TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA
IDENTIFICAR POTENCIALES COMPRADORES DE VIVIENDA ENTRE LOS BENEFICIARIOS DEL SISBEN IV

XIMENA MARULANDA WALLES

JOHN JAIRO BEDOYA SAENZ

Nota de Aceptación

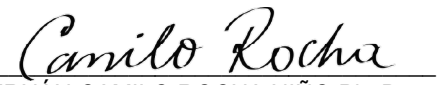
Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

Director
GUSTAVO ADOLFO ARTEAGA BOTERO

Jurado
DIEGO LUIS LINARES OSPINA

Jurado
JULIAN GIL GONZÁLEZ

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en CIENCIAS DE DATOS.



HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Cali, 12 febrero de 2024

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 13 febrero de 2024

Autor: XIMENA MARULANDA WALLES Y JOHN JAIRO BEDOYA SAENZ

Título del Trabajo de Grado: "TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA IDENTIFICAR POTENCIALES COMPRADORES DE VIVIENDA ENTRE LOS BENEFICIARIOS DEL SISBEN IV"

Director: DR. GUSTAVO ADOLFO ARTEAGA BOTERO

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

GUSTAVO ADOLFO ARTEAGA BOTERO
Director del Trabajo de Grado

Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: "Técnicas De Aprendizaje Automático Para Identificar Potenciales Compradores De Vivienda Entre Los Beneficiarios Del Sisben IV".

1. **ÉNFASIS:** N/A
2. **TIPO DE PROYECTO:** Aplicado
3. **ÁREA DE TRABAJO:** Vivienda y analítica de datos.
4. **ESTUDIANTE (S):** Ximena Marulanda Walles y John Jairo Bedoya Sáenz
5. **CORREO ELECTRÓNICO:** xmarulanda@javerianacali.edu.co y jjbedoya@javerianacali.edu.co
6. **DIRECCIÓN Y TELÉFONO:** Carrera 68 # 13b-61, apto 303A – 316 8647139 y Calle 8g # 51-75, apto 301 – 314 7853993 respectivamente.
7. **DIRECTOR:** Dr. Gustavo Adolfo Arteaga Botero.
8. **VINCULACIÓN DEL DIRECTOR (en la universidad):** Planta
9. **CORREO ELECTRÓNICO DEL DIRECTOR:** gustavo.arteaga@javerianacali.edu.co
10. **CO-DIRECTOR(ES) (Si aplica):** No aplica.
11. **GRUPO O EMPRESA QUE LO AVALA (Si aplica**): Alcaldía de Santiago de Cali, Departamento Administrativo de Planeación Distrital.
12. **OTROS GRUPOS O EMPRESAS:** No aplica
13. **PALABRAS CLAVE (al menos 5):** Aprendizaje automático, Vivienda, Sisbén, Ingresos familiares, calidad de vida.
14. **ODS QUE APLICA EL PROYECTO** (Agenda 2030): 11 Ciudades y Comunidades Sostenibles.
15. **FECHA DE INICIO (Desarrollo del proyecto):** 1/11/2022
16. **RESUMEN (máximo 400 palabras).**

La problemática asociada al elevado porcentaje de hogares encuestados en el Sisbén IV de la ciudad Cali – Valle del Cauca que no cuenta con vivienda propia y que desconocen su potencialidad como posibles compradores, es razón por la cual el objetivo de la presente investigación es aplicar técnicas de aprendizaje automático para identificar potenciales compradores de vivienda entre los beneficiarios del Sisbén IV en Santiago de Cali. A partir de la información suministrada por el Departamento Administrativo de Planeación Distrital de la ciudad para el año 2022 se desarrolló un modelo de aprendizaje automático que permita analizar los datos recolectados, identificar patrones y tendencias, y predecir con precisión qué personas tienen mayor probabilidad de convertirse en compradores de vivienda. El conocimiento generado permite respaldar la toma de decisiones eficaces de los organismos privados y gubernamentales del sector vivienda en lo relacionado con el planteamiento de políticas públicas y programas en general en beneficio a la población de estudio.

TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA IDENTIFICAR POTENCIALES COMPRADORES DE VIVIENDA ENTRE LOS BENEFICIARIOS DEL SISBEN IV

Ximena Marulanda Walles
John Jairo Bedoya Saenz

Anteproyecto del Proyecto Aplicado para optar al título de Magister en Ciencia de Datos

Director(a) Dr. GUSTAVO ADOLFO ARTEAGA BOTERO

FACULTAD DE INGENIERÍA Y CIENCIAS MAESTRÍA EN CIENCIA DE DATOS SANTIAGO DE CALI, FEBRERO 13 DE 2024

RESUMEN:

La problemática asociada al elevado porcentaje de hogares encuestados en el Sisbén IV de la ciudad Cali – Valle del Cauca que no cuenta con vivienda propia y que desconocen su potencialidad como posibles compradores, es razón por la cual el objetivo de la presente investigación es aplicar técnicas de aprendizaje automático para identificar potenciales compradores de vivienda entre los beneficiarios del Sisbén IV en Santiago de Cali. A partir de la información suministrada por el Departamento Administrativo de Planeación Distrital de la ciudad para el año 2022 se desarrolló un modelo de aprendizaje automático que permita analizar los datos recolectados, identificar patrones y tendencias, y predecir con precisión qué personas tienen mayor probabilidad de convertirse en compradores de vivienda. El conocimiento generado permite respaldar la toma de decisiones eficaces de los organismos privados y gubernamentales del sector vivienda en lo relacionado con el planteamiento de políticas públicas y programas en general en beneficio a la población de estudio.

TABLA DE CONTENIDO

pág.

INTRODUCCIÓN 7
1. DEFINICIÓN DEL PROBLEMA 8
1.1 PLANTEAMIENTO DEL PROBLEMA 8
1.2 FORMULACIÓN DEL PROBLEMA 9
2. OBJETIVOS DEL PROYECTO 10
2.1 OBJETIVO GENERAL 10
2.2 OBJETIVOS ESPECÍFICOS 10
2.3 RESULTADOS ESPERADOS 10
2.4 ALCANCE 10
3. JUSTIFICACIÓN 12
4. MARCO DE REFERENCIA 13
4.1 MARCO TEÓRICO 13
4.1.1 Ciencias de datos 14
4.1.2 Técnicas de aprendizaje automático 15
4.2 ANTECEDENTES 17
5. METODOLOGÍA 19
6. RESULTADOS 20
6.1 Recolección de datos precisos y actualizados de los hogares y familias pertenecientes al Sisbén IV en Santiago de Cali. 20
6.1.1 Conocimiento del modelo de negocio 21
6.1.2 Métodos de recolección de datos 23
6.1.3 Descripción de los datos 25
6.2 Desarrollo de un modelo de aprendizaje automático para predecir compradores de vivienda 37
6.2.1 Etapa 1: conocimiento de datos 37
6.2.2 Etapa 2: Limpieza de los datos 43
6.2.3 Etapa 3: Definición y entrenamiento del modelo 46

6.3 Evaluación del rendimiento del modelo de aprendizaje automático para garantizar su precisión y eficacia en la identificación de potenciales compradores de vivienda en la población del Sisbén IV en Santiago de Cali. 50
7. CONCLUSIONES 53
REFERENCIAS BIBLIOGRÁFICAS 55

Lista de Tablas

pág.

Tabla 1. Antecedentes demográficos. Estado civil 25
Tabla 2. Salud. Seguridad social. 27
Tabla 3. Mercado laboral. Pensión. 28
Tabla 4. Mercado laboral. Tipo de actividad 29
Tabla 5. Nivel educativo de la población. 30
Tabla 6. Población con bien raíz. 31
Tabla 7. Tipo de vivienda. 32
Tabla 8. Tipo de estrato de energía. 33
Tabla 9. Gasto total promedio. 34
Tabla 10. Ingreso total promedio. 35

Lista de figuras

pág.

Figura 1. Antecedentes demográficos. Estado civil.....26
Figura 2. Salud. Seguridad social. 27
Figura 3. Mercado laboral. Pensión. 28
Figura 4. Mercado laboral. Tipo de actividad. 29
Figura 5. Nivel educativo de la población. 30
Figura 6. Población con bien raíz. 31
Figura 7. Tipo de vivienda. 32
Figura 8. Tipo de estrato de energía. 33
Figura 9. Variable dependiente C 40
Figura 10. Balanceo de datos 45
Figura 11. Estructura del árbol de decisión. 49

INTRODUCCIÓN

En Colombia, el déficit habitacional es un indicador desarrollado por el DANE en el 2009, donde publicó el documento "Metodología Déficit de Vivienda" en el cual se expuso la forma en la que, usando los datos del Censo General de 2005, se midió el Déficit de vivienda en el país. Dicho indicador brinda la posibilidad de identificar la brecha existente entre el número total de hogares y aquellos que viven en condiciones adecuadas [1]

Desde el punto de vista cuantitativo, la Encuesta Nacional de Vida (ECV) para el año 2020 evidenció que el déficit correspondía a un 16,8%, mientras que para el 2021 se incrementó a 18,3% en el Valle del Cauca [2], mostrando así un crecimiento preocupante para la región; igualmente se ha permitido observar diferentes fenómenos que dan cuenta de los procesos del crecimiento demográfico, del desarrollo urbano y de las condiciones sociales con que se están complejizando en las ciudades del país; lo específico de las regiones parece afectar de manera diferencial y, lo económico, adquiere magnitud que hace de los procesos condiciones especiales por ser vistas en detalle.

A nivel nacional, existe un sistema de información de identificación y clasificación de potenciales beneficiarios para los programas sociales conocido como SISBÉN, el cual es un sistema informático creado por el Gobierno Nacional para identificar y clasificar a hogares, familias y personas según sus condiciones de vida, utilizando una encuesta realizada en el hogar como fuente de información. Dicha clasificación es utilizada para focalizar la inversión social, sin embargo, por sí mismo no otorga el acceso a los programas sociales.

En el caso de los programas sociales relacionados con la vivienda, existe en Colombia el programa de subsidios Mi Casa Ya creado desde el 2015 a través del cual se apoya a los colombianos más vulnerables para la adquisición de este tipo de bienes [3]. No obstante, el acceso a dichos subsidios dependerá de la disponibilidad y el puntaje de priorización, así como a las condiciones de vulnerabilidad de la población solicitante.

Finalmente, el presente trabajo genera profundidad sobre aspectos asociados a la vivienda para población SISBEN IV, teniendo en cuenta las técnicas de aprendizaje automático como principal herramienta, que da como resultado aquellos posibles compradores de vivienda teniendo en cuenta sus condiciones de estabilidad laboral y de ingresos como principales variables de estudio, entre otras. Lo anterior aporta a la población y al gobierno la base para futuras políticas públicas, subsidios y proyectos de vivienda dignos enfocados a su bienestar y condición.

1. DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

La necesidad de los usuarios pertenecientes al SISBÉN IV para mejorar su calidad de vida en cuanto vivienda se refiere, es parte fundamental de su supervivencia como sociedad, siendo esto, además, un derecho constitucional establecido en el artículo 51 de la Constitución Política de Colombia donde indica que: "todos los colombianos tienen derecho a vivienda digna. El Estado fijará las condiciones necesarias para hacer efectivo este derecho y promoverá planes de vivienda de interés social, sistemas adecuados de financiación a largo plazo y formas asociativas de ejecución de estos programas de vivienda" [4], por ello, las estructuras creadas por el gobierno para el periodo 2018-2022 en los diferentes programas de vivienda son [5]:

- **Mi Casa Ya**: Este programa beneficia a hogares con ingresos inferiores a 4 salarios mínimos mensuales legales vigentes (SMMLV). Ofrece subsidios monetarios de 20 a 30 SMMLV, dependiendo de los ingresos, y cobertura a la tasa de interés de 4 a 5 puntos porcentuales. Además, proporciona ayuda específica para viviendas de interés prioritario y social en diversas gamas de ingresos, con extensiones específicas para aglomeraciones urbanas de más de un millón de habitantes.

- **Casa Digna Vida Digna**: Este programa tiene como objetivo mejorar la calidad de vida de los usuarios a través de intervenciones en viviendas y su entorno. Se centra en tres áreas: mejora del entorno y construcción de equipamientos sociales, titulación de viviendas y mejoras internas en las viviendas, como construcción de baños, cocinas y reparaciones estructurales.

- **Vivienda de Interés Prioritario para Ahorradores (VIPA)**: Dirigido a personas con ingresos entre 1 y 2 SMMLV, este programa facilita la adquisición de vivienda nueva de interés prioritario urbana. Permite que los hogares posean una casa propia sin comprometer más del 30% de sus ingresos en el pago de las cuotas del crédito hipotecario, lo que implica cuotas mensuales accesibles, incluso para hogares con ingresos de un salario mínimo mensual.

Jurídicamente está constituido en la Sentencia T-409/13 la Corte Constitucional, la cual explica que: "el derecho a la vivienda está íntimamente relacionado con el derecho a la vida en

condiciones dignas y, como tal, es obligación del Estado ofrecer proyectos de vivienda o solución de vivienda a los ciudadanos, ya sea de forma directa o por intermedio de los particulares, procurando garantizar la materialización efectiva del derecho en cuanto: a) la seguridad jurídica de la tenencia; b) disponibilidad de servicios, materiales, facilidades e infraestructura; c) gastos soportables; d) habitabilidad; e) asequibilidad; f) lugar y g) adecuación cultural" [6].

Sin embargo, según datos suministrados por el Departamento Administrativo de Planeación Distrital, para el 2019, el 76% de los hogares encuestados en el SISBÉN IV en la ciudad de Cali no cuentan con vivienda propia ya que pertenecen a la variable tenencia (arriendo o subarriendo, propia pagando, posesión sin título, entre otros) excluyendo aquellos que son propietarios, mostrando así la falta de divulgación y aplicación de estos programas a este sector de la población, generando un alto déficit de vivienda habitacional [7].

Adicionalmente, en el marco del programa Mi Casa Ya, es importante tener en cuenta que para recibir el apoyo de los subsidios otorgados se tomará como referencia el parámetro de categorización de del Sisbén IV, contemplando variables como el ingreso familiar y la capacidad de pago, así como el valor del inmueble elegido, entre otros.

1.2 FORMULACIÓN DEL PROBLEMA

Este estudio se basa en la realización de un modelo predictivo que responda la siguiente pregunta de investigación:

¿Cómo pueden aplicarse las técnicas de aprendizaje automático para la identificación de potenciales compradores de vivienda entre las personas pertenecientes al Sisbén IV en Santiago de Cali?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Aplicar técnicas de aprendizaje automático para la identificación de potenciales compradores de vivienda entre las personas pertenecientes al Sisbén IV en Santiago de Cali.

