**Logotipo

Descripción generada automáticamente**

**Análisis Integral de clasificación de usuarios -**

**SISBEN - Año 2017**

John Byron Alzate Hernández

[Jorge Luis Genes Padilla](mailto:jlgenesp@unal.edu.co)

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

Asesor  
David Manuel Villanueva Valdés, Magíster en Ingeniería de Software

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2023

|  |  |
| --- | --- |
| **Cita** | (Alzate Hernández & Genes Padilla, 2023) |
| **Referencia**  **Estilo APA 7 (2020)** | Alzate Hernández, J.B., & Genes Padilla, J. L. (2023). *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, CohorteVI.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

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

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Dedicatoria**

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.

**Agradecimientos**

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.

**Tabla de contenido**

[Resumen 9](#_heading=h.3rdcrjn)

[Abstract 10](#_heading=h.26in1rg)

[1.](#_heading=h.35nkun2) Descripción del problema 11

[1.1.](#_heading=h.1ksv4uv) Problema de negocio 11

[1.2.](#_heading=h.44sinio) Aproximación desde la analítica de datos 11

[1.3.](#_heading=h.z337ya) Origen de los datos 11

[1.4.](#_heading=h.3j2qqm3) Métricas de desempeño 11

[2.](#_heading=h.4i7ojhp) Objetivos 12

[2.1.](#_heading=h.2xcytpi) Objetivo general 12

[2.2.](#_heading=h.3whwml4) Objetivos específicos 12

[3.](#_heading=h.2bn6wsx) Datos 13

[3.1.](#_heading=h.qsh70q) Datos originales 13

[3.2.](#_heading=h.3as4poj) Datasets 13

[3.3.](#_heading=h.1pxezwc) Analítica descriptiva 13

[Referencias 22](#_heading=h.1jlao46)

**Lista de tablas**

**Tabla 1** Descripción de las variables.14

**Lista de figuras**

**Figura 1** Distribución de los usuarios de acuerdo a la categoría de Puntaje.19

**Figura 2** Gráfica de distribución de usuarios por estrato. 19

**Figura 3** Distribución de usuarios según la cobertura de necesidades básicas. 20

**Figura 4** Distribución de usuarios según la cobertura de necesidades secundarias.21

**Siglas, acrónimos y abreviaturas**

**SISBEN** Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales

**ML** Machine Learning

**DNP** Departamento Nacional de Planeación

**RMSE** Root Mean Square Error

**MSE** Mean Squared Error

**PCA** Principal Component Analysis

# Resumen

El proyecto aborda el análisis de la Base de datos SISBEN 2017, apuntando a extraer métricas de desarrollo urbano y poblacional para la ciudad de Medellín, con el fin de generar el puntaje SISBEN, dato que será usado como variable de salida para los diferentes modelos que serán empleados, tanto analíticos como estadísticos y descriptivos.

El propósito primordial es generar un proceso de categorización de datos para el dataset, con base en la variable de salida anteriormente descrita y con la finalidad de evaluar si el sistema actual propuesto está realizando correctamente la clasificación del puntaje SISBEN a sus beneficiarios o si, por el contrario, es posible mejorar el procedimiento actual. Para tal fin, se implementarán herramientas estadísticas (Python, Jupyter Notebook, scikit-learn), con la finalidad de evaluar qué modelos de ML son apropiados para el desarrollo y optimización del objetivo del proyecto, permitiendo tener un mejor criterio al momento de la toma de decisiones en clasificación poblacional y de puntaje SISBEN por las partes interesadas.

Incluye al final de dicha página la dirección de los repositorios GitHub.

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

# Abstract

The project addresses the analysis of the SISBEN 2017 database, aiming to extract metrics related to urban and population development for the city of Medellín. The primary goal is to generate the SISBEN score, which will be used as the output variable for various models, both analytical and statistical, as well as descriptive.

