Jiang, F., Chen, W., & Zhang, J. (2020). Analyzing driving factors of land values in urban scale based on big data and non-linear

machine learning techniques. *Land Use Policy*, *94*. https://doi.org/10.1016/j.landusepol.2020.104537

Otchere, D. A., Ganat, T. O. A., Ojero, J. O., Tackie-Otoo, B. N., & Taki, M. Y. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering*, *208*. https://doi.org/10.1016/j.petrol.2021.109244

Piergallini, A. (2020). Demographic change and real house prices: a general equilibrium perspective. *Journal of Economics/ Zeitschrift Fur Nationalokonomie*, *130*(1), 85–102. https://doi.org/10.1007/s00712-019-00670-y

Porkodi, R. (2014). Comparison of Filter Based Feature Selection Algorithms : an Overview. *International Journal of Innovative Research in Technology & Science(IJIRTS)*, *2*(2).

Ravi, S., & Larochelle, H. (2017). Optimization as a model for few-shot learning. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*.

Sandeep Kumar, E., Talasila, V., Rishe, N., Suresh Kumar, T. V., & Iyengar, S. S. (2019). Location identification for real estate investment using data analytics. *International Journal of Data Science and Analytics*, *8*(3). https://doi.org/10.1007/s41060-018-00170-0

Yakub, A. R. A., Hishamuddin, M. A., Kamalahasan, A., Jalil, R. B. A., & Folake, A. F. (2020). The effect of adopting micro and macro-economic variables on real estate price prediction models using ann: A systematic literature review. In *Journal of Critical Reviews* (Vol. 7, Issue 11). https://doi.org/10.31838/jcr.07.11.88

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proceedings, Twentieth International Conference on Machine Learning*, *2*.

Zhu, Z., & He, K. (2022). Prediction of Amazon's Stock Price Based on ARIMA, XGBoost, and LSTM Models. *Proceedings of Business and Economic Studies*, *5*(5). https://doi.org/10.26689/pbes.v5i5.4432

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). A Comparative Analysis of XGBoost. *https://www.researchgate.net/publication/337048557_A_Comparative_Analysis_of_XGBoost*.

Henriksson, E., & Werlinder, K. (2021). Housing Price Prediction over Countrywide Data: A comparison of XGBoost and Random Forest regressor models. *Independent thesis*.

Anexos

A. Descripción de las variables

Los archivos anexos del listado se encuentran en el siguiente enlace de GitHub ya que son 1971.

- Lista de variables: https://github.com/Daniel1388/Towns_ML/raw/main/Resumen/Variables_Modelo_Clasificaci%C3%B3n.xlsx
- Importancia de variables en reducción de la dimensionalidad Clasificación: https://github.com/Daniel1388/Towns_ML/blob/main/Resultados/Variables_Importantes_selector_Clasificacion.csv
- Importancia de variables en reducción de la dimensionalidad Regresión: https://github.com/Daniel1388/Towns_ML/blob/main/Resultados/Variables_Importantes_selector_Regresion.csv
- Lista de variables XGBoost Clsificación: https://github.com/Daniel1388/Towns_ML/blob/main/Resultados/Variables_Importantes_XGBoost_Classificacion.csv
- Lista de variables XGBoost Regresión: https://github.com/Daniel1388/Towns_ML/blob/main/Resultados/Variables_Importantes_XGBoost_Regresion.csv
- Lista de variables por categoria: https://github.com/Daniel1388/Towns_ML/blob/main/Resultados/Variables_categorias.csv

B. Archivos de Github

Los Scripts del proyecto se encuentran en Github: https://github.com/Daniel1388/Towns_ML

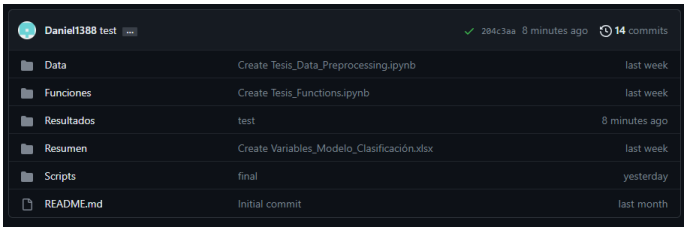

Ilustración 17. carpetas GitHub

C. Proporción de municipios por categoría para cada departamento

Municipio	Grandes Creciendo	Desacelerado	Emergentes	Esperanza Futura	Estáticos	Total Municipios
Bogotá	100%	0%	0%	0%	0%	1
Amazonas	50%	0%	50%	0%	0%	2
Cundinamarca	28%	3%	16%	32%	22%	116
La Guajira	27%	0%	53%	13%	7%	15
Vichada	25%	0%	75%	0%	0%	4
Antioquia	25%	6%	41%	19%	10%	125
Valle del Cauca	24%	29%	24%	10%	14%	42
Atlántico	22%	0%	30%	26%	22%	23
Risaralda	21%	0%	14%	36%	29%	14
Casanare	21%	0%	11%	32%	37%	19
Córdoba	20%	7%	30%	17%	27%	30
Meta	17%	10%	48%	14%	10%	29
Tolima	17%	4%	15%	30%	34%	47
Quindío	17%	17%	33%	17%	17%	12
Santander	14%	1%	41%	26%	17%	87
Cesar	12%	8%	60%	16%	4%	25
Sucre	12%	0%	46%	23%	19%	26
Caldas	11%	15%	30%	30%	15%	27
Huila	11%	0%	11%	43%	35%	37
Cauca	10%	0%	26%	33%	31%	42
Bolívar	9%	0%	35%	17%	39%	46
Caquetá	7%	0%	87%	7%	0%	15
Chocó	7%	10%	53%	17%	13%	30
Magdalena	7%	0%	30%	17%	47%	30
Boyacá	6%	2%	22%	30%	41%	123
Norte de Santander	5%	8%	15%	28%	45%	40
Nariño	3%	2%	22%	30%	44%	64
Putumayo	0%	0%	54%	15%	31%	13
Vaupés	0%	0%	33%	67%	0%	3
Arauca	0%	14%	29%	29%	29%	7
Guaviare	0%	0%	0%	25%	75%	4
Guainía	0%	0%	0%	0%	100%	1