Análisis Integral de Clasificación de Usuarios - SISBEN Año 2017

John Byron Alzate Hernández
Jorge Luis Genes Padilla

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesora
Daniela Serna Buitrago
Especialista en Analítica y Ciencia de Datos

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2024

Cita	(Alzate Hernández & Genes Padilla, 2024)
Referencia **Estilo IEEE (2023)**	Alzate Hernández, J.B., & Genes Padilla, J. L. (2024). *Análisis Integral de Clasificación de Usuarios - SISBEN Año 2017*. Trabajo de grado especialización. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte VI.

Centro de Investigación Ambientales y de Ingeniería (CIA).

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.
Decano: Julio Cesar Saldarriaga Molina.
Jefe departamento: Diego José Luis Botia Valderrama

Dedicatoria

A nuestras dos familias, cuyo apoyo nos permitió avanzar en el desarrollo del presente proyecto de forma más tranquila y centrada. Así mismo a nuestros profesores y asesores de la especialización, por su apoyo en la construcción de una línea de guía adecuada para la elaboración y desarrollo del tema del presente proyecto.

Agradecimientos

A nuestros compañeros de la especialización, con quienes en conjunto vivimos un proceso de aprendizaje que nos permitió compartir experiencias y conocimientos, ayudarnos y retroalimentamos, pero sobre todo volvernos grandes amigos y profesionales más comprometidos y competentes.

Tabla de contenido

Lista de tablas

Lista de figuras

Siglas, acrónimos y abreviaturas

SISBEN	Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales
ML	Machine Learning
DNP	Departamento Nacional de Planeación
UdeA	Universidad de Antioquia
Esp.	Especialista
MB	Megabytes
CRISP-DM	Cross-Industry Standard Process for Data Mining
Ovr	One-vs-Rest
Lbfgs	Limited-memory Broyden–Fletcher–Goldfarb–Shanno
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
AWS	Amazon Web Services
AI	Artificial Intelligence

Resumen

El impacto de las tecnologías de análisis de datos en la planificación urbana y social es una cuestión cada vez más relevante a nivel global. Un enfoque innovador en este campo es la aplicación de técnicas avanzadas de clasificación a los datos poblacionales, como se evidencia en el análisis de la Base de Datos SISBEN 2017 para el área metropolitana de Antioquia, enfocándose particularmente en Medellín. Este análisis emplea características demográficas y socioeconómicas transformadas en datos cuantitativos, que luego se analizan mediante modelos avanzados de aprendizaje automático.

El objetivo es optimizar la asignación del puntaje SISBEN, un indicador crucial para la distribución de recursos y servicios. La metodología no solo busca validar la eficacia del sistema existente, sino también explorar posibles mejoras para una clasificación más precisa y equitativa de los beneficiarios. Este enfoque cuantitativo abre nuevas perspectivas para una toma de decisiones más informada en el contexto de la política social y la gestión urbana, reflejando cómo la intersección entre la tecnología y las ciencias sociales puede enriquecer la comprensión y mejora de los sistemas de clasificación poblacional en la era digital.

Palabras clave: SISBEN, f1-score, Puntaje, ML, Precisión

Repositorios de GitHub:

https://github.com/johnbyronA/Monografia_Sisben_2017
https://github.com/jorgegenes23/Monografia_Sisben_2017

Abstract

The impact of data analysis technologies on urban and social planning is an increasingly relevant issue globally. An innovative approach in this field is the application of advanced classification techniques to population data, as evidenced in the analysis of the SISBEN 2017 Database for the Metropolitan Area of Antioquia, focusing particularly on Medellín. This analysis uses demographic and socioeconomic features transformed into quantitative data, which are then processed through advanced machine learning models.

The aim is to optimize the allocation of the SISBEN score, a crucial indicator for the distribution of resources and services. The methodology not only seeks to validate the effectiveness of the existing system but also explores possible improvements for a more accurate and equitable classification of beneficiaries. This quantitative approach opens new perspectives for more informed decision-making in the context of social policy and urban management, reflecting how the intersection between technology and social sciences can enrich the understanding and improvement of population classification systems in the digital age.

Keywords: SISBEN, f1-score, Score, ML, Accuracy

GitHub Repositories:

https://github.com/johnbyronA/Monografia_Sisben_2017
https://github.com/jorgegenes23/Monografia_Sisben_2017

Metodología CRISP-DM

La metodología **CRISP-DM**, diseñada para maximizar el valor extraído de manera organizada y efectiva. Se destaca por su adaptabilidad a diversos entornos y tipos de proyectos. Los elementos fundamentales a considerar son:

- **Revisión Permanente**: Mantener una constante evaluación de los modelos desarrollados, garantizando que cumplen con los criterios de calidad y efectividad establecidos.
- **Comprensión Empresarial Primordial**: Hacer énfasis en la importancia de entender a fondo el entorno empresarial antes de proceder con los análisis técnicos.
- **Colaboración con Interesados**: Comprender las expectativas y necesidades de los stakeholders es vital para definir correctamente el problema y asegurar la relevancia y utilidad de los resultados.
- **Proceso Cíclico**: Adoptar un enfoque no lineal que permita iteraciones para ajustarse a nuevos hallazgos o cambios durante el proyecto, facilitando una mejor adaptación.

CRISP-DM se estructura en seis etapas claves que abarcan desde el análisis inicial del problema hasta la implementación final del modelo. Estas etapas son:

1. **Análisis del Problema**: Se identifican y definen los problemas y objetivos del proyecto, explorando el contexto empresarial y los datos que están disponibles.
2. **Exploración de Datos**: Se realiza un análisis preliminar de los datos para evaluar su calidad y determinar su pertinencia para los objetivos del proyecto.
3. **Transformación de Datos**: Esta fase implica la limpieza y modificación de los datos para asegurar que están en formato óptimo para el modelado.
4. **Desarrollo de Modelos**: Se elige y calibra el modelo estadístico o de machine learning más adecuado basado en el problema y los datos procesados.
5. **Validación del Modelo**: El modelo se somete a pruebas para verificar su eficacia y su capacidad para abordar y resolver el problema especificado inicialmente.
6. **Despliegue del Modelo**: Finalmente, el modelo se implementa en un entorno operativo real donde se monitorea continuamente su desempeño para garantizar su eficiencia y utilidad continuas.

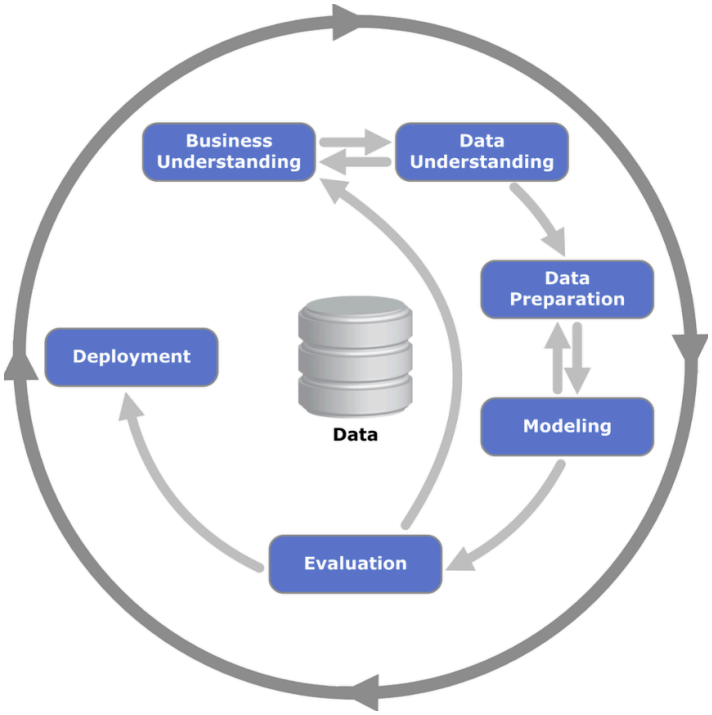

Figura 1. Ciclo vital Metodología CRISP-DM.