The main purpose is to establish a data categorization process for the dataset based on the aforementioned output variable. This is done to evaluate whether the current proposed system correctly classifies the SISBEN score for its beneficiaries or, conversely, if it is possible to improve the current procedure. To achieve this, statistical tools (Python, Jupyter Notebook, scikit-learn) will be implemented to assess which machine learning models are suitable for achieving and optimizing the project's objective. This will provide better insights for decision-making in population classification and SISBEN score determination for the stakeholders involved.

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

# Descripción del problema

El problema del proyecto se centra en la inexactitud en la clasificación urbana, evidenciada por la falta de un análisis más preciso sobre los parámetros de calidad de vida a nivel local en la ciudad de Medellín para determinar el Puntaje SISBEN, variable necesaria para determinar qué ciudadanos, con base en sus condiciones de vida diarias, pueden acceder a subsidios y beneficios sociales.

## Problema de negocio

Actualmente, la medición del Puntaje SISBEN (dato fundamental para la clasificación poblacional con la finalidad del acceso a beneficios sociales) carece de fundamentos sólidos y precisos a causa de la presencia múltiple de valores atípicos dentro de las variables que componen el modelo, alterando así la variable de salida (Puntaje SISBEN). Lo anterior ocasiona que, desde el punto de vista social, los recursos y subsidios sean desaprovechados o usados por un sector de la población que no los necesita en realidad.

## Aproximación desde la analítica de datos

Se pretende desarrollar modelos predictivos con el fin de detectar comportamientos y valores atípicos que evidencien que el modelo actual de clasificación por puntaje SISBEN es impreciso. Lo anterior puede atribuirse a situaciones como información suministrada de forma errada por los usuarios y la falta de procesos de control y verificación en la recolección de la información sensible. Dado lo anterior, se pretende generar un análisis predictivo que permita mejorar la clasificación del puntaje SISBEN con base en las variables estadísticamente más aportantes.

## Origen de los datos.

Se pretende estudiar la información del SISBEN (Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales) del año 2017. El autor de dicha base de datos es la Subdirección de Información y Evaluación Estratégica del Departamento Nacional de Planeación.

## Métricas de desempeño

Se usarán como métricas de desempeño las siguientes:

* **RMSE (Root Mean Square Error):** Es la raíz cuadrada del promedio de los errores cuadrados entre las predicciones y los valores reales. Cuanto menor sea el RMSE, más preciso es el modelo; ofrece una medida consolidada del error del modelo en términos de la variable objetivo.
* **F1-Score:** Es la media armónica entre precisión y recall. Ofrece un balance entre la capacidad del modelo para detectar positivos verdaderos y evitar falsos positivos. Un F1-Score alto indica una precisión y recall equilibrados.
* **Recall (Sensibilidad):** Es la proporción de positivos verdaderos detectados por el modelo en relación con todos los casos positivos reales. Es crucial en situaciones donde no detectar un positivo (falso negativo) tiene consecuencias graves, como en diagnósticos médicos.
* **Error Cuadrático Medio (MSE):** Mide la calidad de un estimador o un modelo de predicción. Es siempre positivo y cuanto más cerca de cero esté, mejor será.
* Las métricas de negocio (desempeño del proceso de negocio que usaría tus modelos) (HABLAR CON DAVID)

Aumento en Seguridad alimentaria

Aumento en Niveles de seguridad

Los puntos anteriores no son directamente medibles por nuestro modelo. Sin embargo, sí son impactados por la solución y la mejora propuesta en el presente proyecto. Debido a que ante una mejor clasificación poblacional para la obtención de beneficios sociales, se obtienen directamente variaciones en las tres medidas anteriores, será menester de una entrega posterior el consultar las fuentes de información específicas que contienen las mediciones exactas de los tres puntos anteriores y sus factores/variables, con el fin de encontrar puntos en común para establecer una relación entre una medida adecuada del puntaje SISBEN y el cambio en las variables ya mencionadas.

* Un juicio sobre qué valor mínimo de las métricas son necesarias en el contexto de aplicación. Por ejemplo: a partir de x% de acierto (métrica de ML) , se estima un ROI de USD y (métrica de negocio) que es lo mínimo para cubrir la operación, despliegue, etc. de los modelos. (HABLAR CON DAVID)

# Objetivos

## Objetivo general

Proponer una solución con el fin de reducir la inexactitud y minimizar errores en la asignación de puntajes SISBEN en los estratos 1 al 4 del Área Metropolitana de la ciudad de Medellín, mejorando así la equidad en la distribución de recursos y beneficios sociales.