2.2 OBJETIVOS ESPECÍFICOS

- Recolectar datos precisos y actualizados de los hogares y familias pertenecientes al Sisbén IV en Santiago de Cali.
- Desarrollar un modelo de aprendizaje automático que permita analizar los datos recolectados, identificar patrones y tendencias, y predecir con precisión qué personas tienen mayor probabilidad de convertirse en compradores de vivienda.
- Evaluar el rendimiento del modelo de aprendizaje automático para garantizar su precisión y eficacia en la identificación de potenciales compradores de vivienda en la población del Sisbén IV en Santiago de Cali.

2.3 RESULTADOS ESPERADOS

El presente estudio busca identificar aquellos hogares (individuos o grupo) que tienen estabilidad laboral, ya sea como independientes o contratos laborales por obra o labor, que podrían ser candidatos para acceder a compra de vivienda sea nueva o usada en las condiciones actuales del mercado. Sin embargo, se esperaría que el alcance de éste pueda ser aún mayor y de gran utilidad para entes gubernamentales o sector privado en el desarrollo de futuras políticas públicas enfocadas a este segmento de la población, brindando subsidios o ayudas especiales para compra de vivienda o constructores puedan realizar proyectos de vivienda dignos para este sector de la población y así mejorar su calidad de vida.

2.4 ALCANCE

La investigación se centra en realizar un análisis multivariado de datos que permita segmentar la población de Santiago de Cali encuestada a través del Sisbén, es decir, en clasificar la población por un conjunto de características similares. Como segunda instancia se busca identificar los grupos candidatos para acceder a la compra de vivienda propia, ya sea nueva o usada según las

condiciones del mercado y del hogar a través de un árbol de decisión que permita generar un modelo de predicción.

Las dificultades identificadas se relacionan con la captura de información y la inconsistencia de los datos en la encuesta SISBÉN, dado que se presentan datos faltantes en el componente de "capacidad de generación de ingresos" siendo esta una de la característica más relevante para conocer la estabilidad laboral del hogar.

3. JUSTIFICACIÓN

El Sisbén es un sistema de identificación que nació de la necesidad de reconocer y ubicar la población más vulnerable del país, para direccionar y focalizar los recursos de los programas sociales, de manera eficiente y eficaz. Desde su puesta en marcha en 1995 hasta la fecha, se han establecido cuatro versiones, las cuales responden a modificaciones en el enfoque, que conlleva cambios en la estructura del cuestionario y en la manera en que se establece el puntaje, que permite ser acreedores de beneficios sociales y económicos. El Sisbén IV articula las versiones anteriores y permite identificar personas en exclusión social y productiva, así como la capacidad de la generación de ingresos de los hogares según sus condiciones sociales y económicas.

La vivienda es un derecho fundamental y una necesidad básica para todas las personas, sin embargo, muchas veces el acceso a una vivienda digna se ve limitado por factores económicos y sociales. En este contexto, el Sisbén IV es una herramienta importante para identificar a las personas y familias más vulnerables y brindarles acceso a programas sociales y de vivienda.

A pesar de lo anterior, actualmente el proceso de identificación de potenciales compradores de vivienda en la población del Sisbén IV en Santiago de Cali no cuenta con una herramienta que le permita gestionar la gran cantidad de información que genera. En este sentido, la aplicación de técnicas de aprendizaje automático puede ser una solución eficaz y precisa para identificar a las personas con mayor probabilidad de convertirse en compradores de vivienda, lo que a su vez permitiría a las autoridades y empresas del sector enfocar sus esfuerzos y recursos de manera más efectiva y eficiente.

Por lo tanto, la justificación de este estudio radica en la necesidad de mejorar la identificación de potenciales compradores de vivienda en la población del Sisbén IV en Santiago de Cali, a través de la aplicación de técnicas de aprendizaje automático, con el fin de garantizar el acceso a la vivienda digna como un derecho fundamental para todas las personas.

4. MARCO DE REFERENCIA

4.1 MARCO TEÓRICO

Víctor Brodersohn (1999), en el documento *"Focalización de Programas de Superación de la Pobreza"* establece que la focalización de las políticas y programas de superación de la pobreza surge como respuesta al incremento del nivel de pobreza en América Latina, en el marco de la crisis económica y social de los años 80s y, en el nuevo papel del Estado que propenden por la privatización de empresas y organismos públicos, en la lógica de la modernización de la gestión pública [8].

> "La idea de focalización intenta responder en buena medida a este imperativo de modernizar y racionalizar la política social estatal... Doble desafío, combinar la modernización de la política social estatal con la democratización de la demanda social mediante el estímulo a la gestión ciudadana." [9]

En cuanto a la definición de focalización establece, que es una modalidad de intervención pública que se direcciona a asegurar que un programa/proyecto se dirija en exclusividad a una población objetivo, por lo tanto, la focalización parte de una formulación de estrategias de intervención pública centradas en la eficacia [8].

Con relación a los aspectos técnicos y operacionales de la focalización, Raczynski, Dagmar en el documento "*Focalización de Programas Sociales: Lecciones de la Experiencia Chilena"* a partir de las conclusiones de un estudio basado en la experiencia de nueve países latinoamericanos (incluido Chile), identifica tres dimensiones centrales [10]:

Mecanismos e instrumentos para seleccionar los beneficiarios: Distinguen tres mecanismos: si priorizan a demanda, que corresponde a beneficiarios directos; la oferta, que características de las entidades y/o organismos que ofrecen prestaciones, servicios o subsidios; áreas geográficas, donde la ubicación denota pobreza por lo tanto las personas y familias ubicadas en dichas zonas tienen el derecho a recibir el servicio o subsidio. Pero otro, lado establece que, la focalización puede definirse "a partir de la aplicación de un instrumento específico de selección de beneficiarios, o a través del propio diseño del programa. En este último caso, el programa se diseña de tal forma que la entrada al mismo sea "atractiva" sólo para los beneficiarios, ya sea por su modalidad específica de funcionamiento -como gratuidad o alternativas de copago…" [10, p. 219].

Costos de la focalización: Relaciona dos tipos de costos: Costo administrativo, que se subscribe a las tareas de elaboración, aplicación, actualización y control de los instrumentos de selección (fichas, cuestionarios, entrevistas domiciliarias, mapas de pobreza, entre otros). Costos psicosociales: se relaciona con el costo de la estigmatización social y de estimular la dependencia de la población de los programas del Estado, que coarta y limita la iniciativa y creatividad de los beneficiarios para establecer soluciones a sus problemas.

Errores de focalización: Se identificaron dos errores: Error de Inclusión, el cual se presenta cuando se filtra el beneficio hacia segmentos sociales excluidos del grupo destinatario; Error de Exclusión, cuando se establece una brecha entre el universo del grupo destinatario del programa y los integrantes efectivamente cobijado por él.

Una de las conclusiones resultado del estudio, establece que, para focalizar de manera efectiva un programa social, es fundamental "disponer de información fidedigna sobre la realidad que el programa se propone modificar" el tipo de información debe trascender la clasificación de pobre y asociarse con los objetivos del programa y las modalidades y mecanismos de intervención que privilegia. Lo que conlleva a la eficiencia de los recursos y eficacia en los resultados [10].

4.1.1 Ciencias de datos

La ciencia de datos es esencial para explorar datos, generar conocimiento y alcanzar objetivos como la identificación de patrones y comportamientos para la toma de decisiones y predicciones. Su crecimiento ha sido notable debido al acceso extendido a grandes volúmenes de datos y su procesamiento en tiempo real. Este campo requiere técnicas avanzadas para abordar desafíos prácticos como la escalabilidad, robustez ante errores y adaptabilidad a modelos dinámicos. La colaboración interdisciplinaria, incluyendo áreas como computación, estadística, matemáticas e ingeniería, es crucial para desarrollar nuevos algoritmos, técnicas y estructuras que faciliten la captura, almacenamiento y procesamiento eficiente de datos [11].

Los datos son hechos recopilados en una base de datos, y los patrones son expresiones que describen subconjuntos de esos datos. La búsqueda de modelos y patrones en una base de datos implica un proceso iterativo e interactivo. Los patrones descubiertos deben ser válidos, novedosos y potencialmente útiles. Se requiere definir medidas cuantitativas para evaluar la precisión, utilidad y beneficio de los patrones. La minería de datos utiliza técnicas para encontrar patrones y establecer medidas de interés que consideren la validez, utilidad y simplicidad. El objetivo final es incorporar el conocimiento obtenido en sistemas reales, tomar decisiones informadas o registrar la información para su posterior uso [11].

La inteligencia artificial (IA) es esencial en la minería de datos y el análisis de patrones. Su importancia radica en su capacidad para automatizar procesos, mejorar la precisión en la identificación de patrones, gestionar grandes volúmenes de datos de manera eficiente y reconocer estructuras complejas. La adaptabilidad y flexibilidad de la IA, su búsqueda constante de optimización y su aplicabilidad en tiempo real la destacan como un catalizador clave en la ciencia de datos [12]. De acuerdo con esto, la IA enriquece la capacidad de los sistemas para procesar, comprender y utilizar información de manera avanzada y efectiva, ofreciendo soluciones más precisas y relevantes, incluyendo la identificación de potenciales compradores de vivienda.

En esto contexto, Machine Learning es una rama de la Inteligencia Artificial que se centra en crear algoritmos con la capacidad de aprender sin requerir una programación explícita. Elimina la necesidad de programar escenarios y excepciones específicos, ya que el algoritmo aprende automáticamente al ser alimentado con grandes cantidades de datos. Hay dos tipos de aprendizajes: el supervisado y el no supervisado [12].

4.1.2 Técnicas de aprendizaje automático

El aprendizaje automático, como rama de la inteligencia artificial, desempeña un papel fundamental al permitir la interpretación de patrones complejos presentes en una amplia variedad de fenómenos en disciplinas como la política, las finanzas, la biología, la ciencia y muchas otras [11]. Esta capacidad de análisis de patrones posibilita la predicción de diversas situaciones, proporcionando herramientas valiosas para la toma de decisiones informadas en contextos diversos.

De acuerdo con la definición de Mitchell, el aprendizaje automático, también conocido como machine learning, se centra en la construcción de programas informáticos capaces de mejorar automáticamente a través de la experiencia [12]. En otras palabras, implica procesos automatizados que extraen patrones significativos de los datos, permitiendo que los sistemas aprendan y se adapten de manera autónoma con el tiempo.

Dentro de este campo, existen dos tipos fundamentales de aprendizaje que se emplean para diversos propósitos: las técnicas de aprendizaje supervisado y las técnicas de aprendizaje no supervisado [13].

Aprendizaje supervisado: estas técnicas de aprendizaje buscan encontrar el mejor clasificador posible para un problema determinado, en donde existe un algoritmo considerado el responsable de encontrar dicha solución, denominado algoritmo de clasificación que a través de la inferencia

encuentra un modelo asociado a cada ejemplo de entrada considerando la asociación entre las variables X e Y. En este tipo de aprendizaje existen unas técnicas de clasificación que se enfocan en predecir una respuesta cualitativa por medio del análisis de datos y el reconocimiento de patrones, siendo las más utilizadas las siguientes [13]:

Dentro del Aprendizaje Supervisado, varias técnicas de clasificación se destacan por su eficacia en la predicción de variables cualitativas. Entre ellas, se encuentran [13]:

Regresión logística: Utilizada para modelar la probabilidad de que una variable dependiente pertenezca a una categoría particular. Es efectiva en problemas de clasificación binaria.

Análisis discriminante lineal: Se enfoca en encontrar la combinación lineal de características que maximiza la separación entre múltiples clases. Es útil cuando las clases tienen distribuciones normales.

K vecinos más cercanos (KNN): Este método clasifica un punto de datos según la mayoría de las clases de sus vecinos más cercanos. Es versátil y se adapta bien a diferentes tipos de datos y estructuras.

Árboles de decisión: Representan una estructura de decisión en forma de árbol, donde cada nodo representa una decisión basada en características particulares. Son fáciles de entender y visualizar.

Redes Neuronales: Modelan el funcionamiento del cerebro humano para el reconocimiento de patrones. Consisten en capas de nodos interconectados que aprenden y adaptan sus pesos a través del entrenamiento.

Máquinas Vectoriales de Soporte (SVM): Buscan encontrar el hiperplano óptimo que mejor separa las clases en un espacio multidimensional. Son eficientes en problemas de clasificación tanto lineales como no lineales.

Estas técnicas se seleccionan según la naturaleza del problema y la naturaleza de los datos. La elección adecuada de la técnica de clasificación depende de diversos factores, como la complejidad del problema, la cantidad de datos disponibles y la interpretabilidad del modelo resultante. En conjunto, estas herramientas ofrecen un amplio abanico de opciones para abordar problemas de clasificación en el Aprendizaje Supervisado.

Aprendizaje no supervisado: estas técnicas consisten en aprender sin la presencia de un maestro [13], es decir, consiste en que el algoritmo descubre por sí mismo las características, patrones o categorías en los datos de entrada para obtener de forma codificada los datos de salida [14]. De

alguna manera, en este tipo de aprendizaje, la tarea consiste en encontrar patrones en los datos o predecir nuevos valores, lo cual puede entenderse como el aprendizaje basado en un modelo probabilístico de los datos en donde a la máquina no se le proporciona ningún conjunto de salidas deseadas, a pesar de lo cual es capaz de estimar un modelo que represente la distribución de probabilidad para una nueva entrada [15].

El desarrollo de un modelo en Machine Learning involucra diversas fases, cada una crucial para asegurar la eficacia y la precisión del algoritmo. Estas fases, que abarcan distintos ámbitos de aplicación, se detallan a continuación [12]:

Fases de Desarrollo:

1. Fase de Entrenamiento:

En esta etapa inicial, se dispone de una gran cantidad de datos. Se selecciona una porción de estos datos para entrenar el algoritmo, proporcionándole información suficiente para que identifique patrones relevantes. Durante el entrenamiento, el algoritmo ajusta sus parámetros y estructuras internas con el objetivo de hacer predicciones precisas [12].

2. Fase de Prueba:

La porción restante de los datos se reserva para la fase de prueba. Aquí, el algoritmo se somete a interrogantes y evaluaciones para determinar su nivel de aprendizaje y la precisión de sus respuestas. Si las predicciones no coinciden con los resultados esperados, se deben realizar ajustes, como agregar más datos de entrenamiento o modificar el método utilizado. Un rendimiento del 80% a 90% de respuestas correctas generalmente indica un buen grado de aprendizaje, permitiendo validar la utilidad del algoritmo [12].