El ciclo del modelo está diseñado para ser flexible y adaptativo, permitiendo que los proyectos fluyan de manera dinámica a través de sus diferentes etapas.

Las conexiones entre cada fase, representadas por flechas, ilustran la interdependencia y el intercambio continuo de información. Por ejemplo, los descubrimientos durante la fase de despliegue pueden revelar la necesidad de revisar las etapas de comprensión del negocio, promoviendo así una regresión temporal para refinar los enfoques iniciales. Esta metodología iterativa y cíclica promueve una adaptabilidad superior frente a cambios o desafíos no anticipados, garantizando que el proyecto se mantenga relevante y efectivo frente a demandas y circunstancias nuevas que surjan.

1. Comprensión del negocio

El Sistema de Identificación y Selección de Beneficiarios de Programas Sociales (SISBEN) es una herramienta crítica en Colombia para la focalización y asignación de subsidios y beneficios sociales. Su objetivo principal es identificar a las personas y familias en situación de vulnerabilidad económica para asegurar que los recursos del estado se distribuyan de manera equitativa y eficiente.

Funcionamiento del SISBEN:

El SISBEN clasifica a la población en diferentes niveles según su situación socioeconómica, utilizando un puntaje basado en diversas variables como ingresos, condiciones de vivienda, acceso a servicios públicos y bienes poseídos, entre otros. Estos datos se recogen mediante encuestas y se procesan para asignar un puntaje que determina la elegibilidad para distintos programas sociales.

Importancia del Sistema:

1. **Justicia Social:** Permite que los recursos del estado lleguen a quienes más los necesitan, asegurando una distribución equitativa y justa.
2. **Eficiencia:** Optimiza el uso de los recursos públicos, evitando su desperdicio en individuos que no cumplen los criterios de vulnerabilidad.
3. **Transparencia:** Promueve la transparencia en la asignación de subsidios y beneficios sociales, reduciendo la posibilidad de corrupción y malversación de fondos.

A pesar de los beneficios que ofrece el SISBEN, existen varios desafíos que pueden comprometer su eficacia:

1. Precisión de los datos: La calidad de los datos es fundamental para el correcto funcionamiento del sistema. Si las personas proporcionan información inexacta o falsa sobre sus ingresos y condiciones de vida, esto puede llevar a una clasificación errónea y a la asignación indebida de beneficios.

2. Detección de alteraciones: las alteraciones son una preocupación constante. Las personas pueden intentar manipular el sistema proporcionando información incorrecta para obtener

beneficios a los que no tienen derecho. Esto no solo desvirtúa el propósito del sistema, sino que también priva de recursos a aquellos que realmente los necesitan.

3. Complejidad y Diversidad de Datos: El sistema maneja una gran cantidad de variables categóricas y numéricas, lo que añade complejidad al proceso de análisis y clasificación. Esto requiere herramientas y técnicas avanzadas para procesar y analizar los datos de manera efectiva.

Estos desafíos subrayan la necesidad de un análisis más preciso y detallado sobre los parámetros de calidad de vida a nivel local, y así asegurar que el sistema funcione de manera justa y eficiente en la ciudad de Medellín.

Con base en lo anterior, se escoge reevaluar la medición del Puntaje SISBEN, un dato fundamental para la clasificación poblacional con la finalidad del acceso a beneficios sociales, la cual enfrenta serios desafíos debido a la presencia de múltiples valores atípicos dentro de las variables que componen el modelo. Estos valores atípicos distorsionan la variable de salida (Puntaje SISBEN), resultando en clasificaciones inexactas y poco representativas de la realidad socioeconómica de los hogares. Esta situación no sólo mina la eficiencia del sistema, sino que también perpetúa la desigualdad al privar a los verdaderos necesitados de los beneficios a los que tienen derecho. Por tanto, es crucial implementar métodos de análisis y modelos predictivos más robustos que puedan identificar y mitigar estas anomalías, obteniendo como resultado una mejor clasificación poblacional y de paso, asegurando una distribución más justa y efectiva de los recursos públicos.

2. Objetivos

2.1. Objetivo general

Desarrollar un sistema de clasificación basado en machine learning para el puntaje SISBEN, que optimice la toma de decisiones en cuanto a la asignación de beneficios sociales al identificar con mayor precisión la situación socioeconómica de los individuos, minimizando errores y asegurando que los datos estudiados sean precisos y reflejen mejor la realidad.

2.2. Objetivos específicos

- Identificar las variables que más influyen en la asignación del puntaje SISBEN.
- Implementar técnicas de análisis de datos y modelos predictivos para identificar valores atípicos en los datos del SISBEN, que faciliten la integridad y la justicia en la distribución de beneficios.
- Aplicar herramientas de machine learning y técnicas de preprocesamiento de datos para mejorar la eficiencia y precisión en el análisis de las variables socioeconómicas escogidas que influyen en el puntaje SISBEN.
- Utilizar métricas de evaluación y técnicas de validación cruzada para garantizar que los modelos desarrollados sean robustos y precisos, proporcionando una base sólida para la toma de decisiones en la asignación de beneficios sociales.

3. Comprensión de los datos

3.1. Datos originales

La base de datos se encuentra en la página de MEData, estrategia de datos de la ciudad de Medellín, para acceder a ella no es necesario tener usuario y contraseña para ingresar, lo que garantiza total accesibilidad a los datos. Esta posee información demográfica de los habitantes de la ciudad de Medellín, es obtenida en formato .CSV y pesa aproximadamente 118 MB de peso.

Para efectos de la presente investigación, se seleccionaron variables que son consideradas directamente relevantes para el cálculo del puntaje SISBEN y que impactan significativamente en la clasificación socioeconómica, como ingresos y acceso a servicios públicos, lo anterior se hace debido a que se considera que existe una alta redundancia y correlación de variables en los datos debido a similitudes en el contexto de la información, como se puede apreciar en los siguientes ejemplos:

- **PARED y PISO**: Eliminadas ya que la calidad de la vivienda se puede representar suficientemente con la variable **VIVIENDA**.
- **TVCABLE y HORNO**: Se eliminaron estas variables, ya que **TVCOLOR** y **CALENTA** ya cubren suficientemente los electrodomésticos y servicios de entretenimiento.
- **AIRE y MOTO**: Se eliminaron ya que la propiedad de un vehículo se cubre suficientemente con la variable **AUTO1**.

A continuación se presentarán los datos utilizados para el análisis general, así como también una descripción detallada y el tipo de dato:

Nombre Columna	**Tipo**	**Descripción de la variable**
COMUNA	Object	Código de la comuna donde se encuentra ubicada la unidad de vivienda.
TELÉFONO	Object	La vivienda cuenta con servicio de teléfono. 1=SI, 2=NO.
VIVIENDA	Object	Tipo de vivienda. 1 = Casa o apartamento 2 = Cuarto 3 = Otro tipo de unidad de vivienda 4 = Casa Indígena