## Objetivos específicos

* Identificar las variables que más influyen en la asignación de dicho puntaje. Una métrica útil para tal identificación es el PCA (Análisis de Componentes Principales).
* Desarrollar un modelo de Machine Learning que sea capaz de asignar un puntaje SISBEN más exacto, de acuerdo a los datos recopilados. Métricas y herramientas estadísticas como la Validación Cruzada y Selección de Hiperparámetros son métricas que nos pueden servir para cumplir dicho objetivo.
* De igual forma, dicho modelo debe ser capaz de identificar la fuente de la inexactitud en los puntajes asignados anteriormente. Herramientas estadísticas como matrices de correlación son útiles para dicha identificación.
* Evaluar la precisión de los resultados del modelo desarrollado, con el fin de proponerlo como una alternativa de solución al DNP. Se utilizarán métricas como el Accuracy, útiles para la evaluación de dicha precisión.

# Datos

## Datos originales

La base de datos se encuentra en la página de MEData, estrategia de datos de la ciudad de Medellín, que busca la apropiación, apertura y uso de los datos como herramienta de gobierno, acción ciudadana y toma de decisiones. El dataset anteriormente mencionado, creado por el Departamento Nacional de Planeación, posee información de los puntajes SISBEN asignados a los habitantes de los estratos 1 al 4, de la ciudad de Medellín. La información se encuentra disponible al público, ya sea en páginas de bases de datos como MEData, así como directamente en la página oficial del Departamento Nacional de Planeación; para acceder a ella no es necesario tener usuario y contraseña para ingresar, basta con entrar a la página y estará completamente disponible, lo cual garantiza total accesibilidad.

Para efectos de investigación, se tuvieron en cuenta un total de 52 columnas, así como también un total de 100.000 registros. A continuación se presentarán los datos utilizados, así como también una descripción detallada y el tipo de dato:

**Tabla 1.** Descripción de las variables.

|  |  |  |
| --- | --- | --- |
| **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 |
| PARED | Object | Material predominante de las paredes exteriores. 0 = Sin paredes 1 = Bloque, ladrillo, piedra, madera pulida 2 = Tapia pisada, adobe 3 = Bahareque 4 = Material prefabricado 5 = Madera burda, tabla, tablón 6 = Guadua, caña, esterilla, otro vegetal 7 = Zinc, tela, cartón, latas, desechos, plásticos. |
| PISO | Object | Material predominante de los pisos. 1 = Alfombra o tapete, mármol, parqué, madera pulida y lacada 2 = Baldosa, vinilo, tableta o ladrillo 3 = Cemento o gravilla 4 = Madera burda, madera en mal estado, tabla o tablón 5 = Tierra o arena 6 = Otro. |
| 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. |
| BASURA | Object | La vivienda cuenta con servicio público de basura. 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. |
| ELIMBASURA | Object | ¿Cómo eliminan principalmente la basura en esta unidad de vivienda? 1 = La recogen los servicios de aseo 2 = La entierran 3 = La queman 4 = La tiran al patio, lote, zanja o baldío 5 = La tiran al río, caño, quebrada, laguna 6 = La eliminan de otra forma. |
| THOGAR | Object | Total de hogares de vivienda. Esto equivale a la cantidad de grupos de personas que comparten una misma vivienda. |
| TENEVIV | Object | Este hogar vive en: 1 = Arriendo 2 = Propia pagando 3 = Propia pagada 4 = Otra condición. |
| USOSANI | Object | El servicio sanitario es 0 = No tiene 1 = De uso exclusivo del hogar 2 = Compartido con otros hogares |
| SANITAR | Object | El servicio sanitario que utilizan es: 0 = No tiene 1 = Inodoro con conexión a alcantarillado 2 = Inodoro con conexión a pozo séptico 3 = Inodoro sin conexión a alcantarillado ni a pozo séptico 4 = Letrina, bajamar. |
| DUCHA | Object | Tiene ducha o regadera conectada a acueducto. 1=SI, 2=NO. |
| AGUA | Object | El agua para consumo la obtienen principalmente de: 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 |
| LLEGA | Object | ¿El agua llega al hogar los siete días de la semana? 1=SI, 2=NO. |
| SUMINIS | Object | ¿Los días en que llega el agua, el suministro es continuo las 24 horas? 1=SI, 2=NO. |
| PREPARAN | Object | En dónde preparan los alimentos las personas de este hogar. 0 = En ninguna parte (no tiene cocina) 1 = En un espacio exclusivo para cocinar 2 = En un espacio NO exclusivo para cocinar |
| COCINA | Object | La cocina o sitio para preparar los alimentos es. 1 = De uso exclusivo del hogar 2 = Compartida con otros hogares. |
| COCINAN | Object | Cuál combustible o fuente de energía utilizan principalmente para cocinar. 0 = Ninguno 1 = Electricidad 2 = Gas natural domiciliario 3 = Gas propano (en cilindro o pipeta) 4 = Kerosene, petróleo, gasolina, cocinol, alcohol 5 = Carbón minera. |
| ALUMBRA | Object | Qué tipo de alumbrado utilizan principalmente. 0 = Ninguno 1 = Eléctrico 2 = Solar, Bioenergía, Otros 3 = Kerosene, Petróleo, Gasolina 4 = Vela. |
| USOTELE | Object | El servicio telefónico es: 0 = No tiene 1 = De uso exclusivo del hogar 2 = Compartido con otros hogares |
| 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. |
| TVCABLE | Object | El hogar cuenta con servicio de televisión por cable o parabólica: 1=SI, 2=NO. |
| CALENTA | Object | El hogar cuenta con calentador de agua o ducha eléctrica: 1=SI, 2=NO. |
| HORNO | Object | El hogar cuenta con horno microondas: 1=SI, 2=NO. |
| AIRE | Object | El hogar cuenta con aire acondicionado: 1=SI, 2=NO. |
| COMPUTADOR | Object | El hogar cuenta con computador: 1=SI, 2=NO. |
| MOTO | Object | El hogar cuenta con moto para uso del hogar: 1=SI, 2=NO. |
| AUTO1 | Object | El hogar cuenta con automóvil para uso del hogar: 1=SI, 2=NO. |
| SEXO | Object | Sexo. |
| PARENTES | Object | Parentesco con el jefe del hogar 1 = Jefe 2 = Cónyuge o compañera (o) 3 = Hijos 4 = Nietos 5 = Padres 6 = Hermanos 7 = Yerno o nuera 8 = Abuelos 9 = Suegros 10 = Tíos 11 = Sobrinos 12 = Primos 13 = Cuñados 14 = Otros parientes 15 = No parientes 16 = Servicio doméstico o cuidandero 17 = Pariente servicio doméstico o cuidandero |
| ESTCIVIL | Object | Estado civil. |
| CONYUVIVE | Object | Si el cónyuge vive en este hogar, escriba el número de orden. |
| DISCAPA | Object | Por enfermedad, accidente o nacimiento tiene alguna de las siguientes condiciones de forma permanente. 0=Sin discapacidad, 1=Física, 2=Intelectual, 3=Mental, 4=Psicosocial, 5=Sensorial, 6=Auditiva, 7=Visual |
| CARNET | Object | En salud es afiliado a: 0 = Ninguna 1 = Instituto de Seguros Sociales - ISS 2 = Regímenes especiales (Fuerzas Militares, Policía Nacional, Universidad Nacional, Ecopetrol, Magisterio) 3 = EPS contributiva DISTINTA a 1 ó 2 4 = EPS subsidiada (ARS - Administradora de Régimen Subsidiado). |
| EMBARAZA | Object | ¿Está embarazado o ha tenido hijos? 1=SI, 2=NO. |
| ASISTE | Object | Asiste a centro educativo: 1=SI, 2=NO. |
| TIPOESTA | Object | Tipo de establecimiento 0 = Ninguno 1 = Centros de atención u hogares ICBF 2 = Guardería, sala cuna, preescolar, jardín infantil público 3 = Guardería, sala cuna, preescolar, jardín infantil privado 4 = Escuela, colegio, técnico universitario o universidad pública 5 = Escuela, colegio, técnico universitario o universidad privada 6 = SENA 7 = Secundaria técnica pública |
| GRADO | Object | Último año aprobado: |
| NIVEL | Object | Nivel educativo alcanzado: |
| INGRESOS | Float64 | Total de ingresos mensuales: |
| PERCIBE | Object | Percibe ingresos (laborales, arriendos, subsidios, transferencias, en especie) |
| PUNTAJE | Float64 | Puntaje es un valor entre cero (0) y cien (100). |
| EDAD | Int64 | Edad del usuario. |
| AgeCategory | Object | Categoría de edad a la que pertenece el usuario. |
| Puntaje\_Categorico | Object | Categoría de puntaje a la que pertenece el usuario. |

**Tabla 1.** Descripción de las variables.

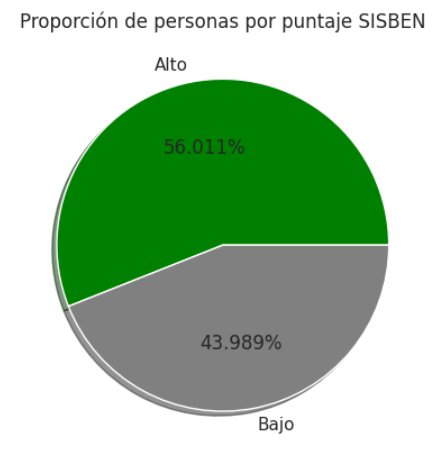
## Datasets

Se realiza en primera instancia los procesos de limpieza, descripción y exploración de los datos. Se pretende utilizar un conjunto de datos compuesto por un 80% de datos de entrenamiento, y un 20% de prueba. Para ello se utilizará la librería scikit-learn de Python.

## Analítica descriptiva

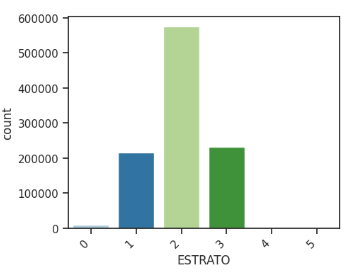
Como se mencionó anteriormente, el dataset se compone de 52 columnas. Cabe resaltar que del dataset original, se tomaron un total de 50 columnas, y fueron creadas dos variables: **AgeCategory,** la cual contiene el rango de edad al que pertenece el usuario, y **Puntaje\_categorico,** la cual contiene el rango de puntaje al que pertenece el usuario. Teniendo esto en cuenta, se toman como variables categóricas, aquellas que son de tipo object, y como variables numéricas las variables de tipo float64 e int64.

Como variable de salida, tenemos la variable **Puntaje\_categorico,** la cual toma valores de Bajo (correspondiente a los usuarios con puntaje entre 0 y 50) y Alto (correspondiente a los usuarios con puntaje entre 51 y 100). La **Figura 1** nos muestra que existe un leve desbalanceo entre ambas categorías, con el 43.989% de los usuarios ubicados en la categoría de puntaje Bajo, y con el 56.011% de los usuarios ubicados en la categoría de puntaje Alto.



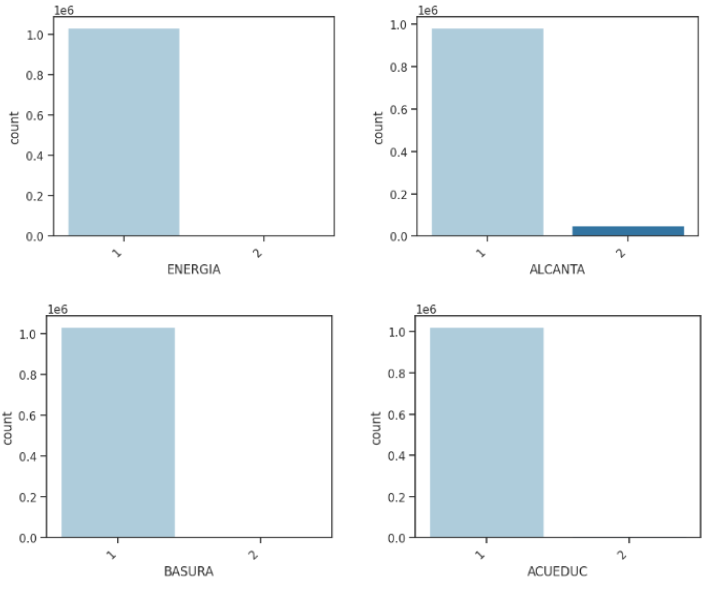
**Figura 1.** Distribución de los usuarios de acuerdo a la categoría de Puntaje.

Para la exploración de datos para las variables categóricas, se utilizaron diagramas de barras para poder visualizar y hacer un análisis más preciso de la distribución de los datos. La **Figura 2** nos muestra que la Zona 1 está compuesta por personas que ocupan los estratos entre 1 y 3, siendo el estrato 2 el que posee más usuarios:

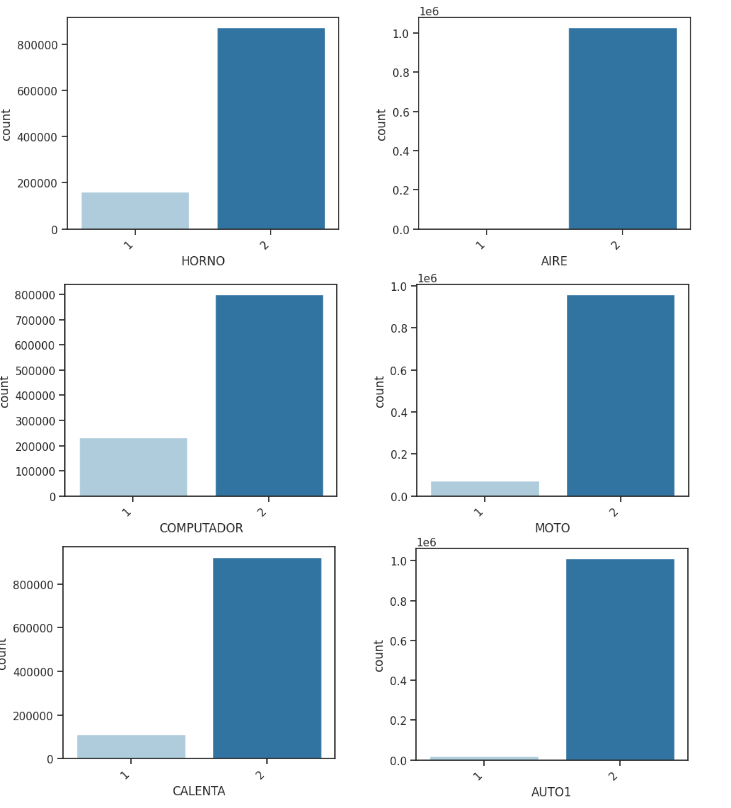


**Figura 2.** Gráfica de distribución de usuarios por estrato.

De acuerdo a la **Figura 3 y 4** podemos evidenciar que los usuarios de la Zona 1, con usuarios en su mayoría de estrato 2, tienen en su mayoría las necesidades básicas cubiertas, las cuales son indispensables para su subsistencia. Por otro lado, las necesidades secundarias, aquellas de las cuales se podría prescindir y que dependen del nivel socioeconómico, no son cubiertas. Esto se puede ver explicado por el tema de que al ser una zona con un alto índice de personas en estrato medio-bajo, no poseen los recursos suficientes para destinar a aspectos como el ocio y artículos de lujo.



**Figura 3.** Distribución de usuarios según la cobertura de necesidades básicas.



**Figura 4.** Distribución de usuarios según la cobertura de necesidades secundarias.

# 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/