4.2 ANTECEDENTES

Con la Ley 388 de 1991, se definieron las condiciones para que el ordenamiento territorial adquiriera dimensiones de la función social de la propiedad y las superficies [16]. En lo urbano, el POT se encargó de recopilar diferentes insumos que, a manera de diagnóstico, crearon insumos para observar las formas del desarrollo. Desde este instrumento, el déficit habitacional se asoció a algunas de las variables evaluadas en el formulario SISBÉN IV, en el cual, se determina desde el orden cuantitativo las ideas de infraestructuras faltante, es decir, lo mínimo requerido en una propiedad para que ésta sea habitada, donde incluye al menos un baño completo (ducha, lavamanos e inodoro), una cocina (mesón en aluminio, lavaplatos, y estufa a gas) y conexiones de energía y acueducto para futuras remodelaciones, por debajo de esto indica que la vivienda no

tiene lo básico para ser habitado y, es en estas condiciones, donde muchos usuarios pertenecientes al SISBÉN IV viven actualmente.

Para el año 2003 en Costa Rica, por ejemplo, la falta de trabajo en equipo de los entes gubernamentales ha traído consecuencias en cuanto a las políticas públicas en lo que a vivienda respecta en ese país. Cada organismo de control cumple con sus propios informes, pero no generan una alianza con las entidades relacionadas de esta área y no se logra prevenir situaciones que ponen en riesgo la calidad de vida de sus habitantes. Dado lo anterior, se indica lo siguiente: "En este trabajo se presenta una metodología para la creación de Indicadores para la toma de decisiones en el sector vivienda, utilizando técnicas de minería de datos. Se analizaron las necesidades de información del sector vivienda, la información disponible, las técnicas de minería de datos existentes, se determinaron las más apropiadas y se finalizó con un prototipo utilizando el programa SQL Server 2000" [17].

En el ámbito inmobiliario de Madrid, España, se llevó a cabo un proyecto de minería de datos denominado "Minería de Datos en portal inmobiliario". Este proyecto utilizó datos extraídos del portal Idealista para mejorar el acceso a los resultados de búsqueda de viviendas en venta o alquiler. Se empleó la técnica de aprendizaje automático no supervisado conocida como mapa auto-organizativo, utilizando atributos como precio, localidad y número de habitaciones para agrupar las viviendas en un mapa bidimensional [18].

Inicialmente, se recopilaron todas las viviendas de Madrid a través del API de Idealista y se aplicó un filtro seleccionado por el usuario. Luego, se entrenó el mapa, calculando repetidamente la distancia euclídea entre las viviendas y las celdas del mapa. Este proceso permitió asignar cada vivienda a una celda específica en el mapa. Posteriormente, el mapa bidimensional resultante facilitó al usuario la navegación y selección de viviendas de interés, reduciendo la cantidad total de resultados y optimizando la experiencia de búsqueda [18].

5. METODOLOGÍA

Para este caso de estudio se usó: Selección, limpieza y transformación de datos, minería de datos y evaluación y presentación de resultados.

Fases del proyecto:

- **Selección de datos:** Obtención de las fuentes externas a través del Departamento Administrativo de Planeación Distrital.
- **Limpieza de datos:** Se utilizarán herramientas para la detección de duplicados, detección de valores atípicos y técnicas para solucionar datos faltantes.
- **Transformación de datos:** Creación de nuevos atributos que permiten conocer factores como la parte social y económica, de los usuarios beneficiarios del programa de SISBÉN IV.
- **Minería de datos:** Búsqueda y descubrimiento de patrones insospechados aplicando tareas de hallazgo como clasificación, clustering, patrones secuenciales, asociaciones, entre otros.
- **Evaluación y presentación de resultados:** Se requiere interpretar y evaluar los patrones descubiertos con el fin de consolidar el conocimiento descubierto e incorporarlo en otro sistema para posteriores acciones o para confrontarlo con conocimiento previamente descubierto. La visualización de datos se realizará por medio de gráficos, mapas y tablas.

6. RESULTADOS

6.1 Recolección de datos precisos y actualizados de los hogares y familias pertenecientes al Sisbén IV en Santiago de Cali.

Según datos suministrados por el Departamento Administrativo de Planeación Distrital, para el 2019, el 76% de los hogares encuestados en el SISBÉN IV en la ciudad de Cali no cuentan con vivienda propia ya que pertenecen a la variable tenencia (arriendo o subarriendo, propia pagando, posesión sin título, entre otros) excluyendo aquellos que son propietarios, mostrando así la falta de divulgación y aplicación de estos programas a este sector de la población, generando un alto déficit de vivienda habitacional.

Adicionalmente, en el marco del programa Mi Casa Ya, es importante tener en cuenta que para recibir el apoyo de los subsidios otorgados se tomará como referencia el parámetro de categorización de del Sisbén IV, contemplando variables como el ingreso familiar y la capacidad de pago, así como el valor del inmueble elegido, entre otros.

La pregunta de investigación de este trabajo plantea cómo se pueden aplicar técnicas de aprendizaje automático para identificar potenciales compradores de vivienda entre las personas pertenecientes al Sisbén IV en Santiago de Cali.

El Sisbén IV es un sistema de clasificación socioeconómica utilizado en Colombia para identificar y priorizar a las personas y familias más vulnerables, con el fin de ofrecerles programas de asistencia y beneficios sociales. Por lo tanto, identificar a potenciales compradores de vivienda dentro de este grupo es un desafío.

El aprendizaje automático, también conocido como machine learning, es un campo de la inteligencia artificial que se enfoca en desarrollar algoritmos y modelos capaces de aprender y tomar decisiones sin ser programados explícitamente. Estas técnicas pueden ser aplicadas en diversos campos, incluyendo el análisis de datos y la predicción.

En el contexto de la pregunta de investigación de este trabajo, se propone utilizar técnicas de aprendizaje automático para analizar datos relacionados con las personas pertenecientes al Sisbén IV en Santiago de Cali y determinar cuáles de ellas tienen un mayor potencial para

convertirse en compradores de vivienda. Esto implica recopilar y analizar datos demográficos, socioeconómicos y otros factores relevantes que puedan influir en la capacidad y disposición de una persona para comprar una vivienda.

Al respecto la técnica de aprendizaje automático que se utiliza incluye la utilización de algoritmos de clasificación para etiquetar a las personas del Sisbén IV como posibles compradores de vivienda o no. Esto se basa en datos históricos y características relevantes, como ingresos, historial crediticio, edad, estado civil, entre otros.

6.1.1 Conocimiento del modelo de negocio

Se ha avanzado en el conocimiento del modelo de negocio, en este caso, a partir de dos elementos fundamentales: el sistema Sisbén y el segmento de la vivienda en el contexto de los derechos fundamentales de las personas en Colombia.

El Sisbén es un sistema de identificación que se implementó en Colombia con el propósito de reconocer y ubicar a la población más vulnerable del país. Su objetivo principal es dirigir y focalizar de manera eficiente y eficaz los recursos de los programas sociales hacia quienes más lo necesitan. Desde su lanzamiento en 1995, el Sisbén ha experimentado varias modificaciones y actualizaciones, dando lugar a cuatro versiones diferentes.

El Sisbén IV es la última versión del sistema y ha sido diseñado para articular y mejorar las versiones anteriores. Esta versión ha introducido cambios significativos en el enfoque y la metodología de evaluación, lo que ha llevado a modificaciones en la estructura del cuestionario y en la forma en que se establece el puntaje.

El puntaje del Sisbén es un indicador numérico asignado a cada hogar o persona evaluada, el cual se calcula sobre la base de una serie de variables socioeconómicas. Este puntaje determina el nivel de vulnerabilidad y exclusión social de una persona o familia, así como su capacidad para generar ingresos. Cuanto más bajo sea el puntaje, mayor será el nivel de vulnerabilidad y, por lo tanto, mayor será la elegibilidad para acceder a beneficios sociales y económicos.

El Sisbén IV se ha enfocado en identificar a las personas en exclusión social y productiva, lo que significa que no solo se tienen en cuenta las condiciones socioeconómicas, sino también la capacidad de generar ingresos de los hogares. Esto permite tener una visión más completa de la

situación de cada individuo o familia y facilita la implementación de políticas y programas de apoyo específicos para su situación particular.

En el contexto de la pregunta de investigación planteada, el uso de técnicas de aprendizaje automático para la identificación de potenciales compradores de vivienda entre las personas pertenecientes al Sisbén IV en Santiago de Cali busca aprovechar la información recopilada en el sistema, como el puntaje del Sisbén y otras variables socioeconómicas, para desarrollar modelos predictivos y personalizados que puedan identificar a aquellas personas con mayor capacidad y disposición para adquirir una vivienda. Esta aplicación del aprendizaje automático podría contribuir a la generación de oportunidades de vivienda y al mejoramiento de las condiciones de vida de las personas pertenecientes al Sisbén IV en Santiago de Cali.

En cuanto a la vivienda es un derecho fundamental reconocido a nivel internacional y una necesidad básica para todas las personas. Sin embargo, el acceso a una vivienda digna a menudo se ve limitado por una variedad de factores económicos y sociales, como ingresos bajos, falta de acceso a crédito, discriminación, falta de viviendas asequibles y otros obstáculos.

En este contexto, el Sisbén IV juega un papel crucial al ser una herramienta importante para identificar a las personas y familias más vulnerables. A través de la evaluación socioeconómica que realiza, el Sisbén IV permite determinar el nivel de necesidad y exclusión social de los individuos y hogares. Esto facilita la focalización y direccionamiento de los recursos de los programas sociales y de vivienda hacia aquellos que más lo necesitan.

A pesar de la importancia del Sisbén IV en la identificación de personas y familias más vulnerables y su potencial acceso a programas de vivienda, actualmente el proceso de identificación de potenciales compradores de vivienda dentro de esta población en Santiago de Cali enfrenta desafíos en la gestión de la gran cantidad de información generada.

El Sisbén IV recopila datos socioeconómicos de una amplia población, lo que implica un volumen significativo de información que debe ser procesada y analizada de manera eficiente. El manejo manual de esta información puede ser complejo y propenso a errores, lo que dificulta la identificación precisa de aquellos individuos y hogares con mayor probabilidad de convertirse en compradores de vivienda.

En este contexto, las técnicas de aprendizaje automático pueden ofrecer una solución eficaz y precisa para gestionar y analizar la gran cantidad de datos generados por el Sisbén IV. Estas

técnicas permiten desarrollar algoritmos y modelos que aprenden de los datos existentes y pueden realizar predicciones y clasificaciones en función de patrones identificados en los datos.

Aplicar técnicas de aprendizaje automático en el contexto de la identificación de potenciales compradores de vivienda entre la población del Sisbén IV en Santiago de Cali implicaría utilizar algoritmos de clasificación y análisis predictivo. Estos algoritmos podrían analizar variables como ingresos, historial crediticio, edad, estado civil, entre otros, para identificar patrones que indiquen una mayor probabilidad de que una persona perteneciente al Sisbén IV se convierta en un comprador de vivienda.

Al utilizar técnicas de aprendizaje automático, las autoridades y empresas del sector inmobiliario podrían optimizar sus esfuerzos y recursos, enfocándolos de manera más efectiva en aquellos individuos y hogares con una alta probabilidad de ser compradores de vivienda. Esto permitiría diseñar estrategias de promoción y financiamiento más personalizadas, así como ofrecer opciones de vivienda más adecuadas a las necesidades y capacidades de cada persona.

Es importante destacar que la aplicación de técnicas de aprendizaje automático requiere contar con datos de calidad y actualizados, así como garantizar la privacidad y la ética en el manejo de la información personal. Además, estos modelos deben ser validados y ajustados de manera continua para asegurar su precisión y efectividad en la identificación de potenciales compradores de vivienda dentro de la población del Sisbén IV en Santiago de Cali.

6.1.2 Métodos de recolección de datos

Se tuvo acceso a la base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali.

Análisis de las bases existentes

En la primera etapa del programa para realizar aprendizaje supervisado, se enfoca en el entendimiento de los datos antes de proceder con cualquier preprocesamiento. A continuación, se detalla el análisis que se realizó en esta etapa:

Cantidad de registros: Es importante conocer la cantidad total de registros en el conjunto de datos. Esto proporciona una idea del tamaño de la muestra y la cantidad de instancias con las que se trabajará.

Atributos por registro: Se debe determinar la cantidad de atributos o características que están presentes en cada registro. Esto ayuda a comprender la complejidad del conjunto de datos y la información disponible para el análisis.

Tipos de atributos: Es necesario identificar el tipo de cada atributo. Los atributos pueden ser numéricos (continuos o discretos), categóricos, binarios u otros tipos de datos. Esta información es esencial para seleccionar las técnicas de preprocesamiento y modelado adecuadas.

Medidas de centralidad: Calcular medidas de centralidad, como la media o la mediana, para cada atributo numérico proporciona información sobre los valores típicos o representativos en el conjunto de datos. Esto ayuda a comprender las características y distribuciones de los atributos.

Medidas de dispersión o desviación: Calcular medidas de dispersión, como la desviación estándar o el rango, para los atributos numéricos puede revelar la variabilidad de los datos. Esto es útil para identificar atributos con alta variabilidad o baja dispersión, lo que puede ser relevante para el análisis posterior.

Además de los puntos anteriores, se mencionan algunos pasos previos al preprocesamiento de datos:

Datos faltantes: Detectar si hay registros que contienen valores faltantes en alguna columna o atributo. Esto es importante para determinar si es necesario aplicar técnicas de imputación o eliminación de datos faltantes.

Datos atípicos: Identificar la presencia de valores atípicos o anómalos en los atributos. Los valores atípicos pueden afectar el rendimiento del modelo y deben ser tratados adecuadamente, ya sea eliminándolos o aplicando técnicas de transformación.

Desbalance de clases: Determinar si existe un desbalance significativo entre las diferentes clases o categorías en el conjunto de datos. El desbalance de clases puede afectar la capacidad del modelo para aprender correctamente y puede requerir técnicas de muestreo o ponderación para abordar este problema.

Por lo tanto, la primera etapa del programa se enfoca en comprender a fondo los datos antes de realizar cualquier modificación. Esto implica conocer la estructura del conjunto de datos, calcular medidas de centralidad y dispersión, identificar atributos redundantes, así como detectar datos faltantes, atípicos y desbalance de clases. Este análisis inicial sienta las bases para el preprocesamiento y modelado subsiguiente.