ENERGIA	Object	La vivienda cuenta con servicio público de energía eléctrica. 1=SI, 2=NO.
ALCANTA	Object	La vivienda cuenta con el servicio público de alcantarillado. 1=SI, 2=NO.
GAS	Object	La vivienda cuenta con el servicio público de gas. 1=SI, 2=NO.
ACUEDUC	Object	La vivienda cuenta con el servicio público de acueducto. 1=SI, 2=NO.
ESTRATO	Object	Estrato socioeconómico de la vivienda.
TENEVIV	Object	Este hogar vive en: 1 = Arriendo 2 = Propia pagando 3 = Propia pagada 4 = Otra condición.
AGUA	Object	1 = Acueducto 2 = Pozo con bomba 3 = Pozo sin bomba, jagüey 4 = Agua lluvia 5 = Río, quebrada, manantial, nacimiento 6 = Pila pública 7 = Carrotanque 8 = Aguatero 9 = Donación
NEVERA	Object	El hogar cuenta con nevera o enfriador: 1=SI, 2=NO.
LAVADORA	Object	El hogar cuenta con lavadora: 1=SI, 2=NO.
TVCOLOR	Object	El hogar cuenta con televisor: 1=SI, 2=NO.
CALENTA	Object	El hogar cuenta con calentador de agua o ducha eléctrica: 1=SI, 2=NO.
COMPUTADOR	Object	El hogar cuenta con computador: 1=SI, 2=NO.
AUTO1	Object	El hogar cuenta con automóvil para uso del hogar: 1=SI, 2=NO.
SEXO	Object	Sexo.
ESTCIVIL	Object	Estado civil.
ASISTE	Object	Asiste a centro educativo: 1=SI, 2=NO.
NIVEL	Object	Nivel educativo alcanzado:
INGRESOS	Float64	Total de ingresos mensuales:
PERCIBE	Object	Percibe ingresos (laborales, arriendos, subsidios, transferencias, en especie)
PUNTAJE	Float64	Valor entre cero (0) y cien (100).
EDAD	Int64	Edad del usuario.

Tabla 1: Definición de variables del sistema.

3.2. Analítica descriptiva

El Dataset resultante se compone de un total de 22 variables originales y fueron creadas dos variables adicionales: **AgeCategory,** la cual contiene el rango de edad al que pertenece el usuario y **Puntaje_categorico,** la cual contiene el rango de puntaje al que pertenece el usuario. Teniendo esto en cuenta, se toman para la transformación a variables categóricas, aquellas que son de tipo object, y para transformación a variables numéricas, aquellas de tipo **float64** e **int64**.

Como variable de salida, se establece la variable **Puntaje_categorico,** la cual toma valores de Bajo (correspondiente a los usuarios con puntaje entre 0 y 50 puntos) y Alto (correspondiente a los usuarios con puntaje entre 51 y 100 puntos). A continuación y para propósitos de análisis descriptivo, se tiene el gráfico de torta para la variable **Puntaje SISBEN** y el gráfico de dispersión para la variable **INGRESOS**, los cuales permitirán identificar de forma visual el comportamiento de los puntajes asignados y los ingresos entre los diferentes estratos de la población.

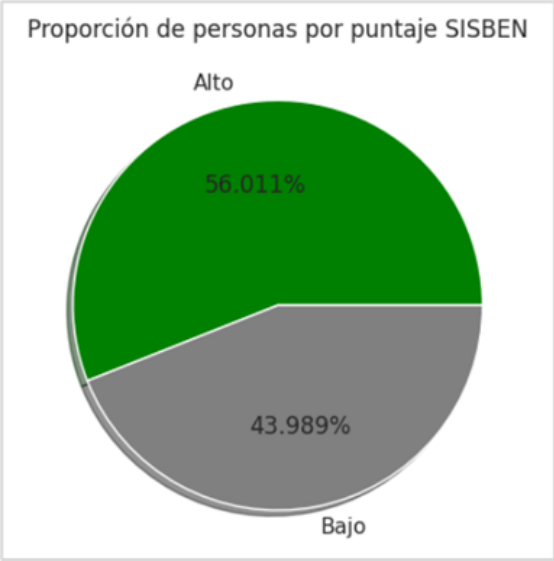

Figura 2. Distribución de los usuarios de acuerdo a la categoría de Puntaje.

La **Figura 2** muestra la distribución de usuarios según la categoría del puntaje SISBEN. Un 56.011% de los usuarios se encuentran en la categoría "Alto", mientras que el 43.989% restante se clasifica en la categoría "Bajo".

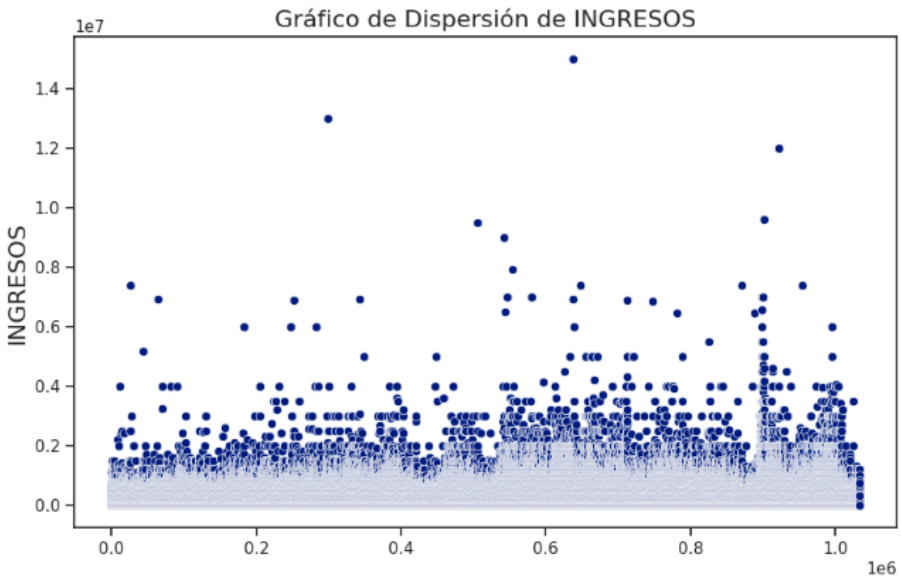

Figura 3. Gráfico de dispersión de la variable INGRESOS vs. Cantidad de usuarios.

La **Figura 3** muestra la distribución de los ingresos de los usuarios del SISBEN. Los puntos en el gráfico representan la cantidad de observaciones o individuos en el dataset, discriminados por sus ingresos. Se observa una gran concentración de puntos en valores bajos de ingresos, con algunos valores atípicos que se alejan considerablemente hacia ingresos muy altos, indicando posibles desviaciones en los datos (En el gráfico de dispersión, la escala de medición de ingresos para el eje Y, está en decenas de millones de pesos, donde 0.2 equivale a 2 Millones de pesos, y 1.4 a 14 millones de pesos).

Se escogieron las **Figuras 2 y 3** para esta sección porque se consideró que la información otorgada por ellas es crucial para entender la distribución inicial de los datos y la categorización de la población objetivo, las otras variables, su comportamiento y su desglose gráfico, pueden ser encontradas en el archivo **Codigo_Monografia_SISBEN_2017_JohnAlzate_JorgeGenes** con extensión **.ipynb** anexo a la presente monografía para su entrega final a la universidad.

Adicionalmente, se escoge realizar un análisis descriptivo de la variable INGRESOS vs. otras variables del modelo, consideradas de las más sensibles, según sus resultados en matriz de correlación, el cual puede ser visto a continuación.

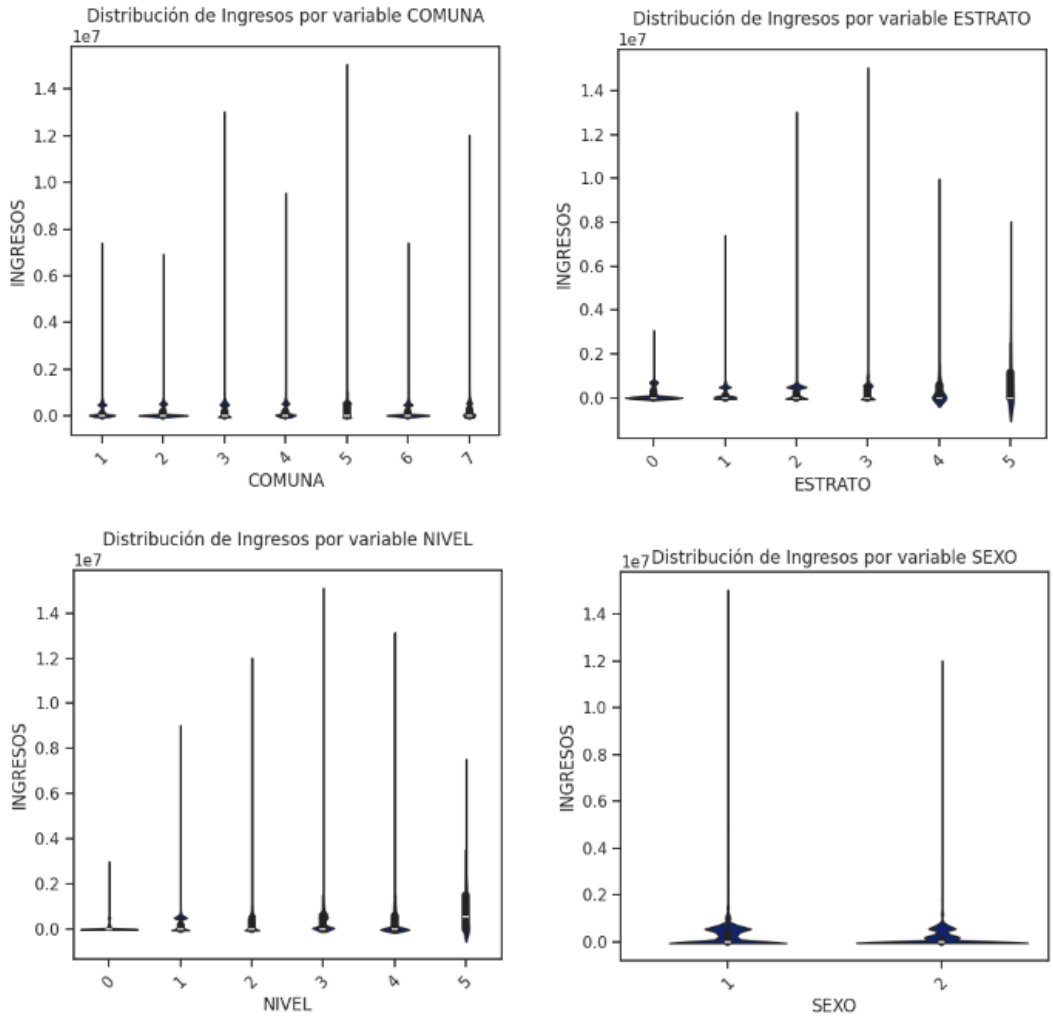

Figura 4. Gráficos de violín de Variables vs. INGRESOS.

Distribución de Ingresos por Comuna: Crucial para identificar diferencias económicas significativas entre diferentes áreas geográficas.

Distribución de Ingresos por variable Estrato: Esencial para entender cómo la categorización socioeconómica impacta los ingresos, ayudando a desarrollar estrategias de inclusión y equidad económica.

Distribución de Ingresos por variable Nivel: Importante para identificar cómo la educación influye en la capacidad de ingresos.

Distribución de Ingresos por variable Sexo: Fundamental para detectar brechas de género en los ingresos.

Para entender mejor las relaciones entre las variables socioeconómicas en el contexto del SISBEN, se emplea la matriz de correlación. Esta herramienta estadística permite identificar cómo estas variables se interrelacionan entre sí, destacando posibles patrones y dependencias.

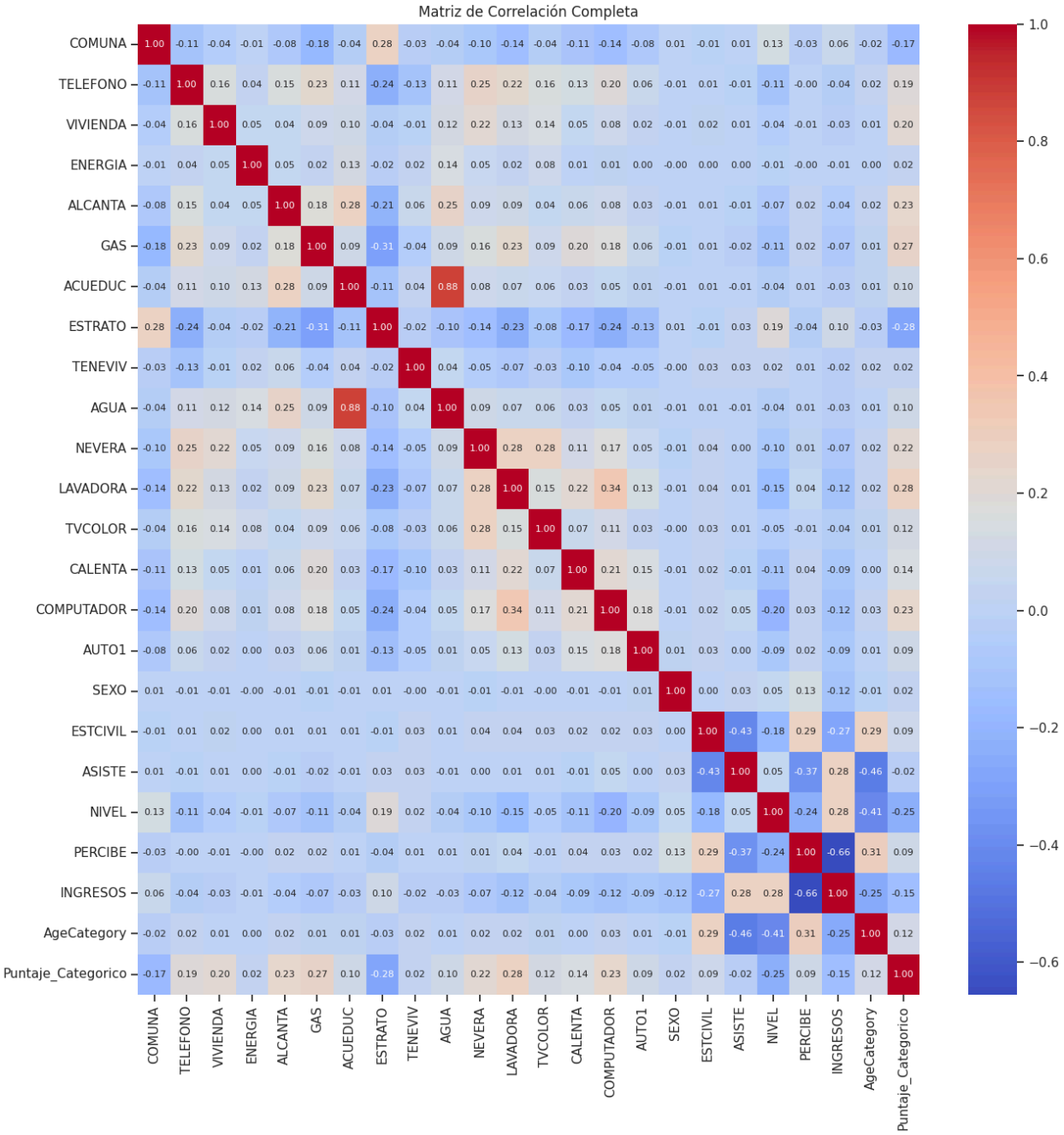

Figura 5. Matriz de correlación de los datos SISBEN.

En el análisis se detectó que algunas variables como **ACUEDUC** (Disponibilidad de Acueducto) y **AGUA** (Disponibilidad del servicio de agua) presentan alta correlación, lo cual podría afectar la precisión de los modelos predictivos, complicando la interpretación y potencialmente inflando los errores de estimación.