6.1.3 Descripción de los datos

A continuación, se presenta la descripción de los datos de la muestra a la cual se accedió para el desarrollo del trabajo. En total se obtuvieron 389.980 registros, considerando los siguientes antecedentes demográficos:

La población de Santiago de Cali, según los datos obtenidos, se caracteriza por ser diversa en términos de estado civil. La mayoría de la población se encuentra en la categoría de "Soltero(a)," lo que sugiere una población relativamente joven o una alta proporción de personas que no han formalizado una relación conyugal. Aproximadamente una cuarta parte de la población está en "Unión libre," lo que indica una proporción significativa de personas que conviven en relaciones no matrimoniales formalizadas. Además, un porcentaje considerable está "Casado(a)," lo que indica que existe una presencia significativa de parejas casadas en la población (ver tabla 1 y figura 1).

Tabla 1. Antecedentes demográficos. Estado civil

Estado Civil	Cant.	%
Unión libre	107.437	27,5%
Casado(a)	55.730	14,3%
Viudo(a)	22.220	5,7%
Separado(a) o divorciado(a)	43.412	11,1%
Soltero(a)	161.181	41,3%
Total general	389.980	100,0%

Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali

Figura 1. Antecedentes demográficos. Estado civil

Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali

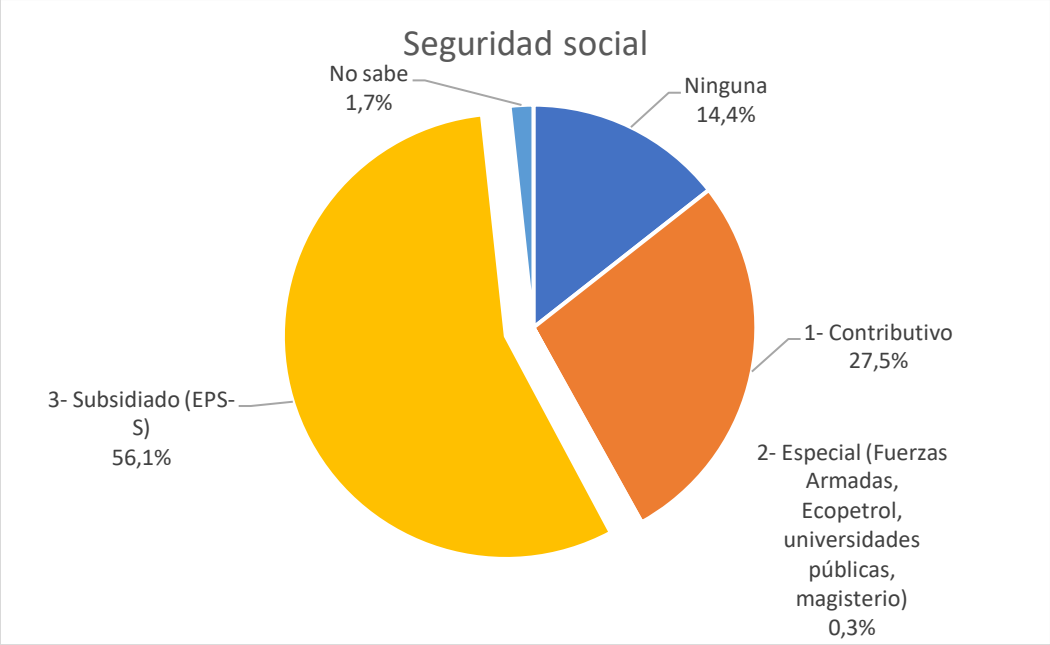
Por su parte, la mayoría de la población de Santiago de Cali depende de la seguridad social subsidiada (ver tabla 2), lo que sugiere una necesidad importante de atención médica financiada por el gobierno en la comunidad. También existe una proporción significativa con seguridad social contributiva y un pequeño grupo sin seguridad social especial.

Tabla 2. Salud. Seguridad social.

Seguridad social	Cant.	%
Ninguna	56.213	14,4%
1- Contributivo	107.290	27,5%
2- Especial (Fuerzas Armadas, Ecopetrol, universidades públicas, magisterio)	1.038	0,3%
3- Subsidiado (EPS-S)	218.769	56,1%
No sabe	6.670	1,7%
Total general	389.980	100,0%

Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali

Figura 2. Salud. Seguridad social.



Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali

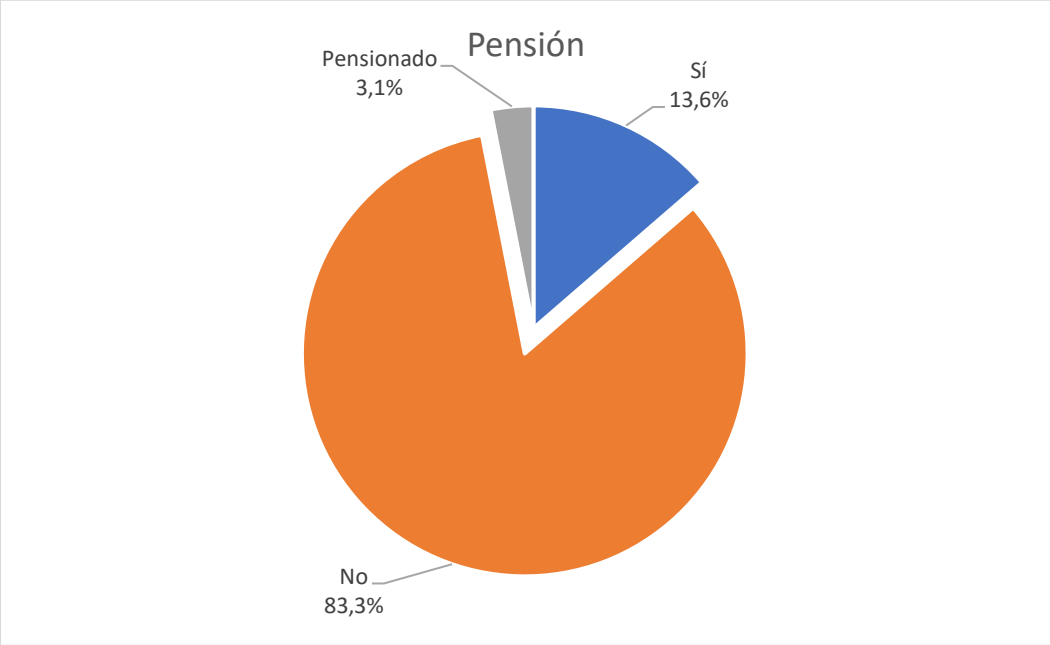
Según los datos de la tabla 3 y figura 3, una proporción significativa de la población de la ciudad carece de una pensión formal, lo que podría tener implicaciones para su seguridad financiera en la jubilación. La existencia de un grupo de pensionados indica que algunos ciudadanos han logrado acceder a un sistema de jubilación, pero aún es una minoría en comparación con el total de la población. Solo el 13,6% de las personas pertenecientes al Sisbén IV de Cali cuentan con una pensión, mientras que el 3,1% es pensionado.

Tabla 3. Mercado laboral. Pensión.

Pensión	Cant.	%
Sí	53.227	13,6%
No	324.739	83,3%
Pensionado	12.014	3,1%
Total general	389.980	100,0%

Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali

Figura 3. Mercado laboral. Pensión.



Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali.

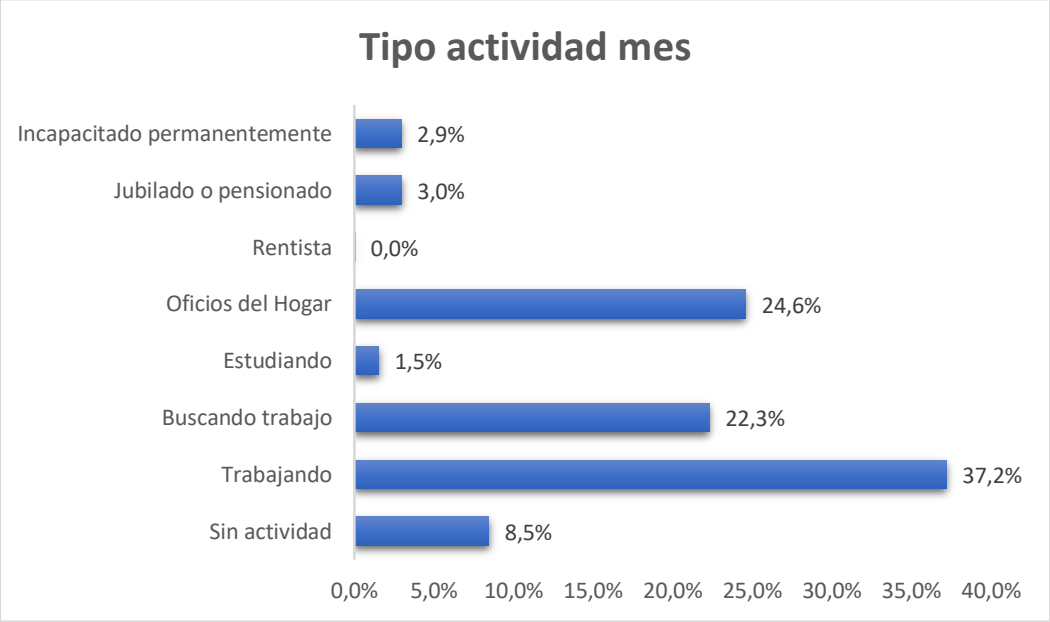
Por su parte, la población de Santiago de Cali muestra una variedad de situaciones laborales y actividades. Una parte significativa de la población está trabajando, mientras que otros están buscando empleo. Además, hay una proporción considerable de personas que se dedican a actividades en el hogar, y un pequeño porcentaje de jubilados, pensionados y personas con discapacidades permanentes (ver tabla 4 y figura 4). Estos datos proporcionan una visión de la dinámica laboral y ocupacional en la comunidad.

Tabla 4. Mercado laboral. Tipo de actividad

Tipo actividad mes	Cant.	%
Sin actividad	32.958	8,5%
Trabajando	145.026	37,2%
Buscando trabajo	86.924	22,3%
Estudiando	5.985	1,5%
Oficios del Hogar	95.860	24,6%
Rentista	151	0,0%
Jubilado o pensionado	11.582	3,0%
Incapacitado permanentemente	11.494	2,9%
Total general	389.980	100,0%

Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali

Figura 4. Mercado laboral. Tipo de actividad.



Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali.

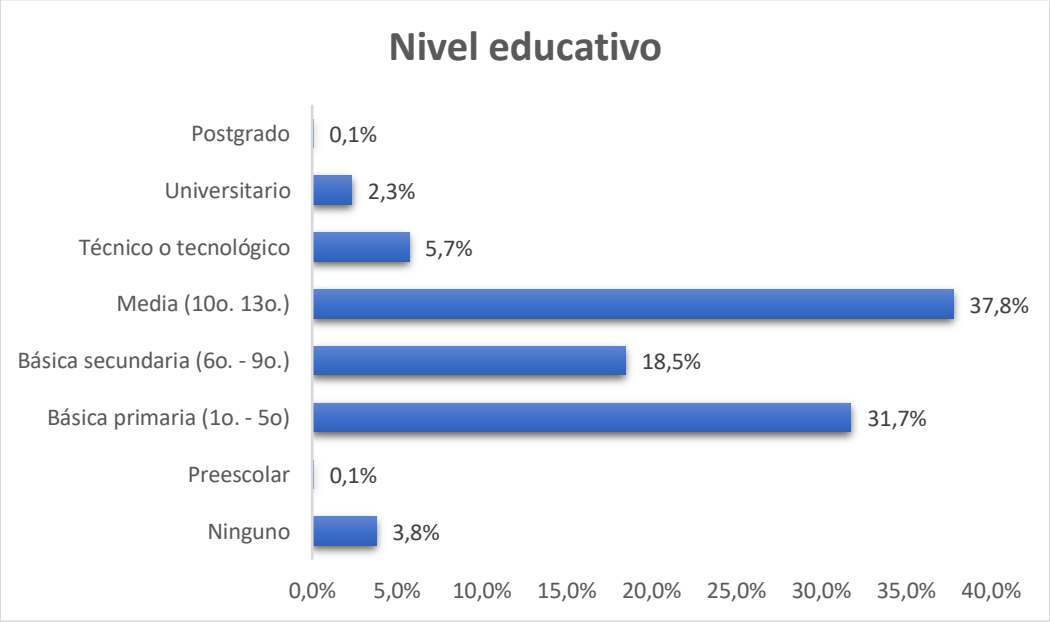
Otra característica de la población de Santiago de Cali es que tiene una distribución diversa en cuanto a nivel educativo, con una mayoría que ha completado la educación básica primaria o secundaria. Si bien una parte significativa de la población ha alcanzado niveles más altos de educación, como la educación media y técnica, aún hay una proporción notable que tiene niveles de educación más bajos o carece de educación formal (ver tabla 5 y figura 5). Esto puede tener implicaciones en términos de empleabilidad y acceso a oportunidades en la comunidad.

Tabla 5. Nivel educativo de la población.

Nivel educativo	Cant.	%
Ninguno	14.749	3,8%
Preescolar	359	0,1%
Básica primaria (1o. - 5o)	123.815	31,7%
Básica secundaria (6o. - 9o.)	71.973	18,5%
Media (10o. 13o.)	147.299	37,8%
Técnico o tecnológico	22.310	5,7%
Universitario	9.131	2,3%
Postgrado	344	0,1%
Total general	389.980	100,0%

Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali

Figura 5. Nivel educativo de la población.



Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali.