4. Preparación de los datos

Para garantizar la calidad y relevancia de los datos antes de aplicar los modelos de Machine Learning, se consideraron varias técnicas de preprocesamiento, entre ellas:

1. **Tratamiento de Valores Nulos**: Se investigó si había valores nulos en las variables numéricas y categóricas. Al no haber valores nulos, no se requirió hacer un tratamiento de los mismos.
2. **Conversión de Tipos de Datos**: Se aseguraron que las variables categóricas estuvieran en el tipo de datos correcto (categoría), y se convirtieron las variables numéricas a float64 para un manejo adecuado en los modelos.
3. **Codificación de Variables Categóricas**: Se emplearon técnicas de codificación como One-Hot Encoding y Label Encoding, asegurando que los modelos de ML puedan procesar estas variables correctamente.
4. **Escalado de Variables**: Para que los algoritmos de ML funcionen de manera eficiente, se aplicó el escalado MinMaxScaler a las variables numéricas, normalizando sus rangos entre 0 y 1.
5. **Aumentación de Datos**: Se emplearon técnicas de sobremuestreo como el Random OverSampler para balancear el dataset, igualando la cantidad de ejemplos en cada clase y mejorando la capacidad del modelo para generalizar.
6. **División del Dataset**: Finalmente, el dataset se dividió en conjuntos de entrenamiento y prueba (70% y 30% respectivamente) para validar la eficacia de los modelos.

5. Modelado

Se consideraron y evaluaron diversos modelos de Machine Learning para determinar cuál se desempeñaba mejor en la clasificación del puntaje SISBEN. Los modelos considerados fueron:

1. **Regresión Logística**:
    - **Configuración**: Se usó la configuración multi_class="ovr" con solver ‘lbfgs’.
    - **Motivación**: Es un modelo interpretativo y eficiente para problemas de clasificación binaria.
2. **Random Forest**:
    - **Configuración**: Variando el número de árboles (n_estimators), la profundidad máxima (max_depth), y el criterio de división (gini y entropy).
    - **Motivación**: Proporciona robustez y mejora la precisión.
3. **Árboles de Decisión**:
    - **Configuración**: Configurado con una profundidad máxima de 5, con 30 nodos terminales y utilizando el criterio 'gini'.
    - **Motivación**: Facilita la interpretación y visualización de la toma de decisiones.
4. **HistGradient Boosting Classifier**:
    - **Configuración**: Optimización de hiperparámetros como la tasa de aprendizaje (learning_rate), el número de iteraciones (max_iter), y la profundidad máxima (max_depth).
    - **Motivación**: Combina múltiples árboles de decisión para mejorar la precisión y manejar datos complejos de manera eficiente.

Cada uno de estos modelos fue evaluado utilizando métricas de desempeño como la exactitud (accuracy), el área bajo la curva ROC (AUC), la precisión, y el recall. La elección del mejor modelo se basó en un balance entre estas métricas, la capacidad del modelo para manejar los datos y su interpretabilidad.

Para evaluar el desempeño de los modelos de Machine Learning y su impacto en el negocio, se utilizaron diversas métricas calculadas con funciones de las bibliotecas sklearn y matplotlib. A continuación, se describen las principales métricas y las funciones utilizadas para calcularlas:

Métricas de desempeño de Machine Learning

Exactitud (Accuracy):

- **Descripción**: Proporción de predicciones correctas sobre el total de predicciones.
- **Función**: accuracy_score de sklearn.metrics

Precisión (Precision):

- **Descripción**: Proporción de verdaderos positivos sobre el total de positivos predichos.
- **Función**: precision_score de sklearn.metrics

Recall (Sensibilidad):

- **Descripción**: Proporción de verdaderos positivos sobre el total de positivos reales.
- **Función**: recall_score de sklearn.metrics

F1-Score:

- **Descripción**: Media armónica de la precisión y el recall, útil para balancear entre falsos positivos y falsos negativos.
- **Función**: f1_score de sklearn.metrics

Área Bajo la Curva ROC (AUC-ROC):

- **Descripción**: Mide la capacidad del modelo para diferenciar entre clases. Cuanto más cerca de 1, mejor.
- **Función**: roc_auc_score de sklearn.metrics

Matriz de Confusión:

- **Descripción**: Muestra la cantidad de predicciones correctas e incorrectas desglosadas por cada clase.
- **Función**: confusion_matrix de sklearn.metrics

5.1. Regresión Logística OVR - lbfgs

El primer modelo utilizado fue una Regresión Logística con el método One-vs-Rest (OVR) y el solver 'lbfgs' para la clasificación de las variables del dataset SISBEN. Este modelo fue entrenado con el 70% de los datos y evaluado con el 30% restante, con el objetivo de predecir el puntaje SISBEN categorizado en "Alto" y "Bajo" utilizando las variables socioeconómicas vistas anteriormente.

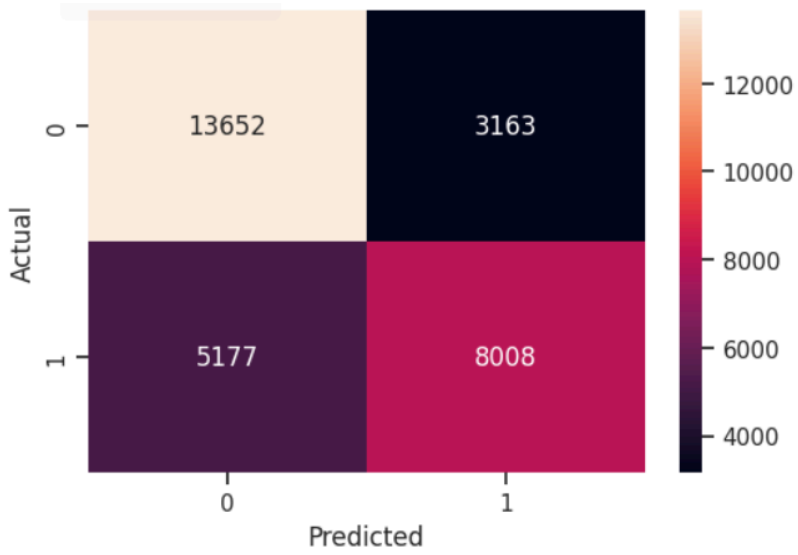

Figura 6. Matriz de confusión modelo Regresión Logística OVR - lbfgs.

	precision	recall	f1-score	support
0	0.73	0.81	0.77	16815
1	0.72	0.61	0.66	13185
accuracy			0.72	30000
macro avg	0.72	0.71	0.71	30000
weighted avg	0.72	0.72	0.72	30000

Figura 7. Resultados de las métricas del modelo Regresión Logística OVR - lbfgs.

La matriz de confusión mostrada en la **Figura 6** muestra que 13,652 individuos fueron correctamente clasificados como "Bajo" (verdaderos negativos), mientras que 3,163 personas fueron incorrectamente clasificadas como "Bajo" cuando deberían haber sido "Alto" (falsos positivos). Además, 5,177 personas fueron incorrectamente clasificadas como "Alto" (falsos negativos) y 8,008 individuos fueron correctamente identificados como "Alto" (verdaderos positivos). De igual forma, la **Figura 7** muestra que el modelo alcanzó una precisión del 72%, indicando que el 72% de las clasificaciones del puntaje SISBEN fueron correctas. El f1-score de 0.77 para "Bajo" y 0.66 para "Alto" sugiere que, similar al primer modelo, este modelo tiene un mejor desempeño en identificar correctamente los casos de "Bajo" que los de "Alto".

5.2. Random Forest

Para el modelo de Random Forest, se realizó una búsqueda de hiperparámetros utilizando GridSearchCV para identificar la mejor configuración posible. El modelo fue entrenado y evaluado con el 70% y 30% de los datos, respectivamente. La configuración utilizada fue `max_depth=10`, `max_features=7`, `n_estimators=150`, y `criterion='gini'`, alcanzando una precisión de validación cruzada del 72.45%.

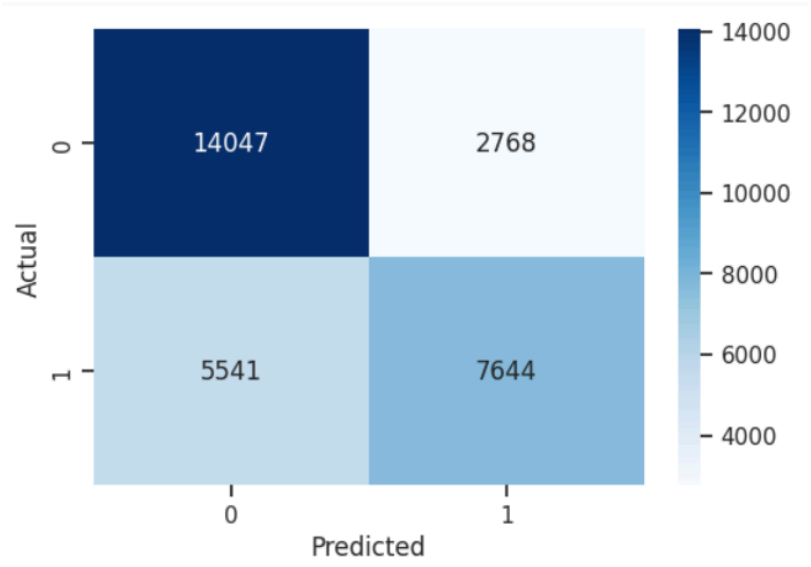

Figura 8. Matriz de confusión modelo Random Forest.

Reporte de clasificación para Random Forest:

	precision	recall	f1-score	support
0	0.72	0.84	0.77	16815
1	0.73	0.58	0.65	13185
accuracy			0.72	30000
macro avg	0.73	0.71	0.71	30000
weighted avg	0.72	0.72	0.72	30000

Figura 9. Resultados de las métricas del modelo Random Forest.

La matriz de confusión mostrada en la **Figura 8** muestra que 14,047 individuos fueron correctamente clasificados como "Bajo" (verdaderos negativos), mientras que 2,768 fueron incorrectamente clasificados como "Bajo" cuando deberían haber sido "Alto" (falsos positivos). Asimismo, 5,541 personas fueron incorrectamente clasificadas como "Alto" (falsos negativos) y 7,644 fueron correctamente identificadas como "Alto" (verdaderos positivos). Desigual forma, la **Figura 9** muestra que este modelo obtuvo una precisión del 72% en el conjunto de prueba, lo que indica que el 72% de las predicciones fueron correctas. El f1-score fue de 0.77 para "Bajo" y 0.65 para "Alto", lo que sugiere que el modelo es más efectivo en identificar correctamente los casos de "Bajo" que los de "Alto".

Estos resultados indican que el modelo de Random Forest tiene un rendimiento adecuado, pero aún existen áreas de mejora, especialmente en la correcta clasificación de individuos que deberían recibir un puntaje "Alto" en el SISBEN, por lo cual se continuará la experimentación de más modelos.

5.3. Árboles de Decisión

Se empleó también un modelo de Árbol de Decisión con una profundidad Depth máxima de 5 niveles y el criterio de Gini para clasificar las variables del dataset SISBEN, el cual puede verse en la **Figura 10**. Este modelo también fue entrenado con el 70% de los datos y evaluado con el 30% restante, con el objetivo de predecir el puntaje SISBEN categorizado en "Alto" y "Bajo" utilizando las variables socioeconómicas disponibles.

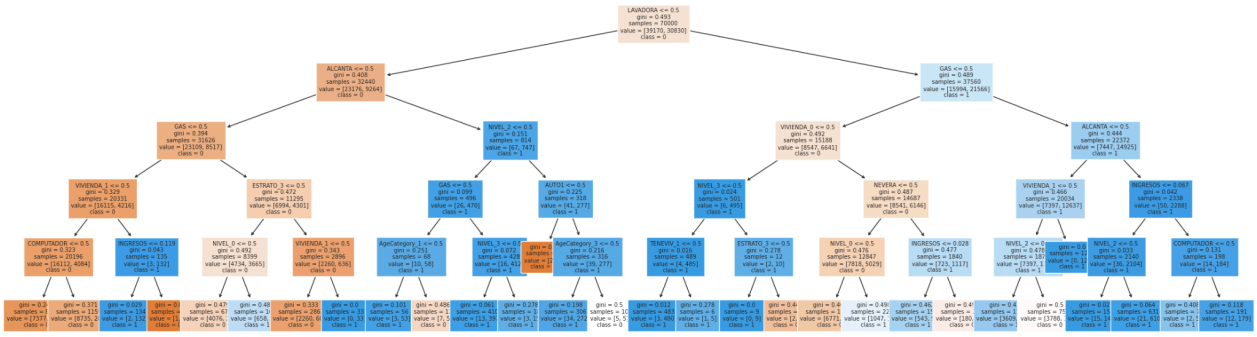

Figura 10. Gráfico de árboles de decisión.

El árbol de decisión muestra cómo diversas variables socioeconómicas se utilizan para predecir el puntaje SISBEN de los individuos. Los principales divisores son la disponibilidad de servicios básicos como alcantarillado (ALCANTA) y gas, seguidos por el nivel educativo (NIVEL), la propiedad de bienes (AUTO1, COMPUTADOR), y la edad (AgeCategory). La mayor parte de los nodos terminales indican que la falta de acceso a estos servicios y bienes está correlacionada con un puntaje SISBEN bajo. La escala de colores ilustra la pureza de cada nodo, donde los nodos azules representan mayormente predicciones correctas de puntaje alto y los nodos naranjas predicciones correctas de puntaje bajo.

Reporte de clasificación para el Árbol de Decisión:

	precision	recall	f1-score	support
0	0.69	0.84	0.76	16815
1	0.72	0.51	0.60	13185
accuracy			0.70	30000
macro avg	0.70	0.68	0.68	30000
weighted avg	0.70	0.70	0.69	30000

Exactitud del modelo de Árbol de Decisión: 0.6970333333333333

Figura 11. Resultados de las métricas del modelo Árboles de Decisión.

En la **Figura 11** pueden verse los resultados de las métricas, en la que se puede constatar que el modelo de Árbol de Decisión alcanzó una precisión del 69.70%, lo que implica que el 69.70% de las clasificaciones del puntaje SISBEN fueron correctas. De igual forma, el f1-score de 0.76 para "Bajo" y 0.60 para "Alto" indica que el modelo tiene un mejor desempeño en identificar correctamente los casos de "Bajo" en comparación con los casos de "Alto".

5.4. HistGradient Boosting Classifier

Para concluir sobre el modelo de HistGradient Boosting Classifier utilizado en la clasificación del puntaje SISBEN, se empleó un Grid Search con validación cruzada para identificar los mejores hiperparámetros. El modelo se entrenó con el 70% de los datos y se evaluó con el 30% restante. El objetivo fue predecir el puntaje SISBEN categorizado en "Alto" y "Bajo" utilizando variables socioeconómicas del dataset.