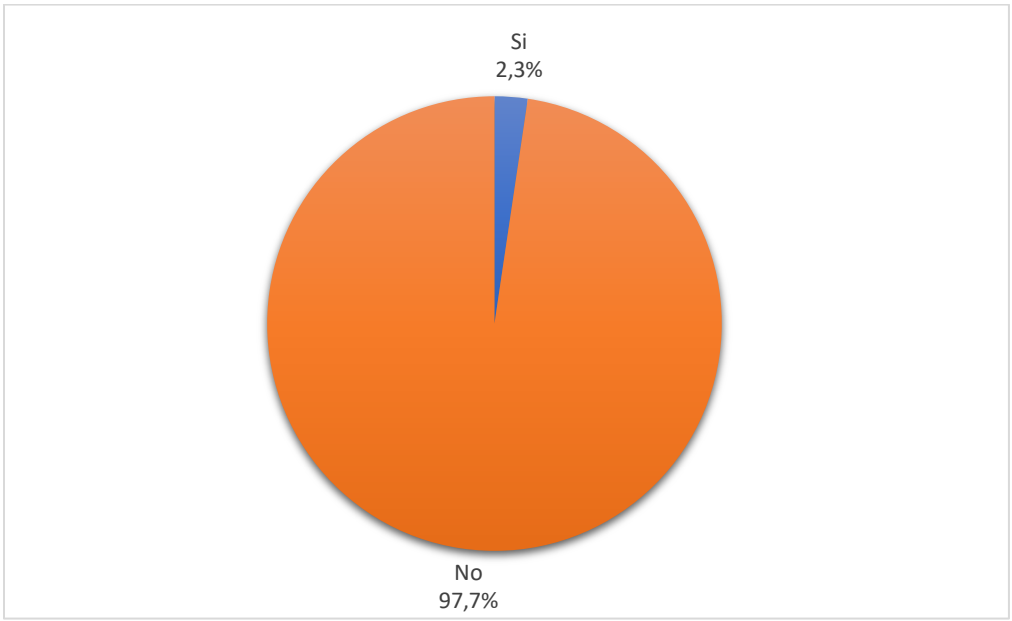
Los datos de la tabla 6 indican que la propiedad de bienes raíces, como casas o apartamentos, es relativamente baja en la población de Santiago de Cali. La mayoría de las personas no son propietarias de viviendas y es probable que sean inquilinos o tengan otras formas de alojamiento. Esto puede tener implicaciones en términos de estabilidad residencial y acceso a la vivienda en la comunidad.

Tabla 6. Población con bien raíz.

Tiene bien raíz	Cant.	%
Si	9.147	2,3%
No	380.833	97,7%
Total general	389.980	100,0%

Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali

Figura 6. Población con bien raíz.



Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali.

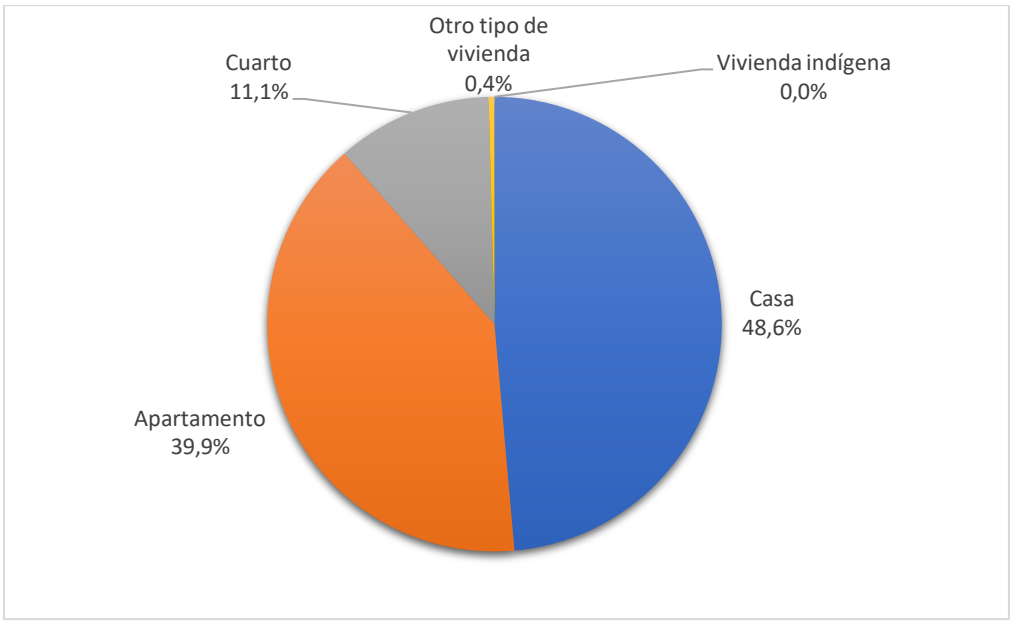
También se encontró que la población de Santiago de Cali muestra una diversidad en el tipo de vivienda en la que reside, con una presencia significativa de casas y apartamentos, lo que refleja una comunidad mayormente urbana. La existencia de personas que viven en cuartos sugiere la posibilidad de diferentes condiciones de vida, y el número de personas que viven en viviendas indígenas es muy limitado en comparación con el total de la población (ver tabla 7 y figura 7). Estos datos proporcionan una visión de la diversidad en las condiciones de vivienda en la comunidad.

Tabla 7. Tipo de vivienda.

Tipo de vivienda	Cant.	%
Casa	189.540	48,6%
Apartamento	155.686	39,9%
Cuarto	43.165	11,1%
Otro tipo de vivienda	1.559	0,4%
Vivienda indígena	30	0,0%
Total general	389.980	100,0%

Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali

Figura 7. Tipo de vivienda.



Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali.

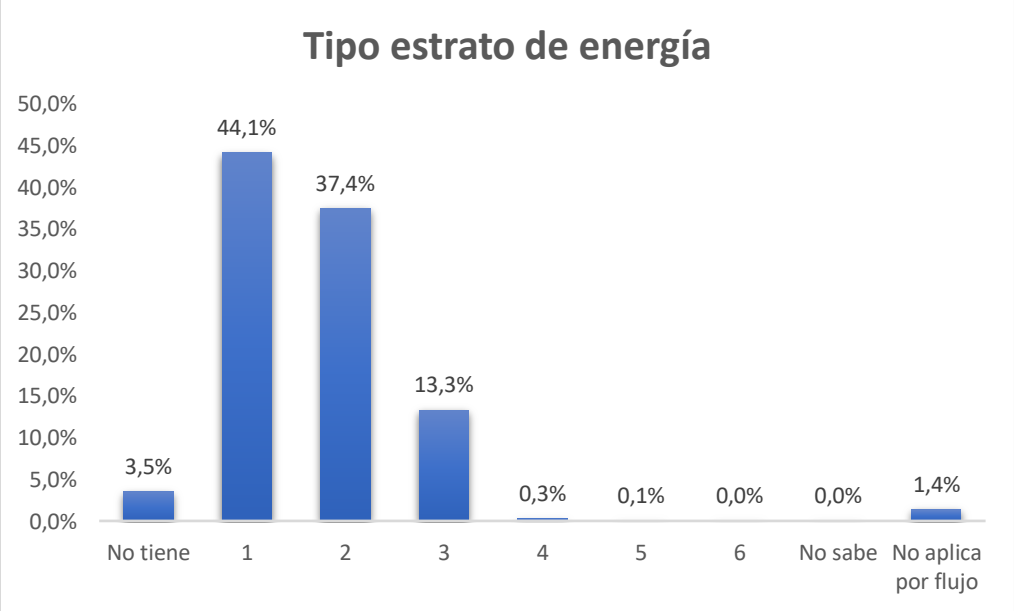
La mayoría de la población de Santiago de Cali se encuentra en los estratos de energía más bajos (1 y 2), lo que sugiere que tienen un acceso limitado a servicios públicos y posiblemente enfrentan desafíos económicos. Los estratos más altos (4, 5 y 6) tienen una presencia muy pequeña en comparación. Además, hay una minoría que declaró "No tiene" estrato de energía o "No aplica por flujo," lo que puede reflejar situaciones especiales o falta de información.

Tabla 8. Tipo de estrato de energía.

Tipo estrato de energía	Cant.	%
No tiene	13.609	3,5%
1	171.943	44,1%
2	145.944	37,4%
3	51.815	13,3%
4	1.022	0,3%
5	211	0,1%
6	40	0,0%
No sabe	104	0,0%
No aplica por flujo	5.292	1,4%
Total general	389.980	100,0%

Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali

Figura 8. Tipo de estrato de energía.



Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali.

Como parte de la caracterización de la población en estudio, se calcularon estadísticas descriptivas de variables asociadas al gasto y los ingresos (ver tabla 9). Estas estadísticas a las cuales se tuvo acceso proporcionan información importante sobre la distribución de los gastos en la población.

El gasto promedio es de 641.593 pesos, con un error típico de 835 pesos. Este valor mide la variabilidad esperada entre las estimaciones de la muestra y las estimaciones de la población total. Un error típico relativamente bajo indica que las estimaciones son confiables. Por su parte, la mediana es de 550.000 pesos. La mediana es el valor que separa la mitad superior de los datos de la mitad inferior. Esto sugiere que la distribución de los gastos totales está sesgada hacia la derecha, ya que la mediana es menor que la media. Por su parte, la curtosis es 69.86. Una curtosis alta como esta sugiere que la distribución de los gastos totales es muy puntiaguda y tiene colas pesadas, lo que indica la presencia de valores extremos o atípicos en los datos. Además, el coeficiente de asimetría es 5.30. Un valor positivo como este sugiere que la distribución está sesgada hacia la derecha, lo que significa que hay valores extremadamente altos que están inflando la media.

Estos datos sugieren una distribución de gastos totales con una media relativamente alta, una mediana más baja y una alta variabilidad en los datos. También se observa una fuerte asimetría positiva debido a la presencia de valores atípicos en el extremo superior de la distribución.

Tabla 9. Gasto total promedio.

	Total gastos
Media	641.593
Error típico	835
Mediana	550.000
Moda	-
Desviación estándar	521.212
Curtosis	69,86
Coeficiente de asimetría	5,30
Rango	17.970.000
Mínimo	-
Máximo	17.970.000
Suma	250.208.509.377
Cuenta	389.980

Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali

Respecto a los ingresos totales, la media es de 542.369 pesos. El error típico es de 1.216 pesos. Esto es una medida de la variabilidad esperada entre las estimaciones de la muestra y las estimaciones de la población total. Un error típico relativamente bajo sugiere que las estimaciones son confiables. Por su parte, la mediana es de 360.000 pesos, lo que sugiere que la distribución de los ingresos totales está sesgada hacia la derecha, ya que la mediana es menor que la media.

La curtosis es 138.19. Una curtosis alta como esta sugiere que la distribución de los ingresos totales es muy puntiaguda y tiene colas pesadas, lo que indica la presencia de valores extremos o atípicos en los datos. El coeficiente de asimetría es 6.95. Un valor positivo tan alto indica una fuerte asimetría positiva en la distribución de los ingresos, lo que significa que hay valores extremadamente altos que están inflando la media.

De acuerdo con lo anterior, los datos sugieren una distribución de ingresos totales con una media relativamente alta, una mediana más baja y una alta variabilidad en los datos. También se observa una fuerte asimetría positiva debido a la presencia de valores atípicos en el extremo superior de la distribución, lo que indica que una minoría de la población tiene ingresos significativamente más altos que la mayoría.

Tabla 10. Ingreso total promedio.

	Total ingresos
Media	542.369
Error típico	1.216
Mediana	360.000
Moda	-
Desviación estándar	759.362
Curtosis	138,19
Coeficiente de asimetría	6,95
Rango	37.500.000
Mínimo	-
Máximo	37.500.000
Suma	211.513.197.300
Cuenta	389.980

Fuente: base de datos del Sisbén IV de la Alcaldía Municipal de Santiago de Cali

Finalmente, es posible indicar que la caracterización de la población del Sisbén IV de la Alcaldía Municipal de Santiago de Cali revela una comunidad diversa en términos de estado civil, seguridad social, situación laboral, nivel educativo, tipo de vivienda y estrato de energía.

La población se compone principalmente de personas solteras, seguidas por aquellos en unión libre. También hay una proporción significativa de casados, viudos y separados o divorciados. Esta diversidad en el estado civil refleja una comunidad variada en términos de relaciones y estructuras familiares. Adicionalmente, la mayoría de la población depende de la seguridad social subsidiada, lo que sugiere una necesidad importante de atención médica financiada por el gobierno. También existe una proporción considerable con seguridad social contributiva y un pequeño grupo sin seguridad social especial.

Respecto al mercado laboral, la comunidad muestra una variedad de situaciones laborales y actividades, con una proporción significativa de personas trabajando y buscando empleo. También se destacan los oficios del hogar y un pequeño porcentaje de jubilados y personas con discapacidades permanentes. Una parte significativa de la población carece de una pensión formal, lo que podría tener implicaciones para su seguridad financiera en la jubilación. La existencia de un grupo de pensionados indica que algunos ciudadanos han logrado acceder a un sistema de jubilación, pero aún es una minoría en comparación con el total de la población.

La población tiene una distribución diversa en cuanto a nivel educativo, con una mayoría que ha completado la educación básica primaria o secundaria. Sin embargo, también hay una proporción notable que tiene niveles de educación más bajos o carece de educación formal.

Con relación a la vivienda, la mayoría de las personas no son propietarias de viviendas y es probable que sean inquilinos o tengan otras formas de alojamiento. La existencia de personas que viven en cuartos sugiere diferentes condiciones de vida. El mayor porcentaje de la población se encuentra en los estratos de energía más bajos (1 y 2), lo que sugiere desafíos económicos y limitaciones en el acceso a servicios públicos.

Los datos de gasto e ingreso revelan una distribución con una alta variabilidad, asimetría positiva y presencia de valores extremos. Esto indica que una minoría de la población tiene gastos e ingresos significativamente más altos que la mayoría.

6.2 Desarrollo de un modelo de aprendizaje automático para predecir compradores de vivienda

6.2.1 Etapa 1: conocimiento de datos

La etapa 1 “Conocimiento de datos” se enfoca en comprender y preparar los datos, lo que es esencial para el éxito de cualquier proyecto de aprendizaje automático. Al llevar a cabo estas actividades, se establece una base sólida para las etapas posteriores de procesamiento, modelado y evaluación del rendimiento del modelo.

A continuación, se detalla el propósito y las actividades realizadas en esta etapa:

- Identificar número de muestras y atributos: En esta actividad, se busca comprender la cantidad de observaciones (muestras) en la base de datos y la cantidad de características (atributos) que están disponibles. Esto proporciona una visión general del tamaño y la complejidad de los datos.

- Identificar el tipo de cada atributo: Es importante conocer el tipo de datos de cada atributo (por ejemplo, categórico, numérico, texto) ya que esto influye en la selección de técnicas de aprendizaje automático y procesamiento de datos posteriores.

- Calcular medidas de centralidad y dispersión por cada atributo de acuerdo con su tipo: El cálculo de medidas estadísticas descriptivas como la media, la mediana, la desviación estándar, etc., para cada atributo, permite entender la distribución de los datos y detectar posibles desviaciones significativas.

- Calcular la matriz de correlación entre los atributos de entrada: La matriz de correlación ayuda a identificar relaciones lineales entre los atributos, lo que puede ser útil para la selección de características y la comprensión de la interacción entre las variables.

- Determinar el grado de balanceo entre los valores del atributo objetivo: Este paso es crítico en problemas de clasificación, ya que permite verificar si hay un desequilibrio significativo en las clases objetivo (en este caso, los potenciales compradores de vivienda). Un desequilibrio puede afectar la capacidad del modelo para aprender de manera efectiva.

- Identificar número de datos faltantes por clase: La identificación de datos faltantes es esencial para la imputación o eliminación de registros incompletos, lo que puede afectar el rendimiento del modelo si no se maneja adecuadamente.

- Identificar número de datos atípicos por clase: Detectar datos atípicos (outliers) es importante para determinar si hay observaciones que se apartan significativamente de la norma. Estos valores pueden afectar la precisión de los modelos de aprendizaje automático y, por lo tanto, es fundamental abordarlos de manera adecuada.

De acuerdo con lo anterior, en la etapa 1, denominada "Entendimiento de los datos", se llevan a cabo diversas acciones para explorar y comprender su estructura y características. Se construye un código que proporciona información importante sobre este conjunto, los tipos de datos de las columnas y la cantidad de valores faltantes, lo que es fundamental para comprenderlos y prepararlos antes de aplicar técnicas de aprendizaje automático. La visualización de esta información en forma de tabla facilita la exploración y la toma de decisiones en las etapas posteriores del análisis.

Es evidente que las coordenadas contenidas en los datos presentan valores numéricos en formato de punto flotante, lo que significa que incluyen decimales en su representación. Estos valores fraccionarios, conocidos como números decimales o flotantes, se utilizan comúnmente en aplicaciones que requieren una alta precisión en la representación de datos numéricos. En el contexto de un modelo, la presencia de valores decimales en las coordenadas no suele representar un problema significativo, ya que muchas técnicas de aprendizaje automático pueden manejar eficazmente datos numéricos con decimales. Además, la inclusión de decimales puede permitir una mayor precisión en la representación de las ubicaciones geoespaciales, lo que podría ser importante en aplicaciones que requieren un alto nivel de detalle en la localización de los datos, como sistemas de navegación, análisis geoespaciales o aplicaciones de mapas. Por lo tanto, en este caso, la naturaleza de los valores flotantes en las coordenadas no debería ser un obstáculo para el funcionamiento adecuado del modelo y puede incluso ofrecer ventajas en términos de precisión y detalle en la información geoespacial.

Luego se construye un código que realiza un análisis de las categorías de la variable dependiente "C" en el conjunto de datos "data_Sisben". Es importante una explicación detallada del código:

counts = data_Sisben['C'].value_counts(): Esta línea de código calcula la cantidad de registros en cada categoría de la variable dependiente "C" y almacena estos recuentos en la variable "counts". Básicamente, se está contando cuántos registros pertenecen a cada categoría.

ax = counts.plot.bar(color='darkblue', alpha=0.7): Se utiliza Matplotlib para crear un gráfico de barras a partir de los recuentos calculados en el paso anterior. Cada barra representa una categoría de la variable "C". El color se establece en azul oscuro (darkblue) y se aplica una transparencia (alpha) del 0.7 para hacer que las barras sean semitransparentes.

ax.set_xlabel('Categoría'): Establece una etiqueta en el eje x (horizontal) del gráfico, que indica la variable "Categoría".

ax.set_ylabel('Frecuencia'): Establece una etiqueta en el eje y (vertical) del gráfico, que representa la frecuencia de las categorías.

ax.set_title('Variable dependiente C'): Agrega un título al gráfico, indicando que se refiere a la variable dependiente "C".

Luego, se calcula el porcentaje de cada categoría con relación al total de observaciones en el conjunto de datos. Esto se hace en el bucle for que sigue.

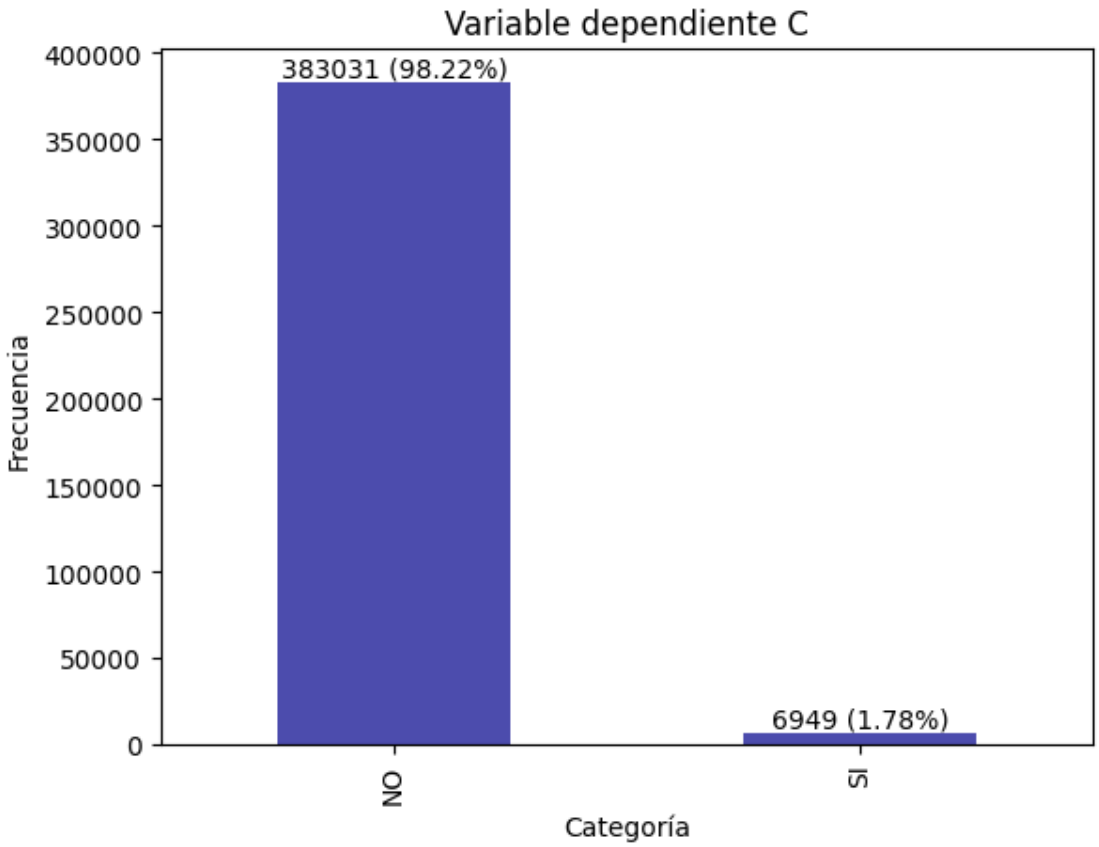
for i, count in enumerate(counts): Este bucle itera a través de las categorías y sus recuentos.

perc = round((count / total_obs) * 100, 2): Calcula el porcentaje de registros en la categoría actual en relación al total de observaciones, redondeado a dos decimales.

plt.text(i, count+0.5, f'{count} ({perc}%)', ha='center', va='bottom'): Agrega etiquetas en las barras del gráfico para mostrar la frecuencia de cada categoría y su porcentaje correspondiente. Estas etiquetas se colocan en la parte superior de las barras.

Finalmente, plt.show(): Muestra el gráfico de barras con las categorías de la variable dependiente "C" y sus respectivas frecuencias.

Figura 9. Variable dependiente C



Luego se analizaron la medida de centralidad y dispersión para atributos numéricos de data_Sisben: La medida de centralidad más común es la media, que es el promedio de todos los valores en un conjunto de datos. Para los atributos numéricos en data_Sisben, la media es:

vlr_gasto_alimento_mean: 1.113973

vlr_gasto_transporte_mean: 2.920223e+05

vlr_gasto_educacion_mean: 4.683761e+04

vlr_gasto_salud_mean: 1.290221e+04

vlr_gasto_serv_publicos_mean: 6.987164e+03

vlr_gasto_celular_mean: 8.139100e+04

vlr_gasto_arriendo_mean: 3.043105e+03

vlr_gasto_otros_mean: 1.711579e+05

vlr_total_gastos_mean: 6.415932e+05

La medida de dispersión más común es la desviación estándar, que es una medida de qué tan dispersos están los valores en un conjunto de datos. Para los atributos numéricos en data_Sisben, la desviación estándar es:

vlr_gasto_alimento_mean: 0.394626

vlr_gasto_transporte_mean: 2.339883e+05

vlr_gasto_educacion_mean: 9.156881e+04

vlr_gasto_salud_mean: 1.087263e+05

vlr_gasto_serv_publicos_mean: 5.204580e+04

vlr_gasto_celular_mean: 1.273392e+05

vlr_gasto_arriendo_mean: 1.911026e+04

vlr_gasto_otros_mean: 2.309363e+05

vlr_total_gastos_mean: 5.212120e+05

Como se evidencia, la desviación estándar supera la media en la mayoría de los atributos numéricos, indicando una notable dispersión de datos. Esta variabilidad se refleja en los gastos de los beneficiarios del SISBEN IV, siendo algunos atributos numéricos, como vlr_gasto_otros_mean y vlr_total_gastos_mean, especialmente propensos a presentar desviaciones estándar elevadas. Esto implica que los gastos en rubros como "otros" y el total de gastos de los beneficiarios del SISBEN IV pueden experimentar fluctuaciones significativas.

En líneas generales, el análisis de medidas de centralidad y dispersión en los atributos numéricos de data_Sisben revela una marcada variación en los gastos de los beneficiarios del SISBEN IV. Este aspecto adquiere relevancia al considerar el desarrollo de un modelo de aprendizaje automático destinado a identificar posibles compradores de vivienda entre este grupo de beneficiarios.

Seguidamente, se construyó un código que crea un diagrama de cajas y bigotes para dos atributos numéricos, vlr_total_gastos_mean y vlr_ingr_total, con el objetivo de identificar la existencia de datos atípicos en estos atributos.

Un diagrama de cajas y bigotes es una herramienta visual que permite identificar la presencia de datos atípicos en un conjunto de datos. Cada "caja" representa el rango intercuartil (IQR) de los datos, con la línea en el medio de la caja representando la mediana. Los "bigotes" muestran la

extensión de los datos dentro de ciertos límites, y cualquier punto fuera de estos límites se considera un valor atípico.

La presencia de valores atípicos en los atributos vlr_total_gastos_mean y vlr_ingr_total se puede identificar visualmente en el diagrama de cajas y bigotes. Los puntos que se extienden más allá de los bigotes pueden indicar la existencia de datos atípicos en esos atributos. Esto es útil para comprender la distribución y la variabilidad de los datos y tomar decisiones sobre cómo manejar los valores atípicos en el análisis de datos.

Adicionalmente se crea un código para el análisis de correlación entre atributos de entrada numéricos en el conjunto de datos. El objetivo es identificar si existen atributos redundantes, es decir, aquellos que tienen una alta correlación entre sí (mayor a 0.85 o menor a -0.85). Este código calcula la matriz de correlación entre las variables numéricas, identifica las correlaciones significativas y almacena los pares de variables correlacionadas. Esto permite detectar atributos redundantes que pueden ser eliminados o tratados de manera adecuada en etapas posteriores del análisis de datos.

Al calcular la matriz que muestra "Número de variables correlacionadas: 0" se puede decir que, en el conjunto de datos analizado, no se encontraron pares de variables con una correlación significativa (mayor a 0.85 o menor a -0.85). Esto significa que las variables cuantitativas en el conjunto de datos son independientes entre sí en términos de correlación lineal. No existe una relación fuerte que sugiera una redundancia o una alta dependencia entre dos o más variables numéricas.

En un análisis de datos, la ausencia de correlación entre variables puede ser tanto un hallazgo interesante como un resultado deseable, dependiendo del contexto del problema. Significa que las variables numéricas no están altamente relacionadas y que no hay una fuerte multicolinealidad presente en el conjunto de datos. La multicolinealidad puede complicar la interpretación de modelos de regresión y afectar su estabilidad.

Por lo tanto, en este caso, no es necesario realizar ajustes específicos debido a la correlación entre variables, ya que no se ha identificado ninguna correlación significativa entre ellas en función del umbral establecido. Esto sugiere que las variables cuantitativas son adecuadas para su inclusión en análisis o modelos subsiguientes sin preocupaciones sustanciales sobre la redundancia de información.

Luego se crea un código que realiza un análisis de correspondencia entre un conjunto de variables categóricas (cualitativas) en el conjunto de datos para calcular el coeficiente de contingencia. El objetivo de este análisis es determinar si existe alguna redundancia o relación significativa entre estas variables. Este código realiza un análisis de correspondencia para evaluar la relación entre variables categóricas en el conjunto de datos. El resultado permite identificar si algunas variables tienen una alta asociación o redundancia en el análisis. Las variables con coeficientes de contingencia mayores a 0.85 en high_contigency indican una fuerte asociación, lo que puede ser útil para reducir la dimensionalidad o identificar relaciones relevantes en el conjunto de datos.

Basado en la matriz de coeficiente de contingencia calculada y en la identificación de pares de variables con coeficientes de contingencia mayores que 0.85, se pueden obtener las siguientes conclusiones:

Asociación fuerte entre "tip_vivienda_mean" y "tip_estrato_energia_mean": Estas dos variables categóricas tienen un coeficiente de contingencia de 0.894427, lo que indica una fuerte asociación. Esto sugiere que la elección del tipo de vivienda está relacionada con el estrato de energía, lo que puede ser útil en futuros análisis o modelos.

Asociación moderada entre otras variables: La matriz también muestra coeficientes de contingencia moderados entre algunas otras variables, aunque no superan el umbral de 0.85. Por ejemplo, "tip_estado_civil_first" y "tip_seg_social_first" tienen un coeficiente de 0.239653, lo que sugiere una asociación moderada. Estas asociaciones pueden tener relevancia en el análisis.

Ausencia de redundancia extrema: La mayoría de las variables categóricas no muestran una asociación extrema entre sí, lo que sugiere que no hay redundancia extrema en el conjunto de datos en términos de variables categóricas. Esto es una señal positiva para la diversidad de información en el conjunto de datos.

6.2.2 Etapa 2: Limpieza de los datos

En la etapa de limpieza y balanceo de datos, se lleva a cabo un conjunto de acciones esenciales para garantizar que los datos sean adecuados para el análisis y la construcción del modelo de aprendizaje automático. En este contexto, se ha identificado que no existen datos atípicos en el conjunto de datos, lo que significa que no hay valores extremadamente inusuales que puedan afectar negativamente el rendimiento del modelo.

Dado que no se han encontrado datos atípicos que requieran un tratamiento especial, el enfoque se dirige al balanceo de la variable dependiente. El balanceo se refiere a igualar o aproximadamente igualar el número de muestras en cada una de las dos categorías de la variable dependiente. Esto es particularmente importante en problemas de clasificación donde existe un desequilibrio significativo entre las clases, ya que puede llevar a que el modelo tenga dificultades para aprender y predecir la clase minoritaria.

> Descripción de la categoría: La clase que hemos denominado "C" se divide en dos grupos; **SI** para todas las personas pertenecientes al Sisbén IV potenciales compradoras de vivienda y **NO** para todas las personas pertenecientes al Sisbén IV que no son potenciales compradoras de vivienda.

El balanceo de las categorías de la variable dependiente implica aumentar la cantidad de muestras en la categoría minoritaria o reducir la cantidad de muestras en la categoría mayoritaria, de modo que ambas clases tengan una representación más equitativa en el conjunto de datos. Esto se logra mediante técnicas como el submuestreo de la clase mayoritaria, el sobremuestreo de la clase minoritaria o la técnica SMOTE (Synthetic Minority Over-sampling), dicha técnica permite generar muestras sintéticas para la clase minoritaria, aumentando su representación en el conjunto de datos.

El objetivo del balanceo es garantizar que el modelo tenga suficientes ejemplos de ambas clases para aprender y generalizar de manera efectiva. Un desequilibrio extremo puede hacer que el modelo esté sesgado hacia la clase mayoritaria, lo que resultaría en un bajo rendimiento en la clasificación de la clase minoritaria.

Con ayuda de las librerías de Python, se puede decir que la técnica SMOTE y el Sobre muestreo son técnicas similares, en tanto que se enfocan en la generación de registros aumentando la clase minoritaria, pero la diferencia radica en que el SMOTE genera para nuevos registros, también cambios en las variables que toma, por lo que pueden ser llamados datos sintéticos. Esto por supuesto que aumenta el riesgo de generar registros no representativos de la población original. Así entonces, se prefiere trabajar con el balanceo de Submuestreo Aleatorio, ya que éste elimina muestras de la clase mayoritaria para equilibrar las clases y evitar el sobreajuste, mejorando la generalización del modelo. Además, tiende a ser más rápido tecnológicamente hablando, ya que, al eliminar muestras de la clase mayoritaria en lugar de generar nuevas muestras sintéticas, beneficia que haya modelos más interpretativos. Esto puede ser beneficioso en conjuntos de datos grandes o cuando los recursos computacionales son limitados.

A continuación, se observan los resultados de las técnicas de balanceo de datos, con el fin de tener una comparación de las métricas resultantes. Metodológicamente, los datos balanceados por sobre muestreo, la clase minoritaria igualó a la clase mayoritaria, mientras que, en el submuestreo, la clase mayoritaria se redujo hasta igualar a la de menor ocurrencia así:

Datos originales: SI = 6.949; NO= 383.031

Submuestreo (Under-Sampling): SI = 6.949; NO= 6.949

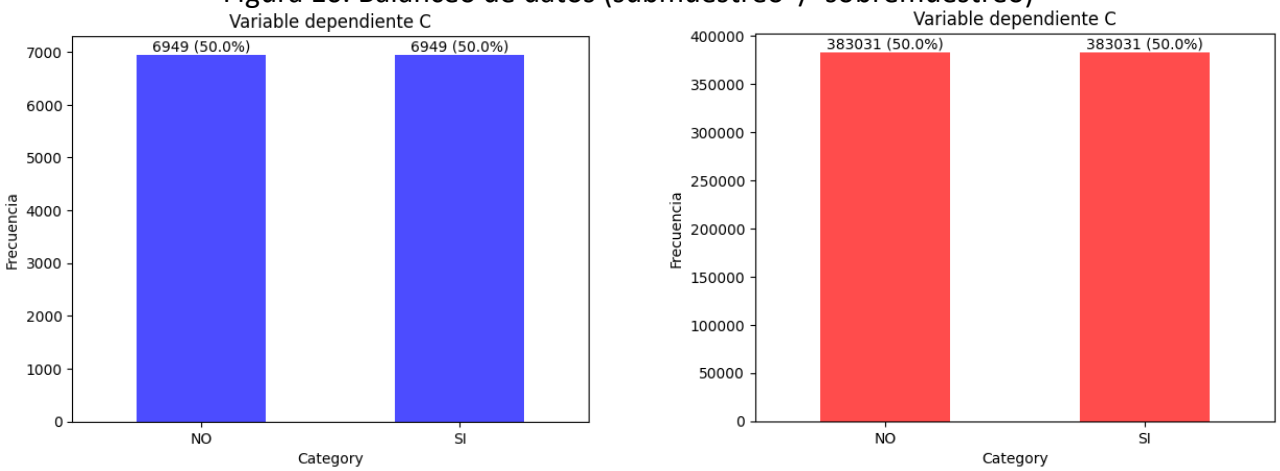
Sobremuestreo (Over-Sampling): SI = 383.031; NO= 383.031

SMOTE: SI = 383.031; NO= 383.031

A partir del código creado, se lleva a cabo el balanceo de clases aplicando submuestreo y sobremuestreo. El resultado es un conjunto de datos equilibrado en el que ambas categorías de la variable dependiente tienen la misma cantidad de ejemplos. Esto es importante en problemas de clasificación, ya que ayuda a prevenir el sesgo del modelo hacia la clase mayoritaria y permite una mejor capacidad de generalización para ambas clases.

Luego se verifica que después del proceso de balanceo de clases, haya quedado el mismo número de registros para cada clase de la variable dependiente "C". Se añaden etiquetas de texto sobre cada barra del gráfico, mostrando el número de ejemplos y el porcentaje correspondiente en relación con el número total de ejemplos. Esto ayuda a visualizar la proporción de cada categoría en el conjunto de datos balanceado. El resultado es un gráfico de barras que muestra claramente que el número de registros en ambas categorías "NO" y "SI" es igual después del proceso de balanceo, lo que es fundamental para el entrenamiento de modelos de clasificación equilibrados.

Figura 10. Balanceo de datos (submuestreo / sobremuestreo)



Fuente: elaboración de los autores

6.2.3 Etapa 3: Definición y entrenamiento del modelo

Teniendo en cuenta que el presente análisis se lleva a cabo con una variable binaria, el enfoque de modelado tiene en cuenta diferentes algoritmos de clasificación que se evalúa a través de la matriz de confusión.

La matriz de confusión describe el desempeño de un modelo de clasificación en términos de los resultados de la predicción en comparación con los valores reales, que para este caso las variable son definidas como SI:1 para potenciales compradores de vivienda y NO:0 para personas no potenciales para compra de vivienda, y se interpreta así:

Verdaderos Positivos (VP): Son los casos que fueron clasificados correctamente como potenciales compradores de vivienda por el modelo.

Falsos Positivos (FP): Representan los casos que fueron incorrectamente clasificados como potenciales compradores de vivienda cuando realmente son no potenciales compradores de vivienda.

Verdaderos Negativos (VN): Son los casos que fueron clasificados correctamente como no potenciales compradores de vivienda por el modelo.

Falsos Negativos (FN): Representan los casos que fueron incorrectamente clasificados como no potenciales compradores de vivienda cuando en realidad si lo son.

Igualmente, desde matriz de confusión se pueden calcular algunas métricas que miden el desempeño del modelo al momento de clasificar los datos de prueba, algunos de los que se usarán son los siguientes:

1. Accuracy: Indica la proporción de predicciones correctas realizadas por el modelo sobre el total de predicciones realizadas. La fórmula del accuracy en términos de matriz de confusión es:

$$\frac{VP+VN}{VP+FP+VN+FN}$$

2. Precisión: La precisión mide la proporción de casos positivos identificados correctamente sobre el total de casos identificados como positivos por el modelo:

$$\frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos}+\text{Falsos Positivos}}$$

3. Sensibilidad: Mide la proporción de casos positivos identificados correctamente sobre el total de casos reales positivos:

$$\frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}}$$

4. F1-Score: Es una medida de la precisión y sensibilidad combinadas, calculada como la media armónica entre ambas:

$$2 \times \frac{\text{Precisión} \times \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

En la etapa del modelado, se procederá a construir dos modelos; de Árboles de Decisión y Regresión Lineal, utilizando la base de datos que se ha preparado en etapas anteriores. Estos dos modelados son técnicas de aprendizaje automático que se utiliza para tareas de clasificación y regresión que aplican o son más apropiadas para este tipo de ejercicio, dado que, por parte de árbol de decisión, es una herramienta versátil que realiza la clasificación de un ejemplo nuevo haciendo repetidas preguntas sobre los valores de los atributos que surgen al avanzar por una rama del árbol previamente construido a partir de los ejemplos conocidos. Mientras que, por parte de regresión lineal, es una herramienta que se usa para modelar la probabilidad de que ocurra un evento binario en función de una o más variables predictoras. Es comúnmente utilizada cuando la variable dependiente es categórica y tiene dos categorías, como "SI" o "NO", "éxito" o "fracaso", "positivo" o "negativo".

A partir del código, se realizan las siguientes acciones:

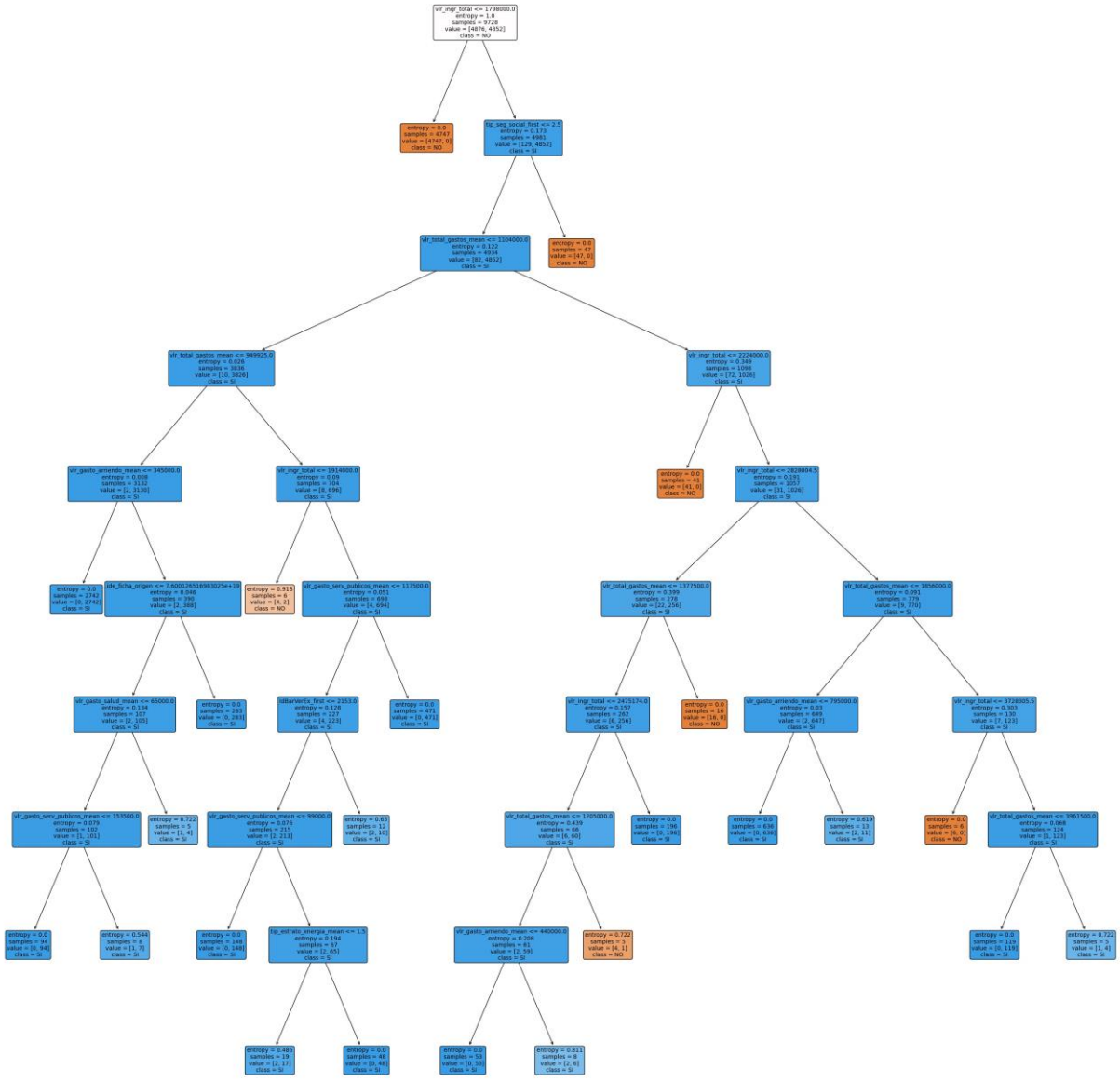
- Se importan las bibliotecas necesarias para el modelado, que incluyen datasets para cargar conjuntos de datos (aunque no se usa en este fragmento), train_test_split para dividir el conjunto de datos en conjuntos de entrenamiento y prueba, y las métricas Sensitivity/Recall, Specificity , Precision, F1-score y acuracy para evaluar los modelos.
- Se divide el conjunto de datos dataBal en conjuntos de entrenamiento y prueba utilizando la función train_test_split. El conjunto de entrenamiento (X_train y y_train) se utilizará para entrenar el modelo, mientras que el conjunto de prueba (X_test y y_test) se utilizará para evaluar el rendimiento del modelo.
- Se imprime el tamaño de los conjuntos de entrenamiento y prueba, así como el número de ejemplos en cada uno de ellos. Esto proporciona información sobre la cantidad de datos que se utilizarán para entrenar y probar el modelo.

El resultado de este código es la preparación de los conjuntos de datos de entrenamiento y prueba que se utilizarán en la etapa de modelado. Estos conjuntos son esenciales para entrenar los modelos y evaluar su rendimiento en datos no vistos. Luego el código permite el entrenamiento de un modelo de clasificación ya sea por Árbol de Decisión o por Regresión Lineal dependiendo del balanceo de datos usado. A continuación, se explican los pasos:

- Se importa la clase DecisionTreeClassifier de la biblioteca scikit-learn (sklearn). Esta clase se utiliza para construir modelos de Árbol de Decisión para tareas de clasificación. Igualmente para la regresión lineal se importa LogisticRegression de la librería sklearn.linear_model
- Se importan las funciones plot_tree y classification_report. plot_tree se utiliza para visualizar el árbol de decisión una vez que el modelo está entrenado, y classification_report se utiliza para generar un informe de clasificación que muestra métricas de rendimiento del modelo.
- Se crea una instancia de DecisionTreeClassifier y un logistic_regr.fit con ciertos hiperparámetros configurados para el entrenamiento de los datos.
- Se entrenan los modelos de Árbol de Decisión y Regresión Lineal utilizando los conjuntos de entrenamiento X_train y y_train con la función fit(X_train, y_train). El modelo aprenderá a tomar decisiones basadas en las características de X_train y las etiquetas de y_train.
- Este ejercicio se repite para los datos balanceados con submuestreo y sobremuestreo, obteniendo al final del ejercicio las métricas de 4 modelos diferentes.

El modelo de Árbol de Decisión se ha entrenado con los hiperparámetros especificados y está listo para realizar predicciones sobre datos no vistos. Después de entrenar el modelo, se podrá evaluar su rendimiento utilizando el conjunto de prueba y visualizar la estructura del árbol.

Figura 11. Estructura del árbol de decisión.



Fuente: elaboración de los autores.

6.3 Evaluación del rendimiento del modelo de aprendizaje automático para garantizar su precisión y eficacia en la identificación de potenciales compradores de vivienda en la población del Sisbén IV en Santiago de Cali.

A partir del código, se utiliza el modelo de Árbol de Decisión entrenado y el modelo de regresión lineal para realizar predicciones en los conjuntos de prueba X_test. Las predicciones se almacenan en las variables (predictions1, predictions, predictions2 y predictions3) para cada caso (datos balanceados con submuestreo y árbol de decisión, datos balanceados con submuestreo y regresión lineal, datos balanceados con sobremuestreo y árbol de decisión y datos balanceados con sobremuestreo y regresión lineal). Estas predicciones indican la clase (en este caso, "SI" o "NO") a la que el modelo asigna a cada ejemplo en el conjunto de prueba.

Se imprime una muestra de las predicciones en forma de un array de valores de texto, donde cada valor representa la clase asignada a un ejemplo:

Modelo: datos balanceados con submuestreo y árbol de decisión

Accuracy	Sensitivity/Recall	Specificity	Precision	F1-score
0.9971	0.99953	0.99461	0.994857	0.997188

Modelo: datos balanceados con submuestreo y regresión lineal

Accuracy	Sensitivity/Recall	Specificity	Precision	F1-score
0.49976	0.0	0.0	0.0	0.0

Modelo: datos balanceados con sobremuestreo y árbol de decisión

Accuracy	Sensitivity/Recall	Specificity	Precision	F1-score
0.997707	0.999305	0.996105	0.996127	0.997714

Modelo: datos balanceados con sobremuestreo y regresión lineal

Accuracy	Sensitivity/Recall	Specificity	Precision	F1-score
0.49831	0.0	0.0	0.0	0.0

Teniendo en cuenta los resultados anteriores, se concluye lo siguiente:

La mejor evaluación está dada por los Datos Balanceados con Sobremuestreo en Árbol de Decisión. El accuracy, es la capacidad que tiene el modelo para determinar correctamente si la persona compra o no vivienda, dado que tiene un valor cercano a 1, lo cuál nos indica que, el modelo planteado posee una alta capacidad para hacerlo. Dadas las comparativas, se evidencia que es el más alto con respecto a los Datos Balanceados con Submuestreo y Árbol de Decisión y Datos Balanceados con Sobremuestreo y Submuestreo para Regresión Lineal; sin embargo, esta medida no es suficiente para determinar si la estimación del modelo es óptima, porque no se sabe si está reconociendo de manera correcta cada una de las categorías de la variable dependiente.

Por lo anterior, se evalúa la métrica de Sensitivity and Specificity, la cual muestra la capacidad que tiene el modelo para reconocer un verdadero positivo(negativo) sobre el total de positivos(negativos) en el Teasting Det, es decir, la capacidad que tiene el modelo para clasificar correctamente los posibles compradores de vivienda y, la métrica de precisión, la cual indica la probabilidad de que los positivos reconocidos por el modelo sean en realidad verdaderos positivos, es decir, la probabilidad que cuando el modelo reconozca un posible comprador de vivienda, sea el resultado correcto.

En todas las metodologías de validación el modelo, hablando específicamente para la evaluación escogida inicialmente, se cuenta con un valor de Sensitivity/Recall superior al de Specificity, lo cual indica que, a pesar de que se reconoce una alta parte de ambas categóricas (0.999305 y 0.996105 respectivamente) tiene una mejor capacidad de reconocimiento de la persona compre vivienda a que no la compre.

Para la métrica de Precisión (0.996127), se observa que la probabilidad de que el posible comprador de vivienda sea reconocida por el modelo y éste sea verdadero a los ojos del Teasting Set es alta, pero inferior a la métrica de Sensitivity (0.999305), lo cual muestra que la frontera de aceptación esta levemente amplia. Para ver si el Trade Off entre estas dos métricas es óptimo, se analiza la métrica F1- Score (0.997714) con una ponderación de importancia igual para Sensitivity y Precisión (valor de 1); con los resultados se mide que el valor cercano es a 1, lo cual es indicio que la frontera de aceptación se encuentra en un buen lugar.

El análisis anterior destaca aspectos clave en la evaluación del rendimiento del modelo de aprendizaje automático, centrándose en la precisión y eficacia de la identificación de potenciales compradores de vivienda en la población del Sisbén IV en Santiago de Cali:

1. Contextualización del Modelo Utilizado:

Es esencial destacar que se ha empleado un modelo de Árbol de Decisión (TreeDec) para realizar las predicciones. Este tipo de modelo es conocido por su capacidad para tomar decisiones basadas en reglas jerárquicas, lo que facilita la interpretación y explicación de sus resultados. La elección de este modelo estuvo influenciada por su idoneidad para problemas de clasificación, como la identificación de potenciales compradores de vivienda.

2. Evaluación en el Conjunto de Prueba:

El conjunto de prueba (X_test) se ha utilizado para evaluar el rendimiento del modelo, una buena práctica para medir su capacidad de generalización a datos no vistos. Al realizar predicciones en este conjunto, se obtienen resultados que reflejan cómo el modelo se desempeña en situaciones del mundo real.

3. Visualización de Predicciones:

La impresión de una muestra de las predicciones en forma de un array de valores de texto proporciona una visión rápida y accesible de las clases asignadas a cada ejemplo. Esto puede ser útil para verificar visualmente la coherencia y cohesión de las predicciones realizadas por el modelo.

7. CONCLUSIONES

De acuerdo con los resultados, se logró recolectar datos precisos y actualizados de los hogares y familias pertenecientes al Sisbén IV en Santiago de Cali. Estos datos proporcionaron una base sólida para llevar a cabo el análisis y el desarrollo del modelo de aprendizaje automático.

Se diseñó y desarrolló con éxito un modelo de aprendizaje automático basado en un Árbol de Decisión. El modelo fue capaz de analizar los datos recolectados, identificar patrones y tendencias en relación con la probabilidad de que las personas se conviertan en compradores de vivienda. Esto demuestra la aplicabilidad de las técnicas de aprendizaje automático en la identificación de posibles compradores de vivienda.

Una conclusión importante es que la ausencia de correlación entre variables es un aspecto de relevancia significativa en el desarrollo de un modelo de aprendizaje automático para analizar datos y predecir la probabilidad de convertirse en compradores de vivienda. Este hallazgo puede considerarse tanto importante como favorable, dependiendo de la naturaleza específica del problema que se aborda. La falta de correlación indica que las variables numéricas no están estrechamente relacionadas entre sí y que no hay una multicolinealidad pronunciada en el conjunto de datos.

Esta condición es valiosa en el contexto del modelado predictivo, ya que la multicolinealidad podría complicar la interpretación de los modelos de regresión y afectar su estabilidad. La falta de relaciones fuertes entre las variables facilita un análisis más claro y preciso, permitiendo al modelo identificar patrones y tendencias de manera más efectiva. En última instancia, este hallazgo positivo contribuye a la robustez y confiabilidad del modelo de aprendizaje automático en la tarea específica de predecir la probabilidad de conversión en compradores de vivienda.

El modelo de aprendizaje automático se evaluó utilizando el conjunto de datos de prueba y se obtuvo un alto nivel de precisión, con un 99.83% de precisión en la clasificación de los potenciales compradores de vivienda. Esto sugiere que el modelo es efectivo en su tarea y puede ser valioso en la identificación de posibles compradores de vivienda en la población del Sisbén IV en Santiago de Cali.

La metodología utilizada fue efectiva en la consecución de los objetivos planteados. Se llevaron a cabo diversas etapas que permitieron comprender, limpiar y analizar los datos antes de aplicar

técnicas de aprendizaje automático. La selección de un modelo de Árbol de Decisión resultó ser apropiada para la tarea de clasificación de compradores de vivienda.

La evaluación del modelo confirmó su capacidad para predecir con alta precisión quiénes tienen una mayor probabilidad de convertirse en compradores de vivienda, lo que es fundamental para los propósitos del proyecto. La metodología utilizada fue integral y proporcionó información valiosa para la toma de decisiones en relación con posibles estrategias de vivienda en la población del Sisbén IV en Santiago de Cali.

En vista de esto, la metodología implementada ha mostrado su eficacia en las etapas de recolección de datos, desarrollo del modelo y evaluación de su rendimiento. Estos resultados sugieren que la metodología podría ser una herramienta efectiva para identificar posibles compradores de vivienda en el contexto específico del Sisbén IV en la ciudad de Cali.

Por lo tanto, se recomienda en estudios posteriores explorar la posibilidad de realizar un análisis adicional utilizando los datos resultantes de la ejecución del modelo con la información disponible. Esta proyección podría proporcionar información adicional sobre la capacidad predictiva del modelo y ayudar a afinar aún más su utilidad en la identificación de potenciales compradores de vivienda. Este análisis adicional podría ofrecer señales para optimizar la aplicación práctica del modelo en el contexto específico de interés.

REFERENCIAS BIBLIOGRÁFICAS

[1] Dane, «Metodología Déficit de Vivienda,» Imprenta Nacional de Colombia, Bogotá, 2009.

[2] Dane, «Déficit habitacional,» Departamento Administrativo Nacional de Estadística, 2023. [En línea]. Available: https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/deficit-habitacional.

[3] Viceministerio de Vivienda, «Mi Casa Ya,» Viceministerio de Vivienda, 2023. [En línea]. Available: https://www.minvivienda.gov.co/viceministerio-de-vivienda/mi-casa-ya.

[4] Secretaría General del Senado, «Artículo 51. Constitución Política de Colombia,» Gaceta Constitucional No. 116 de 20 de julio de 1991, 1991. [En línea]. Available: http://www.secretariasenado.gov.co/senado/basedoc/constitucion_politica_1991_pr001.html#:~:text=ARTICULO%2051.,de%20estos%20programas%20de%20vivienda..

[5] DNP, «Plan Nacional de Desarrollo 2018-2022,» Departamento Nacional de Planeación, 2018. [En línea]. Available: https://www.dnp.gov.co/plan-nacional-desarrollo/Paginas/plan-nacional-de-desarrollo-2018-2022.aspx.

[6] Corte Constitucional, «Sentencia T-409/13,» Corte Constitucional, 2013. [En línea]. Available: https://www.corteconstitucional.gov.co/relatoria/2013/T-409-13.htm#:~:text=T%2D409%2D13%20Corte%20Constitucional%20de%20Colombia&text=El%20concepto%20de%20vivienda%20digna,desarrollar%20su%20proyecto%20de%20vida..

[7] Dane, «Déficit Habitacional Cali,» Departamento Administrativo Nacional de Estadística, 2023. [En línea]. Available: https://geoportal.dane.gov.co/geovisores/sociedad/deficit-habitacional/.

[8] V. Brodesohn, «Focalización de Programas de Superación de la Pobreza,» 1999. [En línea]. Available: http://www.iin.oea.org/cursos_a_distancia/focalizacion_de_los_programas.pdf.

[9] CEPAL, «Focalización y Pobreza: nuevas tendencias en la política social,» 1992. [En línea]. Available: https://repositorio.cepal.org/server/api/core/bitstreams/0e8fb6e6-69c7-4a33-ac68-b6f0a6571ed5/content.

[10] D. Raczynski, «Focalización de Programas Sociales: Lecciones de la Experiencia Chilena,» 2019. [En línea]. Available: https://www.cieplan.org/wp-content/uploads/2019/12/pol-econom-y-soc-chile-demo_Capitulo_10_compressed.pdf.

[11] V. Ortíz, «Aplicación de técnicas de aprendizaje a utomático para la segmentación y clasificación de características sociodemográficas asociadas a tasas de mortalidad infantil utilizando datos reportados por el DANE Colombia entre los años 2008 al 2017,» 2020. [En línea]. Available: https://expeditiorepositorio.utadeo.edu.co/bitstream/handle/20.500.12010/19171/Aplicacion_Modelos_Aprendizaje_Automatico_Ortiz.pdf?sequence=1&isAllowed=y.

[12] T. Mitchell, Machine Learning, New York: McGraw Hill, 1997.

[13] I. Salamanca y E. Junior, «Técnicas de aprendizaje automático aplicadas en los sistemas de predicción,» Revista Tecnología, Investigación y Academia -RedAvanzada –RITA, vol. 8, nº 1, pp. 39-53, 2021.

[14] N. Peláez, «Aprendizaje no supervisado y el algorítmo de Wake - Sleep en redes neuronales,» 2012. [En línea]. Available: http://jupiter.utm.mx/~tesis_dig/11612.pdf.

[15] S. Páez, «Análisis comparativo de herramientas open source para data mining sobre datos públicos del Ministerio de Educaciónde la República del Ecuador,» 2019. [En línea]. Available: http://repositorio.puce.edu.ec/bitstream/handle/22000/17060/Tesis%20Sergio%20P%C3%A1ez%20Doc.pdf?sequence=1&isAllowed=y.

[16] Congreso de Colombia, «Ley 388 de 1997,» 1997. [En línea]. Available: https://www.funcionpublica.gov.co/eva/gestornormativo/norma_pdf.php?i=339.

[17] G. Ortíz, «Establecimiento de un grupo de indicadores para la toma de decisiones en el sector vivienda utilizando técnicas de minería de datos,» Instituto Técnológico de Costa Rica, Cartago, 2003.

[18] M. Colino, «Minería de Datos en portal inmobiliario,» Universidad Carlos III de Madrid, Leganés, 2012.