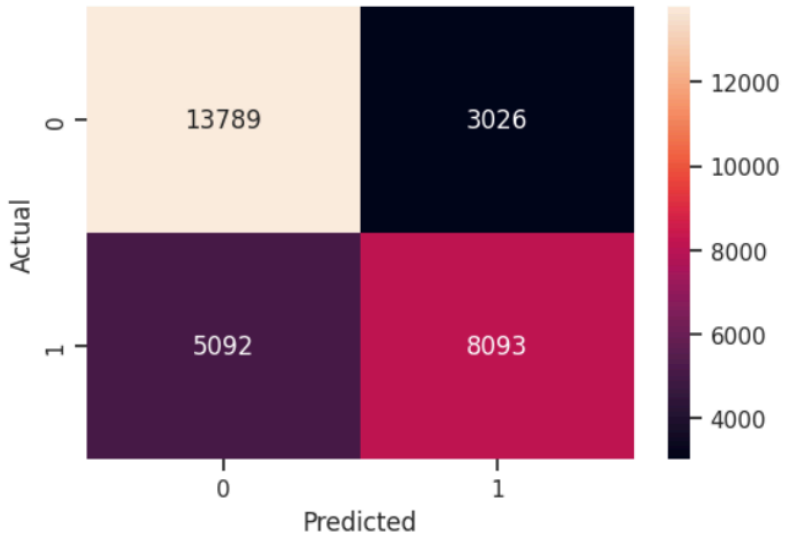

Figura 12. Matriz de confusión modelo HistGradient Boosting Classifier.

La matriz de confusión mostrada en la **Figura 12** evidencia que 13,789 individuos fueron correctamente clasificados como "Bajo" (verdaderos negativos), mientras que 3,026 personas fueron incorrectamente clasificadas como "Bajo" cuando deberían haber sido "Alto" (falsos positivos). Además, 5,092 personas fueron incorrectamente clasificadas como "Alto" (falsos negativos) y 8,093 individuos fueron correctamente identificados como "Alto" (verdaderos positivos).

Reporte de clasificación para HistGradient Boosting Classifier:

	precision	recall	f1-score	support
0	0.73	0.82	0.77	16815
1	0.73	0.61	0.67	13185
accuracy			0.73	30000
macro avg	0.73	0.72	0.72	30000
weighted avg	0.73	0.73	0.73	30000

Figura 13. Resultados de las métricas del modelo HistGradient Boosting Classifier.

Asimismo, la **Figura 13** muestra que este modelo alcanzó una precisión del 73%, indicando que el 73% de las clasificaciones del puntaje SISBEN fueron correctas. El f1-score de 0.77 para "Bajo" y 0.67 para "Alto" refleja que el modelo tiene un desempeño equilibrado en la clasificación de ambos casos.

Lo anterior indica que el modelo de HistGradient Boosting Classifier tiene un desempeño sólido en la clasificación de individuos para el puntaje SISBEN, con una buena precisión general.

6. Evaluación

Para hacer una comparación entre los modelos, y saber cuál es el mejor de ellos, se hace uso de la curva ROC:

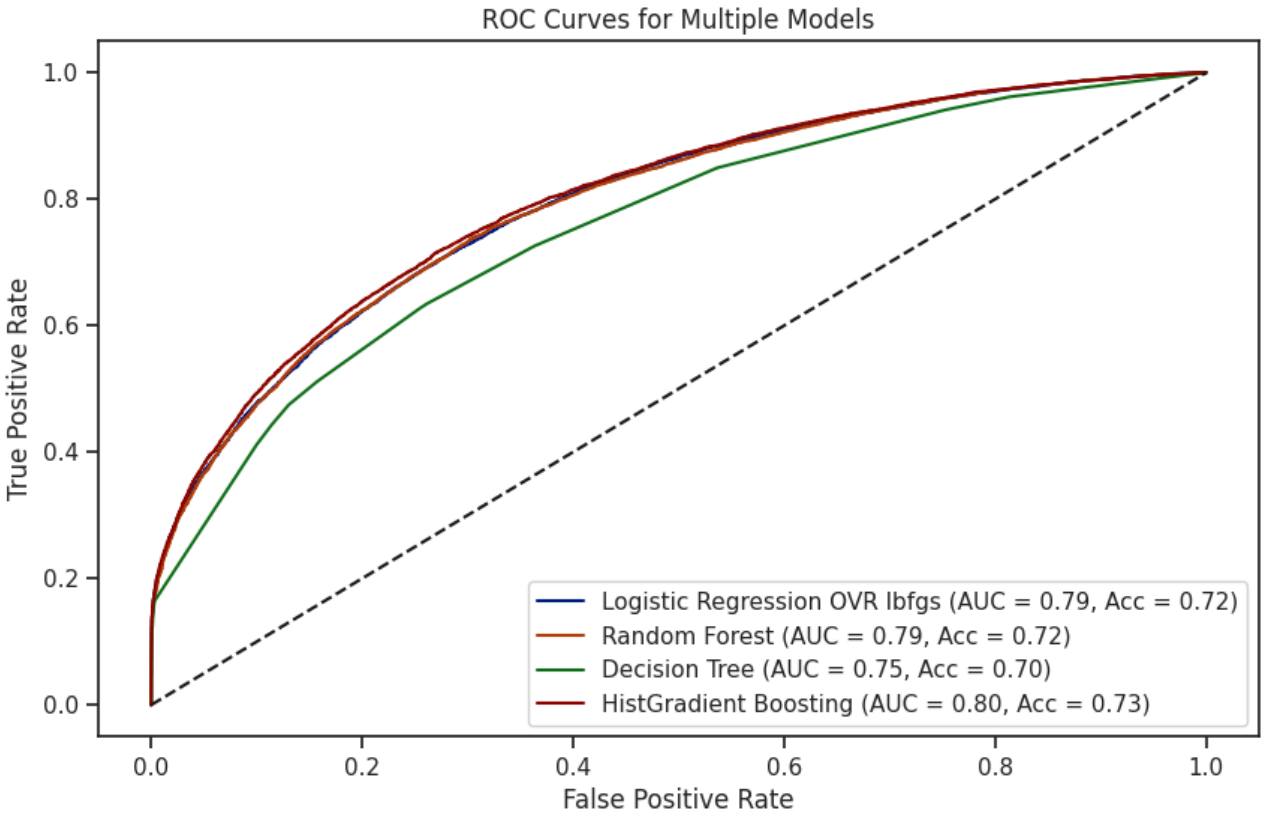

Figura 14. Curva ROC de comparación de efectividad de los modelos.

	MLA Name	MLA Train Accuracy	MLA Test Accuracy	MLA Precision	MLA Recall	MLA AUC
0	Logistic Regression OVR lbfgs	72.3886	72.1767	71.4298	61.1528	79.4529
1	Random Forest	73.1743	72.3567	73.1936	58.5438	79.3533
2	Decision Tree	69.8814	69.7100	71.8257	51.1414	75.4757
3	HistGradient Boosting	74.2271	72.9400	72.7853	61.3804	80.0493

Figura 15. Comparación de las métricas de los cuatro modelos utilizados.

El análisis de la **Figura 14 y Figura 15** muestran que el modelo HistGradient Boosting Classifier es el más adecuado para clasificar correctamente los puntajes SISBEN, con un AUC de 0.80 y una precisión de 0.73. Esto sugiere que este modelo ofrece la mejor capacidad para distinguir entre los individuos que deberían recibir puntajes "Alto" o "Bajo" en el SISBEN, asegurando así una asignación más justa y eficiente de los recursos sociales. Al implementar este modelo, se espera una mejora en la precisión de la clasificación, lo que reduce la posibilidad de errores y garantiza que los beneficios sociales lleguen a quienes realmente los necesitan. La mayor efectividad del HistGradient Boosting Classifier en comparación con otros modelos respalda su elección como la herramienta principal para mejorar el sistema SISBEN.

La evaluación cualitativa de los resultados revela que los modelos Random Forest y Decision Tree presentaron señales de overfitting, dado que sus precisiones de entrenamiento fueron significativamente mayores que las de prueba. En contraste, los modelos de Regresión Logística y HistGradient Boosting mostraron un mejor equilibrio entre entrenamiento y prueba, indicando menor riesgo de overfitting. El HistGradient Boosting destacó por su precisión y AUC, mostrando una fuerte relación entre las métricas de ML y la métrica de negocio, lo que sugiere su alta utilidad para clasificar correctamente los puntajes SISBEN y asegurar una distribución más equitativa de los recursos sociales. Esto resalta su capacidad para mejorar la precisión del sistema y la eficacia en la asignación de beneficios.

Evaluación cualitativa

La evaluación cualitativa de los resultados revela que los modelos Random Forest y Decision Tree presentaron señales de overfitting, dado que sus precisiones de entrenamiento fueron significativamente mayores que las de prueba. En contraste, los modelos de Regresión Logística y HistGradient Boosting mostraron un mejor equilibrio entre entrenamiento y prueba, indicando menor riesgo de overfitting. El HistGradient Boosting destacó por su precisión y AUC, mostrando una fuerte relación entre las métricas de ML y la métrica de negocio, lo que sugiere su alta utilidad para clasificar correctamente los puntajes SISBEN y asegurar una distribución más equitativa de los recursos sociales.

7. Consideraciones para despliegue de producción

Para que el Departamento Nacional de Planeación (DNP) ponga en marcha la producción de los modelos de machine learning, se deben considerar varios aspectos clave en términos de infraestructura, estrategias y operatividad:

1. Se debe establecer un sistema de monitoreo que rastree continuamente el desempeño del modelo escogido, detectando desviaciones en precisión, recall y AUC. Herramientas como Prometheus y Grafana pueden ser útiles para esta tarea.

2. Implementar herramientas como MLflow o Kubeflow para crear pipelines automatizados que faciliten la gestión, reentrenamiento y despliegue de modelos, garantizando que las nuevas versiones se integren sin interrupciones.

3. Dado el costo computacional y la complejidad de los modelos Random Forest y HistGradient Boosting, se recomienda el uso de servicios en la nube como AWS SageMaker o Google Cloud AI Platform, que ofrecen escalabilidad y capacidad de procesamiento en tiempo real.

4. La integración con streams de datos, como Apache Kafka, permitiría la actualización continua del modelo con nuevos datos, mejorando su precisión y relevancia.

Para asegurar la robustez del sistema, es esencial realizar evaluaciones periódicas de reentrenamiento del modelo, utilizando técnicas de validación cruzada para confirmar su estabilidad.

Es también importante considerar la optimización de costos computacionales, posiblemente empleando modelos más ligeros para predicciones rápidas en casos donde la precisión extrema no sea crítica. Además, la implementación debe contemplar medidas de seguridad para proteger los datos sensibles del SISBEN, asegurando el cumplimiento de regulaciones de privacidad y protección de datos.

8. Conclusiones

En resumen, el proyecto logró exitosamente el objetivo de desarrollar un modelo preciso para clasificar el puntaje SISBEN en categorías de alto y bajo, utilizando datos socioeconómicos. A través de un enfoque metodológico que incluyó el balanceo de datos, la creación de variables dummies y la normalización, se construyeron modelos robustos y efectivos. Este esfuerzo no solo demostró habilidades avanzadas en minería de datos y aprendizaje automático, sino que también proporcionó soluciones prácticas para la toma de decisiones en políticas públicas.

El modelo más eficaz, HistGradient Boosting, demostró un rendimiento sólido, aunque existe potencial para mejoras adicionales. Por ejemplo, se podría mejorar la precisión del modelo mediante la inclusión de más datos de entrenamiento y la evaluación de nuevas variables. La implementación de reentrenamiento periódico y la integración de sistemas de monitoreo continuo garantizarían la estabilidad y pertinencia del modelo. Además, la evaluación de nuevas categorías de puntajes y ajustes en los hiperparámetros podría optimizar aún más el rendimiento.

Considerando la aplicación práctica del modelo, es esencial asegurar que el sistema sea fácil de mantener y actualizar. La implementación de pipelines automatizados y el uso de servicios en la nube pueden facilitar estos procesos, permitiendo una rápida adaptación a cambios en los datos o en las necesidades del negocio. La integración con plataformas de análisis en tiempo real también puede mejorar la precisión y relevancia del modelo en la toma de decisiones.

Finalmente, este proyecto ha proporcionado no solo una herramienta valiosa para la clasificación del puntaje SISBEN, sino también una base sólida para futuros desarrollos y mejoras. La experiencia obtenida en la manipulación y creación de modelos de Machine Learning será esencial para abordar desafíos similares en el futuro, mejorando la calidad de vida de los beneficiarios del SISBEN. La implementación de este sistema promete una asignación de recursos más justa y eficiente, alineada con las necesidades reales de la comunidad.

9. Recomendaciones

1. Continuar con la Validación de Modelos: Es recomendable seguir evaluando y ajustando los modelos de Machine Learning con nuevos datos para asegurar que los modelos sigan siendo precisos y relevantes. La validación continua y el monitoreo del rendimiento del modelo deben ser una práctica habitual para adaptarse a posibles cambios en los datos socioeconómicos.
2. Incorporar más variables: Evaluar la posibilidad de incluir nuevas variables que puedan influir en la clasificación del puntaje SISBEN. Factores adicionales como indicadores de salud, educación o empleo podrían proporcionar una visión más completa y mejorar la precisión del modelo.
3. Explorar Técnicas Avanzadas: Continuar explorando y probando técnicas avanzadas de Machine Learning y algoritmos más sofisticados que puedan ofrecer mejoras adicionales en la precisión y eficiencia de la clasificación, como redes neuronales profundas y métodos de aprendizaje no supervisado.
4. Realizar Auditorías Periódicas: Realizar auditorías periódicas del sistema para identificar y corregir cualquier sesgo potencial en los modelos. Asegurar que los modelos no discriminen contra ningún grupo demográfico y que los beneficios se distribuyan de manera justa y equitativa.
5. Fomentar la Colaboración: Fomentar la colaboración entre departamentos y con instituciones académicas para investigar nuevas metodologías y compartir mejores prácticas en el uso de Machine Learning para la clasificación socioeconómica.

Referencias

[1] Python Foundation. (2021). *Python (Versión 3.9)* [Software]. https://www.python.org/

[2] MEDATA, Estrategia de datos de Medellín "*Base de datos SISBEN 2017*". Recuperado el 19 de octubre de 2023. [En línea]. https://medata.gov.co/dataset/1-002-22-000038

[3] SISBEN. (n.d.). *¿Qué es el sisben?* [En línea]. Recuperado el 1 de noviembre de 2023, de https://www.sisben.gov.co/Paginas/que-es-sisben.aspx

[4] Project Jupyter. (n.d.). *Jupyter Notebook* [Software]. https://jupyter.org/

[5] Amat Rodrigo, J. (2020, Octubre). *Regresión lineal con Python. Ciencia de Datos* [En línea]. Recuperado el 19 de octubre de 2023. https://cienciadedatos.net/documentos/py10-regresion-lineal-python

[6] Hunter, J. D., Dale, D., Firing, E., Droettboom, M., & Matplotlib development team. (2023). *Matplotlib (Versión 3.8.1)* [Software]. https://matplotlib.org/

[7] McKinney, W., & Pandas development team. (2023). *Pandas (Versión 2.1.2)* [Software]. https://pandas.pydata.org/

[8] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Oliphant, T. E. (2023). *NumPy (Versión 1.29)* [Software]. https://numpy.org/

[9] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., & SciPy 1.0 Contributors. (2023). *SciPy (Versión 1.11.3)* [Software]. https://scipy.org/

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2023). *Scikit-learn (Versión 1.3)* [Software]. https://scikit-learn.org/

[11] Reichheld, F. F. (2003). The One Number You Need to Grow. Harvard Business Review. Recuperado de https://hbr.org/

[12] Jensen, K. (n.d.). "Cross Industry Standard Process for Data Mining - CRISP-DM". Wikipedia. Trabajo propio (Jensen,K.), basado en documentación de SPSS Modeler CRISP-DM. Recuperado de https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining