Santiago de Cali, 07 de 06 de 2024

Ingeniero:

Juan Carlos Martínez Arias

Director Posgrados de Ingeniería

Facultad de Ingeniería y Ciencias

Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado INDICADOR DE DESARROLLO INTEGRAL EN LOS NIÑOS ENTRE 0 Y 5 AÑOS, EN EL VALLE DEL CAUCA Y 13 MUNICIPIOS DEL NORTE DEL CAUCA, A TRAVÉS DE MODELO PREDICTIVO, el cual será realizado por los estudiantes Wladimir Ávila , Daniel López, Mario Aristizábal con código 8977649,8973969,8975755 perteneciente al énfasis en Sistemas y Computación, bajo la dirección del profesor Mario Julián Mora Cardona.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

Daniel Alejandro Lopez

Daniel Alejandro Lopez
C.C 75.107.733 de Manizales, Caldas

Mario F. Aristizabal

Mario Felipe Aristizábal
C.C 1.053.799.669 de Manizales, Caldas

Wladimir Ávila Huertas
C.C 80.825.479 de Bogotá

Mario Julián Mora Cardona

Director

FICHA RESUMEN

PROYECTO APLICADO – MAESTRÍA EN CIENCIA DE DATOS

TÍTULO: INDICADOR DE DESARROLLO INTEGRAL EN LOS NIÑOS ENTRE 0 Y 5 AÑOS, EN EL VALLE DEL CAUCA Y 13 MUNICIPIOS DEL NORTE DEL CAUCA, A TRAVÉS DE MODELO PREDICTIVO.

1. **ÁREA DE TRABAJO:** Sistema de Información Georreferenciado de Bienestar en la Primera infancia, variables posibles que conducen a un desarrollo integral.
2. **TIPO DE PROYECTO:** Investigación.
3. **ESTUDIANTES:** Wladimir Ávila, Daniel López, Mario Felipe Aristizábal
4. **CORREOS ELECTRÓNICOS:** wladimiravila@gmail.com, mariofelipearisti@gmail.com, danielalejandro234@gmail.com
5. **DIRECCIÓNES Y TELÉFONOS:**
   - Calle 94# 23-64 Apto 604 Bogotá. 3158398864
   - Cra 21 #65c-37 Manizales, Caldas. 3123537264
   - Cra 10 #12-18 Manizales, Caldas. 3163191550
6. **DIRECTOR:** Mario Julián Mora
7. **VINCULACIÓN DEL DIRECTOR:** Profesor Universidad Javeriana Cali
8. **CORREO ELECTRÓNICO DEL DIRECTOR:** mariomora@javerianacali.edu.co
9. **GRUPO O EMPRESA QUE LO AVALA:** Universidad Javeriana de Cali, Riqueza Completa Centro de Investigación
10. **PALABRAS CLAVE:**

    Infancia, niñez, desarrollo integral, bienestar, condiciones, 0-5 años, salud, crecimiento, ciencia de datos, probabilidad, estadística, regresión logística, Modelos predictivos., Sisbén.
11. **FECHA DE INICIO:** 15/05/2023
12. **DURACIÓN ESTIMADA:** 14 meses
13. **RESUMEN:**

El objetivo de este proyecto se enfocó en el desarrollo de modelos predictivos para clasificar y predecir el potencial de desarrollo de niños menores o iguales de 5 años en el Valle del Cauca y norte del Cauca

utilizando los datos del Sisbén. Se analizaron factores socioeconómicos y geográficos que influyen en el desarrollo de la primera infancia, con el objetivo de proporcionar información relevante para la toma de decisiones en políticas públicas y programas sociales dirigidos a esta población vulnerable.

La actualización de la base de datos del Sisbén es fundamental para garantizar la eficacia y equidad en la asignación de subsidios y programas sociales en Colombia. Al mantener la información actualizada, se asegura que los recursos lleguen de manera eficiente a las familias más necesitadas, permitiendo una distribución equitativa y efectiva de los subsidios. Además, una base de datos actualizada proporciona una visión precisa de la situación socioeconómica de la población, facilitando la identificación de necesidades y la planificación de intervenciones enfocadas en el bienestar de los niños y sus familias. Los beneficios de contar con una base de datos del Sisbén actualizada son significativos.

No solo se mejora la focalización de los programas sociales, sino que también se reducen las inequidades y se promueve el desarrollo integral de los niños en situaciones de pobreza y vulnerabilidad. La actualización de la base de datos permite realizar análisis detallados que sirven como base para la formulación de políticas basadas en evidencia, orientadas a mejorar las condiciones de vida de la población infantil en estas regiones específicas de Colombia. En resumen, este proyecto busca aprovechar los datos del Sisbén para desarrollar un modelo predictivo que ayude a comprender y predecir el desarrollo de la primera infancia en el Valle del Cauca y norte del Cauca. La correcta clasificación de categorías en la base de datos del Sisbén garantiza una asignación eficiente de recursos, generando impactos positivos en la calidad de vida y el futuro de los niños en estas regiones.

INDICADOR DE DESARROLLO INTEGRAL EN LOS NIÑOS ENTRE 0 Y 5 AÑOS, EN EL VALLE DEL CAUCA Y 13 MUNICIPIOS DEL NORTE DEL CAUCA, A TRAVÉS DE MODELO PREDICTIVO

Estudiantes:

Wladimir Ávila Huertas

Código: 8977649

Daniel Alejandro López

Código: 8973969

Mario Felipe Aristizábal

Código: 8975755

Proyecto de grado aplicado para optar al título de Magister en Ciencia de Datos

Director:

Mario Julián Mora Cardona

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI

Tabla de contenido

FICHA RESUMEN 2

PROYECTO APLICADO – MAESTRÍA EN CIENCIA DE DATOS 2

1. DEFINICIÓN DEL PROBLEMA 9

1.1. PLANTEAMIENTO DEL PROBLEMA 9

1.2. FORMULACIÓN DEL PROBLEMA 11

2. OBJETIVOS DEL PROYECTO 13

2.1. OBJETIVO GENERAL 13

2.2. OBJETIVOS ESPECÍFICOS 13

3. ALCANCE 15

4. JUSTIFICACION 16

5. MARCO TEÓRICO 19

5.1. Conceptualización y Definición de Modelos Predictivos 19

5.2. Análisis Socioeconómico y Geográfico 22

5.3. Uso de Datos del Sisbén 24

5.4. DECLARACIÓN DE IMPORTANCIA ESTRATÉGICA DEL SISTEMA DE IDENTIFICACIÓN DE POTENCIALES BENEFICIARIOS (SISBÉN IV) Documento CONPES 3877 25

5.5. Definición de un modelo diferencial para 64 zonas geográficas 28

5.6. Transición a Sisbén IV 29

5.7. ¿Para qué sirve el Sisbén y para qué se usa? 30

5.8. ANTECEDENTES 31

6. Metodología de trabajo. 34

6.1. Metodología. 34

6.2. Generalidad del desarrollo del problema 34

6.3. Fase de trabajo. 36

6.4. Conclusiones de las Hipótesis 45

7. CONCLUSIONES 51

8. REFERENCIAS BIBLIOGRÁFICAS 55

Lista de Figuras

ILUSTRACIÓN 1CÁLCULO DEL SISBÉN 26

ILUSTRACIÓN 2 ECUACIÓN DE REGRESIÓN CUANTILICA 27

ILUSTRACIÓN 3METODOLOGIA DE TRABAJO 34

ILUSTRACIÓN 4ALCANCE DEL PROYECTO 35

ILUSTRACIÓN 5ETL CON GLUE AWS 40

ILUSTRACIÓN 6 SCRIPTS GENERADOS EN GLUE AWS 41

ILUSTRACIÓN 7PRESTO EN ATHENA AWS 41

Listas de Tablas

TABLA 1 CLASIFICACIÓN DE PERSONAS POR UBICACIÓN EN SISBÉN 18

TABLA 2 MODELOS DE CLASIFICACIÓN VALLE DEL CAUCA 43

TABLA 3 MODELOS DE CLASIFICACIÓN NORTE DEL CAUCA 44

TABLA 4 MODELOS DE CLASIFICACIÓN VALLE Y NORTE DEL CAUCA 44

INTRODUCCIÓN

El propósito de este proyecto fue abordar un tema de relevancia social y humanitaria; la predicción de las condiciones de vida para niños entre 0 y 5 años utilizando técnicas de ciencia de datos, limitándose al departamento del Valle del Cauca y 13 municipios del norte del Cauca. La primera infancia es una etapa crucial en el desarrollo de un individuo y las condiciones en las que los niños crecen y se desarrollan tienen un impacto significativo en su bienestar futuro.

El proyecto abordó factores vitales que son fundamentales para el desarrollo integral para la primera infancia como lo son:

- Bienestar material
- Salud
- Cuidado, educación y juego
- Bienestar materno
- Seguridad y riesgo

Los factores anteriores al ser exitosos sugieren un desarrollo saludable en la primera infancia y serán sujetos de estudio independientes. Sin embargo, en muchos casos, los recursos limitados y las desigualdades sociales pueden obstaculizar el acceso a estas condiciones fundamentales para los niños más vulnerables.

En este contexto, la ciencia de datos es una herramienta que permite analizar volúmenes sobre los datos y así identificar patrones, tendencias y factores de influencia y utilizar modelos predictivos que puedan comprender y predecir las condiciones de vida de los niños en contextos geográficos y socioeconómicos. Inicialmente se realizó un modelo para cada una de las 5 dimensiones contenidas en el tablero del sistema de información SIG-PI, se formuló un score para cada una de forma independiente, al analizar los resultados obtenidos se evidenciaron sesgos debido a que se utilizaban datos agregados por municipio (siendo este último la unidad de análisis) y estos eran muy pocos para entrenar y probar modelos predictivos, encontrando que en algunos casos la realidad de los niños de la primera infancia no era la más precisa, considerando además que es otra unidad de análisis.

Posteriormente, con el acompañamiento de un experto del Centro de Riqueza Completa se reformuló el modelo descrito anteriormente, utilizando los datos del Sisbén y como etiqueta la categoría de clasificación de los hogares.

El proyecto desarrolló un sistema de predicción basado en técnicas de ciencia de datos y la evaluación de la capacidad de las diferentes técnicas y modelos para predecir la categoría de la clasificación del Sisbén de los niños de la primera infancia, lo que permitió estimar el riesgo en las condiciones de vida a futuro de los niños.

Este proyecto genera conocimiento y herramientas prácticas que puedan ser utilizadas por organizaciones y tomadores de decisiones en la implementación de políticas y programas orientados a mejorar las condiciones de vida de los niños en sus primeros años de vida. Además, se pretende establecer bases sólidas para futuras investigaciones en el campo de la ciencia de datos aplicada a la infancia y el bienestar infantil.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

En las regiones del Valle del Cauca y norte del Cauca en Colombia se encuentra en la intersección de diversos factores que han impactado en su desarrollo económico, social, orden público y político a lo largo de los años. Históricamente, estas regiones han enfrentado desafíos significativos en términos de desigualdad, pobreza, conflictos armados y falta de infraestructura, entre otros aspectos. En el ámbito económico, el Valle del Cauca ha sido tradicionalmente una región importante para la producción agrícola, industrial y comercial en Colombia. Sin embargo, a lo largo de la historia ha experimentado fluctuaciones económicas, impactadas por factores como crisis económicas a nivel nacional e internacional y cambios en los sectores productivos.

En cuanto al norte del Cauca, esta región ha enfrentado desafíos particulares derivados de conflictos armados, presencia de grupos armados ilegales, disputas territoriales y limitaciones en el acceso a servicios básicos. Estos factores han impactado negativamente en el desarrollo social, económico y político de la región, generando desplazamiento de comunidades, violencia y falta de oportunidades para su crecimiento.

Desde el punto de vista social, tanto el Valle del Cauca como el norte del Cauca han enfrentado problemas de pobreza, falta de acceso a servicios de salud, educación, seguridad y exclusión social, estas condiciones han contribuido a la vulnerabilidad de ciertos grupos poblacionales incluyendo a niños menores o iguales de 5 años, quienes se ven afectados en su desarrollo por las condiciones adversas en las que viven.

En términos de orden público, ambas regiones han enfrentado retos relacionados con la presencia de grupos armados, narcotráfico, violencia y delincuencia organizada. La inestabilidad y la inseguridad han impactado en la calidad de vida de la población y han dificultado el acceso a servicios básicos y el desarrollo de programas sociales efectivos.

La importancia de tener la clasificación de los hogares actualizada en la base de datos del Sisbén mejora la eficacia y equidad en la asignación de recursos del Estado, permitiendo así que los programas sociales y subsidios lleguen correctamente a las familias que lo necesitan.

Mantener la base de datos del Sisbén actualizada con una correcta clasificación de los hogares con menos datos para su cálculo, optimiza y garantiza que los recursos públicos sean dirigidos de manera precisa a las poblaciones más vulnerables, evitando desigualdades en la distribución de los beneficios sociales. Optimizando así la gestión de los programas de asistencia social, asegurando que los recursos estatales mejoren las condiciones de vida y el bienestar de quienes están en situaciones de pobreza y vulnerabilidad en el Valle del Cauca y norte del Cauca.

La situación actual del objeto de conocimiento, que en este caso es el bienestar de los niños menores o iguales de 5 años en el Valle y norte del Cauca, presenta síntomas y posibles causas a abordar. A continuación, se describen los principales aspectos de esta situación y se presentan alternativas para superarlas y controlar así el pronóstico planteado.

Síntomas:

- Condiciones de vida precarias: Los niños menores de 5 años en el Valle y norte del Cauca enfrentan condiciones de vida desfavorables que pueden tener un impacto negativo en su desarrollo y bienestar.
- Acceso limitado a servicios básicos: falta de acceso generalizado a servicios esenciales, como atención médica de calidad, nutrición adecuada y otros básicos fundamentales para el crecimiento y desarrollo saludable de los niños.
- Indicadores de salud y nutrición preocupantes: Los indicadores de salud y nutrición de los niños menores o iguales de 5 años en la región reflejan tasas alarmantes de desnutrición, enfermedades prevenibles y otros problemas de salud que pueden tener consecuencias a largo plazo.
- Entorno familiar desfavorable: Muchos niños no están expuestos a un entorno familiar estimulante y seguro, lo que puede afectar negativamente su desarrollo cognitivo, emocional y social.

Posibles Causas:

- Desigualdades socioeconómicas: La región del Valle y norte del Cauca puede presentar disparidades socioeconómicas que afectan la calidad de vida de los niños, especialmente aquellos en situaciones de pobreza o marginalidad.

- Falta de políticas y programas enfocados en la primera infancia: la falta de enfoque y recursos destinados a la atención y el bienestar de los niños menores o iguales de 5 años puede contribuir a la situación actual.
- Limitaciones en la disponibilidad y acceso a servicios: La infraestructura y los recursos necesarios para brindar servicios básicos a todos los niños pueden ser insuficientes, lo que dificulta su acceso y afecta su calidad.

Situaciones futuras si se sostiene dicha situación:

- Mayor vulnerabilidad de los niños: Si no se abordan las condiciones de vida desfavorables, es probable que los niños menores o iguales de 5 años en el Valle y norte del Cauca sigan siendo vulnerables a problemas de salud, malnutrición y retrasos en su desarrollo.
- Impacto en el futuro desarrollo del país: El bienestar de los niños en la primera infancia tiene implicaciones a largo plazo en el desarrollo social, económico y educativo de una nación. Si no se mejora la situación actual, puede haber repercusiones negativas para el futuro del Valle y del norte del Cauca y del país en general.

Alternativas para superar la situación actual:

- Implementación de políticas integrales para la primera infancia: Es necesario establecer políticas y programas específicos que aborden las necesidades de los niños menores o iguales de 5 años, priorizando su bienestar y desarrollo en todas las áreas relevantes.
- Mejora del acceso a servicios básicos: Se deben tomar medidas para garantizar el acceso equitativo y de calidad a servicios esenciales, como atención médica, nutrición adecuada, educación temprana y servicios públicos y de protección infantil.

1.2. FORMULACIÓN DEL PROBLEMA

Desarrollar un modelo predictivo de clasificación basada en datos etiquetados por expertos y mejorados mediante análisis socioeconómico y geográfico de la ubicación del menor, para predecir un score de desarrollo integral de los niños menores o iguales de 5 años en el Valle y norte del Cauca. Esto implica la

identificación de variables relevantes en los datos del Sisbén que influyan en el desarrollo infantil, así como la implementación de técnicas de clasificación/regresión para generar predicciones útiles para la toma de decisiones en políticas públicas y programas sociales dirigidos a esta población vulnerable.

El desafío radica en la limpieza, validación y análisis de los datos del Sisbén, así como en la identificación de patrones y tendencias significativas que puedan aportar insights[1] para mejorar el bienestar y el desarrollo de los niños en estas regiones. La formulación de este problema busca contribuir a la generación de herramientas y conocimientos que permitan una intervención más efectiva y focalizada en el ámbito social, con el fin de mejorar las condiciones de vida y el futuro de los niños en estas áreas específicas de Colombia.

Teniendo en cuenta los análisis previos se desarrolló un modelo predictivo/clasificación que permite obtener la categoría de clasificación del Sisbén, de acuerdo con las variables que influyen dentro de los primeros cinco años de vida. La pregunta es entonces, ¿Cómo actúan las diferentes combinaciones de variables para incidir de manera positiva o negativa en el desarrollo integral de un niño, teniendo en cuenta aspectos familiares, económicos, ambientales o de entorno social, y cuidado?

[1] Insights: aspectos y características relevantes

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

- Desarrollar un modelo predictivo de clasificación/regresión basada en datos etiquetados por expertos y mejorados mediante análisis socioeconómico y geográfico de la ubicación del menor, para predecir un score de desarrollo integral de los niños menores de 5 años en el Valle del Cauca y norte del Cauca.

2.2.OBJETIVOS ESPECÍFICOS

- Recopilar datos relevantes sobre las condiciones de vida de los niños menores de 5 años en el Valle del Cauca y norte del Cauca, incluyendo información demográfica, acceso a servicios básicos, indicadores de salud y nutrición, y datos sobre el entorno familiar, estimulación y desarrollo infantil. variables claves que permiten comprender el contexto estudiado.

- Comprender mediante estadística descriptiva las variables recopiladas, identificando y valorando variables claves que permitan la construcción de un modelo predictivo.

- Construir y validar modelos predictivos utilizando técnicas de aprendizaje automático y análisis estadístico, explorando diferentes enfoques como regresión logística, clasificación o modelos de redes neuronales. Los modelos se entrenan y valoran utilizando datos históricos y técnicas de validación cruzada para así identificar el modelo más acorde para resolver el problema.

- Facilitar la interpretación de los resultados del modelo matemático para apoyar la toma de decisiones en áreas de intervención y construcción de política pública, que permita así definir estrategias para los niños de 0 a 5 años del valle del cauca.

- Seleccionar y definir las variables clave para el modelo predictivo que reflejen con precisión las condiciones de vida de los niños, centrándose en aspectos como acceso a servicios básicos, estado de salud, calidad de la nutrición y estímulos en el entorno familiar.

2.3.RESULTADOS ESPERADOS

- Desarrollo de un modelo: Creación de un modelo predictivo sólido que permita clasificar y predecir el potencial de desarrollo de los niños menores de 5 años en el Valle del Cauca y norte del Cauca, utilizando los datos del Sisbén. Este modelo sería una herramienta valiosa para comprender y anticipar el desarrollo infantil en estas regiones.

- Identificación de factores clave: Se buscaría identificar los factores socioeconómicos y geográficos más relevantes que influyen en el desarrollo infantil en estas áreas específicas de Colombia. Esto proporcionaría información crucial para la toma de decisiones en políticas públicas y programas sociales dirigidos a mejorar las condiciones de vida de los niños y sus familias.

- Mejora en la asignación de recursos: Con un modelo predictivo preciso y basado en datos actualizados del Sisbén, se esperaría una mejora en la asignación de subsidios y ayudas sociales a las familias más necesitadas en el Valle del Cauca y norte del Cauca. Esto contribuiría a una distribución más equitativa y efectiva de los recursos destinados a programas de asistencia social. Información para políticas basadas en evidencia: Los resultados obtenidos a través del análisis de datos y la construcción del modelo predictivo proporcionarían información valiosa para la formulación de políticas públicas basadas en evidencia. Se espera que estos datos respalden la implementación de intervenciones más efectivas y personalizadas en beneficio del desarrollo infantil en estas regiones.

- Impacto en el bienestar infantil: En última instancia, se espera que los resultados de este proyecto tengan un impacto positivo en el bienestar y el desarrollo de los niños menores de 5 años en el Valle del Cauca y norte del Cauca. Al mejorar la focalización de los programas sociales y la asignación de recursos, se busca contribuir a la reducción de las desigualdades y a la promoción de un desarrollo integral de los niños en situación de vulnerabilidad en estas áreas específicas de Colombia.

3. ALCANCE

El proyecto se enfoca en desarrollar modelos predictivos de clasificación para niños de 0 a 5 años en el Valle y norte del Cauca, considerando cambios en las variables que afectan su desarrollo durante los primeros 5 años. Se hará un análisis de variables clave para comprender y predecir su clasificación, por último, su potencial de desarrollo.

Las fuentes de información primarias para este proyecto son los datos del Sisbén, así como otras fuentes secundarias como informes de UNICEF, políticas públicas del gobierno nacional, datos del DANE, observatorios de bienestar infantil y fuentes suministradas por las administraciones locales. El proyecto analiza la relación de las variables identificadas, para proponer modelos predictivos que faciliten la toma de decisiones en políticas públicas y programas sociales, dirigidos a la primera infancia en el Valle del Cauca y norte del Cauca. Se desarrolla un modelo predictivo robusto basado en datos para estimar las condiciones de vida de los niños menores de 5 años en la región, lo que permitirá una asignación más eficaz y equitativa de recursos para las familias más necesitadas.

Es importante tener en cuenta que este proyecto se limita geográficamente al Valle y norte del Cauca, por lo que los resultados y conclusiones estarán enfocados en estas regiones específicas. Además, se reconoce que la implementación de las recomendaciones derivadas del modelo requerirá la colaboración de actores relevantes del gobierno, la sociedad civil y otros sectores, para asegurar que las políticas propuestas puedan tener un impacto efectivo en el desarrollo infantil en estas comunidades.

4. JUSTIFICACION

El modelo del Sisbén, apoyado en la nueva metodología y la encuesta Sisbén IV desarrollada por el Departamento Nacional de Planeación (DNP), está diseñado para ser más eficiente y reducir el margen de error en la categorización de personas. El Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales (Sisbén) es una herramienta fundamental en la focalización del gasto público en Colombia. Su objetivo es identificar y clasificar a los hogares más vulnerables para asignarles los recursos y beneficios de manera eficiente y equitativa.

La eficiencia en la utilización de recursos públicos es crucial para maximizar el impacto de las políticas sociales y asegurar que estos lleguen a quienes realmente los necesitan. La reasignación de recursos de gastos de funcionamiento contribuye a un desarrollo sostenible y a una mejora en la calidad de vida de la población vulnerable.

La encuesta del Sisbén es un componente crucial para recolectar datos socioeconómicos de los hogares. Esta encuesta, tiene una duración promedio de 30 minutos y consta de 92 preguntas, recoge información detallada sobre las condiciones de vida de los hogares. Sin embargo, existen desafíos significativos en la recolección de esta información debido a la cantidad de personas que deben ser encuestadas. [1]

Actualmente, más de 33 millones de personas están registradas en el Sisbén. Es un reto importante para las entidades locales, encargadas de realizar estas encuestas. Adicionalmente, la continua actualización de esta base de datos es esencial para mantener la precisión y relevancia del sistema. [2]

Siendo más eficientes en el tiempo al hacer las encuestas, un mayor número de personas puede tener sus datos actualizados en menos tiempo, lo que permite una focalización más efectiva de los recursos. Esto asegura que los auxilios estén direccionados de una manera más efectiva, mejorando la asignación del gasto social. Teniendo en cuenta que la base de datos del Sisbén se utiliza para identificar y clasificar a la población que puede beneficiarse de diversos programas sociales en Colombia, como, por ejemplo:

- Acceso a Servicios de Salud: Garantizado para todos los clasificados en el Sisbén, incluso aquellos en el grupo D. (Aprox. 7 billones de pesos en el 2021) [3]

- Subsidios de Vivienda: Programas como Mi Casa Ya, que ofrecen subsidios para la adquisición de vivienda propia. (Aprox 10 Billones de pesos en 2024) [4]

- Programas de Alimentación y Nutrición PAE: Incluyendo programas, que brindan apoyo económico a familias y jóvenes en situación de vulnerabilidad. (2 Billones para 2024)[5]

- DPS Departamento de Prosperidad Social: Colombia Mayor, Familias en Acción, Jóvenes en Acción, Red Unidos, Ingreso Solidario, Compensación de IVA, RESA, Mi Negocio, Emprendimientos Colectivos.

- Subsidios de Servicios Públicos: Descuentos en servicios como electricidad y agua para los hogares de bajos ingresos.

- Apoyos Educativos: Becas y subsidios para estudiantes de educación básica, media y superior. (32,000 nuevos créditos primer semestre 2024) [6]

- Agencia Nacional de Tierras: Adjudicación de terrenos baldíos.

- Se vuelve crucial la veracidad da la información contenida en esta base de datos, dado el nivel de recursos que se ejecutan utilizando la base de datos del Sisbén como soporte.

- Así, la actualización periódica fomenta que los datos sean precisos y permitan identificar cambios en las condiciones socioeconómicas de los hogares en el tiempo. Lo que prevé la alteración deliberada en las condiciones socioeconómicas por parte de los encuestados en el momento en que se hacen las encuestas. Al ser un sistema para adjudicar ayudas, hay casos en donde los encuestados deterioran sus condiciones económicas para la entrevista en pro de obtener un puntaje menor. De manera que a estas personas que buscan sacar provecho del sistema, se les dificulte, pues al realizar un mayor número de encuestas en el tiempo se podrá identificar de manera más acertada el nivel socioeconómico real de los hogares encuestados. [7]

Teniendo en cuenta qué, aproximadamente el 23% de las personas viven en áreas diferentes a la cabecera municipal en las regiones estudiadas, esto puede hacer más deficiente el proceso encuestado. A continuación, se detallan los números de las personas que viven en estas zonas.

Tabla 1 Clasificación de personas por ubicación en Sisbén

Clasificación	Nro. Personas	Porcentaje
Cabecera (6,7A,7B,8,11)	2,332,689	**77%**
Centro poblado (5,6,7A,7B,9,10)	354,301	**12%**
Rural disperso (6,7A,7B,9,10)	332,66	**11%**
Total	**3,019,650**	**100%**

Al desarrollar encuestas más eficientes y cortas, respaldadas en modelos que producen resultados precisos mediante un muestreo de encuestas completas, se pueden derivar modelos cada vez más acertados y eficientes. Este enfoque no solo mejora la eficiencia actual del sistema, sino que también establece una base sólida para futuras mejoras. La integración de nuevas tecnologías y metodologías innovadoras continuará siendo fundamental para asegurar la precisión y eficiencia del Sisbén en la identificación y clasificación de beneficiarios.

La importancia de mantener la base de datos lo más actualizada posible dispuso en el documento “El Sisbén IV como instrumento para medir la capacidad de pago y focalizar el gasto en salud” del DNP en donde se estipulo “Sobre este punto varios de los expertos pusieron sobre la mesa la importancia de establecer mecanismos de salida, entre los cuales propusieron incorporar una “fecha de expiración" al puntaje Sisbén de un hogar y, por lo tanto, su estado como elegible o no elegible. Se enfatizó de manera reiterada la necesidad de garantizar mecanismos efectivos de actualización de datos por parte de la población encuestada.” [8]

El Sisbén es una herramienta vital para la focalización del gasto público en Colombia. Su constante actualización y mejora en la metodología buscan garantizar una distribución más eficiente y equitativa de los recursos. Al minimizar el margen de error y asegurar que los beneficios lleguen a quienes realmente los

necesitan, el Sisbén contribuye significativamente a la atención de la población menos favorecida y al desarrollo económico del país. La implementación de tecnologías avanzadas y metodologías innovadoras, junto con un enfoque en la verificación y validación de datos, aseguran que el sistema siga siendo una herramienta eficaz y confiable para la gestión de los programas sociales en Colombia. [9]

5. MARCO TEÓRICO

5.1. Conceptualización y Definición de Modelos Predictivos

Los modelos predictivos son herramientas analíticas esenciales en la ciencia de datos que permiten prever eventos futuros basándose en datos históricos y patrones identificables. La creación de un modelo predictivo implica la aplicación de técnicas estadísticas y de aprendizaje automático para analizar y entender los datos. Los modelos predictivos se dividen en dos categorías principales: modelos de clasificación y modelos de regresión.

Un modelo de clasificación se utiliza para categorizar datos en grupos o clases predefinidas. Por ejemplo, en el contexto de la predicción del desarrollo infantil, un modelo de clasificación podría determinar si un niño está en riesgo de desnutrición o no, basándose en una serie de variables socioeconómicas y geográficas. Estos modelos son especialmente útiles cuando el objetivo es asignar a cada observación a una clase específica.

Por otro lado, los modelos de regresión se utilizan para predecir valores continuos. Un ejemplo de esto sería predecir el peso exacto de un niño a partir de sus características demográficas y socioeconómicas. En este tipo de modelos, la variable de interés es continua y el objetivo es minimizar la diferencia entre los valores predichos y los valores reales observados en los datos de entrenamiento.

La construcción de un modelo predictivo eficaz comienza con la recolección y preparación de datos. En el caso del desarrollo infantil en el Valle y norte del Cauca, los datos del SISBEN proporcionan una base sólida. Estos datos incluyen información detallada sobre las condiciones socioeconómicas de las familias, como el ingreso familiar, el nivel educativo de los padres, las condiciones de vivienda, y el acceso a servicios básicos.

Esta información es crucial para entender los factores que influyen en el desarrollo infantil y para entrenar el modelo predictivo.

El siguiente paso es la selección de características relevantes. No todas las variables disponibles en los datos del SISBEN serán útiles para el modelo. Es necesario realizar un análisis de características para identificar cuáles son las más importantes para la predicción del desarrollo infantil. Este proceso puede involucrar técnicas estadísticas como la correlación, la regresión lineal, colinealidad, clusterizacion, así como métodos más avanzados de aprendizaje automático como los árboles de decisión, regresión logística y las redes neuronales.

Una vez seleccionadas las características, se procede al entrenamiento del modelo. En esta etapa, se divide el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba. El modelo se entrena utilizando el conjunto de entrenamiento y luego se evalúa su rendimiento utilizando el conjunto de prueba. Esta práctica asegura que el modelo no se sobreajuste a los datos de entrenamiento y pueda generalizar bien a nuevos datos.

La validación y evaluación del rendimiento del modelo son etapas cruciales en el desarrollo de modelos predictivos. Se utilizan diversas métricas para evaluar la precisión del modelo, como la precisión, el recall, el F1-score para modelos de clasificación, y el error cuadrático medio (MSE) y el error absoluto medio (MAE) para modelos de regresión. Estas métricas proporcionan una medida cuantitativa del rendimiento del modelo y permiten comparar diferentes modelos y seleccionar el más adecuado para el problema en cuestión.

Para este proceso predictivo se hizo uso de tres tipos de modelos diferentes, el primero de ellos: La regresión logística multiclase, esta es una extensión del modelo de regresión logística que se utiliza para predecir una variable dependiente categórica con más de dos posibles resultados. A diferencia de la regresión logística binaria, que se emplea para clasificar observaciones en una de dos categorías, la regresión logística multiclase maneja situaciones donde las categorías pueden ser tres o más. El modelo de regresión logística multiclase se basa en la generalización del logit, y existen dos enfoques principales para su implementación: el enfoque "uno contra todos" (one-vs-rest) y el enfoque "uno contra uno" (one-vs-one). [10] En el primero, se entrena un modelo de regresión logística binaria para cada clase, donde una clase es el grupo positivo y todas las

demás clases conforman el grupo negativo. En el segundo enfoque, se construye un modelo para cada par de clases. Al igual que en la regresión logística binaria, los parámetros del modelo se estiman utilizando el método de máxima verosimilitud, y las predicciones se realizan seleccionando la clase con la probabilidad más alta entre las calculadas.

La regresión logística multiclase es una técnica estadística utilizada en el aprendizaje automático para predecir una variable dependiente categórica que tiene más de dos posibles resultados. Este modelo es una extensión de la regresión logística binaria, la cual está limitada a predecir variables con solo dos posibles categorías. La regresión logística multiclase aborda esta limitación al permitir la predicción de múltiples categorías. [11]

El segundo de los modelos utilizados fue el modelo Random Forest, un algoritmo de aprendizaje automático basado en árboles de decisión que emplea el método de bagging (bootstrap aggregating) para mejorar la precisión y reducir el sobreajuste. Este algoritmo construye múltiples árboles de decisión durante el entrenamiento y utiliza la media o moda de las predicciones individuales de cada árbol para hacer la predicción final. La aleatorización se introduce en dos niveles: en la selección de subconjuntos de datos (muestras bootstrap) y en la selección de subconjuntos de características para cada división en los árboles, lo cual contribuye a reducir la varianza y mejorar la generalización del modelo. [12]

El método de bagging, o bootstrap aggregating, es una técnica que mejora la estabilidad y la precisión de los modelos de aprendizaje automático. En el caso de los random forests, el bagging se utiliza para crear múltiples árboles de decisión a partir de diferentes subconjuntos de datos, lo que reduce la varianza y evita el sobreajuste. Los random forests introducen una capa adicional de aleatorización al seleccionar subconjuntos aleatorios de características para cada división en los árboles, mejorando así la generalización del modelo y su precisión predictiva.[13]

El tercer modelo utilizado fue XGBoost (Extreme Gradient Boosting), una implementación avanzada del algoritmo de boosting de gradiente, diseñada para ser altamente eficiente, flexible y portátil. Este modelo utiliza un enfoque iterativo en el que cada nuevo modelo intenta corregir los errores de los modelos anteriores, mejorando así la precisión de las predicciones. XGBoost incorpora regularización para prevenir el sobreajuste, optimizando tanto la velocidad de computación como la utilización de memoria mediante técnicas de paralelización y poda de árboles. [14]

Este modelo es particularmente efectivo en conjuntos de datos grandes y complejos, proporcionando un rendimiento superior en comparación con otros algoritmos de aprendizaje automático. XGBoost es una versión paralelizada y optimizada del algoritmo de boosting de gradiente, lo que permite un entrenamiento muy rápido y la capacidad de manejar grandes volúmenes de datos.[15]

Finalmente, es importante considerar la interpretabilidad y la ética en el uso de modelos predictivos. En el contexto del desarrollo infantil, las decisiones basadas en predicciones del modelo pueden tener un impacto significativo en la vida de los niños y sus familias. Por lo tanto, es crucial que el modelo sea transparente y comprensible, y que se utilicen los datos de manera responsable y ética, respetando la privacidad y los derechos de los individuos.

5.2.Análisis Socioeconómico y Geográfico

El análisis socioeconómico y geográfico es fundamental para comprender las disparidades en el desarrollo de la primera infancia y cómo estos factores externos influyen en los resultados de los niños. Diversos estudios han demostrado que las condiciones socioeconómicas y geográficas tienen un impacto significativo en el bienestar y desarrollo de los niños. Factores como el ingreso familiar, la educación de los padres, el acceso a servicios de salud y la calidad de la educación son determinantes clave en el desarrollo infantil.

En el Valle y norte del Cauca, existen disparidades significativas en términos de desarrollo socioeconómico. Las áreas rurales, por ejemplo, suelen tener menos acceso a servicios básicos y oportunidades educativas en comparación con las áreas urbanas. Estas disparidades pueden afectar negativamente el desarrollo infantil, aumentando el riesgo de desnutrición, enfermedades y bajo rendimiento educativo. [16]

El análisis geográfico también es crucial, ya que la ubicación puede influir en factores como la disponibilidad de recursos, la calidad del aire y el acceso a atención médica. La incorporación de datos geográficos en el modelo predictivo permite considerar estas variaciones espaciales y mejorar la precisión de las predicciones. Métodos como la regresión ponderada geográficamente (GWR) pueden ser utilizados para analizar cómo las

relaciones entre las variables socioeconómicas y los resultados del desarrollo infantil varían a lo largo del espacio.

El uso de datos del SISBEN, que proporciona información detallada sobre las condiciones socioeconómicas de las familias, es esencial para este análisis. Estos datos pueden ser utilizados para identificar patrones y tendencias en el desarrollo infantil y para construir un modelo predictivo que tenga en cuenta tanto los factores socioeconómicos como los geográficos.

La integración de análisis socioeconómicos y geográficos en el marco teórico de la tesis permite abordar de manera integral los factores que influyen en el desarrollo infantil. Esto incluye una revisión de la literatura existente sobre cómo las condiciones socioeconómicas afectan el desarrollo infantil, así como un análisis de cómo la ubicación geográfica puede modificar estas relaciones.

Por ejemplo, estudios previos han encontrado que los niños de familias con ingresos más bajos tienden a tener peores resultados de salud y educación. Estas disparidades pueden estar mediadas por una serie de factores, incluyendo la falta de acceso a servicios de salud de calidad, condiciones de vivienda inadecuadas y menos oportunidades educativas. Al considerar también los factores geográficos, se puede obtener una comprensión más completa de cómo estos determinantes se interrelacionan y afectan el desarrollo infantil. [17]

En Colombia, la información contenida en el Sisbén ha sido utilizada para buscar una reducción en las disparidades socioeconómicas y mejorar el bienestar de las familias. Analizar la efectividad de estos programas puede proporcionar información valiosa para el desarrollo de intervenciones más efectivas y focalizadas.

Además, es importante considerar la variabilidad geográfica en la implementación y efectividad de estos programas. Las políticas que son efectivas en áreas urbanas pueden no ser igualmente efectivas en áreas rurales debido a diferencias en infraestructura, acceso a servicios y otros factores contextuales. Por lo tanto, el análisis geográfico es crucial para identificar áreas de alta necesidad y diseñar intervenciones que sean adecuadas para las condiciones locales.

5.3.Uso de Datos del Sisbén

El Sisbén (Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales) es una herramienta clave en Colombia para la identificación y clasificación de las familias según su nivel de pobreza y vulnerabilidad. Los datos del Sisbén son utilizados para la asignación de subsidios y ayudas sociales, y proporcionan una base de datos valiosa para el análisis socioeconómico.

En el contexto de la presente investigación, los datos del Sisbén pueden ser utilizados para desarrollar un modelo predictivo que clasifique y prediga la calificación y el potencial de desarrollo de los niños. Estos datos incluyen información sobre el ingreso familiar, el nivel educativo de los padres, las condiciones de vivienda, el acceso a servicios básicos, entre otros. Esta información es crucial para entender los factores que influyen en el desarrollo de la primera infancia y para construir un modelo que pueda hacer predicciones precisas.

Es importante tener en cuenta la calidad y la integridad de los datos del Sisbén. Antes de utilizar estos datos para entrenar el modelo, se deben realizar procesos de limpieza y validación para asegurar que los datos sean precisos y completos. Esto incluye la eliminación de valores atípicos, el manejo de datos faltantes y la normalización de las variables.

Además, es esencial considerar la ética y la privacidad en el uso de los datos del Sisbén. Los datos personales y sensibles deben ser protegidos y utilizados de manera responsable para asegurar que el modelo predictivo respete los derechos de privacidad de las personas.

Los datos del Sisbén son recolectados a través de encuestas que recopilan información detallada sobre las condiciones de vida de las familias. Esta información incluye variables demográficas, socioeconómicas y geográficas que son cruciales para el análisis del desarrollo de la primera infancia. Por ejemplo, la información sobre el ingreso familiar puede utilizarse para identificar familias en situación de pobreza, mientras que los datos sobre el nivel educativo de los padres informan sobre el capital humano disponible en el hogar.

La utilización de estos datos en el desarrollo de modelos predictivos permite identificar patrones y tendencias en el desarrollo infantil. Por ejemplo, se pueden identificar factores de riesgo que son más prevalentes en ciertas regiones o entre ciertos grupos socioeconómicos. Esta información puede ser utilizada para diseñar intervenciones más efectivas y focalizadas que aborden las necesidades específicas de los niños en diferentes contextos.

5.4. DECLARACIÓN DE IMPORTANCIA ESTRATÉGICA DEL SISTEMA DE IDENTIFICACIÓN DE POTENCIALES BENEFICIARIOS (SISBÉN IV) Documento CONPES 3877

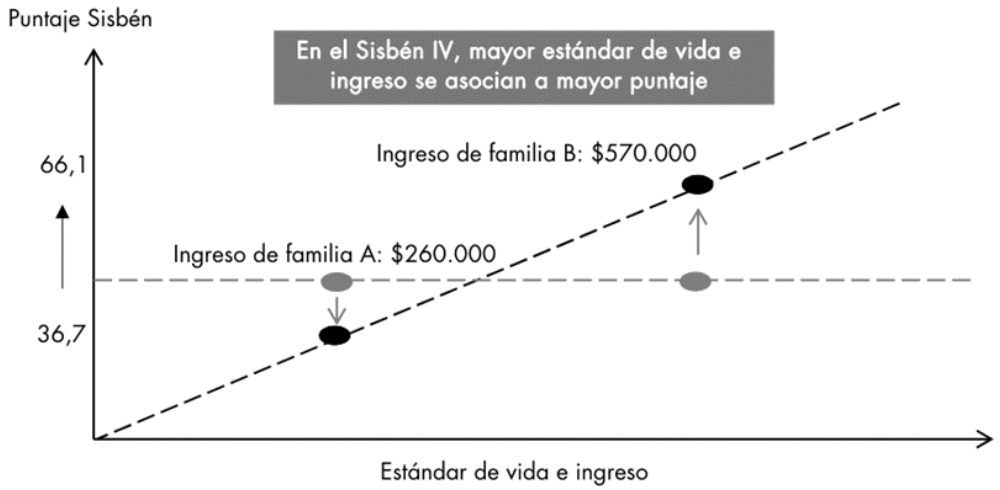
Nuevo enfoque conceptual: inclusión social y productiva

"La movilidad social puede ser percibida como la asociación de dos objetivos deseables: igualdad de oportunidades y meritocracia, los cuales hacen referencia tanto a las características socioeconómicas como al esfuerzo y talento que permiten dichas características. Sin embargo, no puede entenderse la movilidad social como un objetivo alejado de la búsqueda de la igualdad mediante la reducción de la pobreza (Angulo, Gaviria, Azevedo, Páez, 2012). Existe una relación de doble vía entre la situación socioeconómica de los hogares y su capacidad para generar ingreso: el estándar de vida potencia o limita la capacidad de un hogar para generar ingreso, el cual a su vez permite el acceso a una canasta dada de bienes y servicios que determina sus condiciones de vida."

Comprendiendo la complementariedad que existe entre la situación socioeconómica de los hogares y su capacidad para generar ingreso, se propone una metodología para el cálculo del puntaje del Sisbén IV que busque una aproximación integral a la situación real de los hogares desde ambos enfoques. Este nuevo enfoque es el de inclusión social y productiva. La CEPAL define la inclusión como el proceso mediante el cual se logra la igualdad y se superan las disparidades en la productividad, las aptitudes (educación) y el empleo, la segmentación del mercado laboral y la informalidad, que son las causas principales de desigualdad (Bárcena, 2014). Uno de los procesos que son fundamentales en la política pública para promover la inclusión

(tanto social como productiva) es la focalización como mecanismo de identificación de la población para el acceso a los programas sociales que facilitarán dicho proceso de inclusión. En este contexto, el Sisbén, como herramienta para la focalización poblacional, debe incluir un enfoque integral; es decir, en su composición deben incluirse variables que respondan a estos dos tipos de inclusión y generar un puntaje que diferencie hogares con capacidad de generar ingresos y estándar de vida disímiles, como se muestra en la Figura 7.

Ilustración 1Cálculo del Sisbén



Fuente: DNP, 2016.

“Con el fin de actualizar el índice de focalización individual, se inició un proceso de análisis a partir de la pregunta ¿cuál es la principal característica de la población objetivo de los programas sociales? O, en otras palabras, ¿a quiénes debería identificar el instrumento de focalización como potenciales beneficiarios de la oferta social? Los programas buscan beneficiar a la población pobre, definida usualmente como la de baja capacidad de generación de ingresos. Estos programas, a su vez, buscan proveer bienes y servicios básicos a estos hogares. Por tanto, los beneficiarios deben también presentar necesidades en términos de condiciones de vida (salud o educación, por ejemplo). “

“Se evidencia, entonces, la necesidad de incluir en el enfoque del Sisbén III (calidad de vida) el enfoque de ingresos. Así, se podrá contar con un instrumento que permita caracterizar a la población considerando la

complementariedad entre la pobreza monetaria y la multidimensional, y que pueda responder a las necesidades de información de los programas sociales. Nuevo modelo estadístico que incluye capacidad de generar ingresos y nuevas ponderaciones a las variables de calidad de vida Con el fin de aproximar de la mejor manera la relación entre la canasta básica de servicios (calidad de vida) y el nivel de ingresos se emplea la regresión cuantílica [2]como metodología estadística para el cálculo del puntaje. Esta metodología permite tener en cuenta estos dos enfoques en el cálculo del puntaje, aproximando la capacidad de la población para generar ingresos, con el propósito de que la clasificación resultante responda al comportamiento de la pobreza monetaria y multidimensional del país."

Ilustración 2 Ecuación de regresión cuantilica

La ecuación estimada es la siguiente:

$$\ln y_i = X_i \cdot \beta_\tau + \mu_{\tau i},$$
$$Q_\tau\big((W_i|X_i)\big) = X_i \cdot \beta_\tau.$$

"La variable dependiente corresponde al ingreso per cápita del hogar, expresado en escala logarítmica. Las variables independientes Xi corresponden a variables demográficas, mercado laboral, vivienda, tenencia de activos y salud. $\mu\tau$ es un término de error que captura todos los otros factores que podrían afectar, de manera no sistemática, a esta relación. Para la selección de las dimensiones y de las variables se partió de estudios previos sobre determinantes de la presunción de ingresos de los hogares, teniendo en cuenta la información disponible en las fuentes de datos a utilizar. Finalmente, $\beta\tau$ es el vector de parámetros que permite predecir el ingreso de largo plazo a partir de la información de la base de Sisbén. A partir de la estimación de los ingresos de los hogares, se genera un índice con rango de variación entre 0 y 100, con el cual se hace un ordenamiento de menor a mayor."

[2] Regresión cuantílica: Modelo econométrico de tipo de análisis de regresión, basado en la estimación de minimización de desviaciones absolutas ponderadas con pesos asimétricos.

“Como fuente de información para la estimación se usa la Gran Encuesta Integrada de Hogares (GEIH) de 2015 que levanta de manera continua el DANE. Esto, por cuanto esta encuesta: (i) permite obtener estimaciones para un mayor número de desagregaciones geográficas y, en consecuencia, llegar a nivel de departamento por zona, previa agregación de muestras anuales; (ii) es la fuente oficial para las mediciones de los indicadores de mercado laboral, ingresos y pobreza y (iii) es una encuesta continua que permite hacer un seguimiento a las condiciones de vida de los hogares y actualizar las estimaciones de los modelos sin tener que esperar un largo periodo de tiempo como sucede con los censos de población y vivienda, fuente utilizada para el modelo del Sisbén III. El uso de la GEIH de 2015 permite contar con valores actuales de la estimación de los parámetros del modelo. De igual forma, será posible actualizarlos con nuevas versiones de la misma encuesta cuando se considere pertinente para evitar su desactualización. “

5.5. Definición de un modelo diferencial para 64 zonas geográficas

“Para considerar las diferencias regionales en las condiciones de vida del país, el nuevo índice de focalización se estimará por zona geográfica (cabecera y rural) al interior de cada departamento, por lo que se obtienen estimaciones para 64 dominios geográficos. Adicionalmente, se incluirá un modelo propio par Bogotá.”

“Para el uso de este nuevo indicador de focalización individual se ordenará la población en cada dominio según su puntaje de menor a mayor, y se construirán cien grupos de igual tamaño (igual número de personas). De esta forma, en el primer grupo estará la población con menor capacidad de generación de recursos, y en el grupo cien, aquellos con mayor capacidad. La definición de los puntajes de corte para el acceso a los diferentes programas sociales se basará en la definición del número de grupos en cada dominio, desde el más bajo al más alto, que se decida incorporar a cada programa según los objetivos que este persiga. Teniendo en cuenta que las estrategias metodológicas de estimación o aproximación de recursos económicos de los hogares, como cualquier otra metodología utilizada para construir índices de focalización, pueden cometer errores de inclusión y exclusión cuando los programas sociales eligen a sus beneficiarios, deberán usar otros indicadores para optimizar el proceso de focalización individual. Estos pueden ser aquellos recolectados en la ficha de caracterización socioeconómica (por ejemplo, indicadores de educación, salud, trabajo, entre otros), así como información de otras bases de datos (por ejemplo, de los sistemas de información de las

diferentes entidades del Gobierno nacional y otros registros administrativos). Cada programa deberá definir los indicadores y fuentes adicionales que desee considerar en la selección de sus beneficiarios."

5.6. Transición a Sisbén IV

"Para evitar al rezago de información socioeconómica de la población, se prevé poner a disposición de las entidades nacionales y territoriales las bases de datos con la información municipal recolectada bajo la nueva metodología durante el barrido del Sisbén IV. Estas bases servirán como un insumo para la planeación de los programas sociales y de las entidades territoriales. Esta información se dispondrá de manera agregada, para realizar análisis estadísticos que den cuenta a las entidades nacionales y territoriales de la situación actual de la población, a medida que termine el operativo en los distintos departamentos entre diciembre de 2017 y diciembre de 2019. La nueva metodología propuesta requiere de algunas medidas que deben ser tomadas previas a su implementación. La transición al Sisbén IV debe surtir un proceso previo de actualización de la información ya registrada en la base de Sisbén. Para esto se modificará la información con aquella que se recoja en el proceso de levantamiento de la base de Sisbén IV, lo que permitirá entregar al país también los puntajes de Sisbén III actualizados. Adicionalmente, se espera que la actualización del instrumento y de la información de los hogares reflejen tanto los efectos de la política social que ha contribuido a mejorar las condiciones de vida de los colombianos, como los efectos de la desaceleración económica que ha vivido el país en los últimos meses. Lo anterior resultará en cambios en los puntajes de los hogares que hacen parte de la base del Sisbén actualmente. Para promover la actualización de la información por parte del ciudadano y facilitar la transición, las entidades que tengan a su cargo programas sociales, con el acompañamiento técnico del DNP, deberán revisar y hacer explícitos los criterios de entrada, permanencia y salida de los programas. En particular, se deberá desarrollar un contrato social entre el beneficiario y la entidad, en donde se establezca explícitamente el periodo durante el cual se otorgará el apoyo, sin perjuicio de las mejoras en las condiciones de vida y en el puntaje de Sisbén a que haya lugar. Como referente regional, se encuentra el caso de Chile, que determina para cada uno sus programas sociales un tiempo mínimo de permanencia. Por ejemplo, el Bono al Trabajo de la Mujer que entrega un aporte monetario durante cuatro años, así como el Apoyo al Empleo (Sistema Chile Solidario) que promueve la formación y capacitación para el trabajo y tiene una permanencia de 4 meses. Durante el periodo de recolección de información, las entidades líderes de los programas que utilizan el Sisbén como mecanismo de focalización deberán definir, con el acompañamiento

técnico del DNP, el esquema de transición del Sisbén III al Sisbén IV. Este esquema deberá incluir: (i) un análisis del impacto del cambio del índice y del puntaje en la población potencialmente beneficiaria de los programas; (ii) la definición de las poblaciones que serían objeto de transición, y de los criterios de ingreso, permanencia y egreso con los cuales se define la transición; (iii) la definición de mecanismos y de plazos para la adopción del nuevo Sisbén; estos últimos no deberán superar el periodo de recolección de la información; y (iv) la definición de los puntos de corte de acceso a los programas. El esquema de transición definido será descrito en un documento técnico que cada programa pondrá a disposición de sus beneficiarios, una vez la nueva versión sea adoptada. En cuanto a la relación entre el Sisbén y el régimen subsidiado de salud, y en el marco del trabajo de transición al Sisbén IV, el Gobierno en cabeza del DNP, el Ministerio de Salud y Protección Social y el Ministerio de Hacienda y Crédito Público, deberá analizar la pertinencia de mantener el uso del instrumento de focalización en este programa en particular, los impactos técnicos para operación del programa, los impactos fiscales (posibles riesgos en diminución de recaudos) y los impactos operativos (analizar nuevos mecanismos de presunción de ingresos). Este análisis deberá estar terminado en diciembre de 2018.” [18]

5.7. ¿Para qué sirve el Sisbén y para qué se usa?

- Facilitar la clasificación de los posibles beneficiarios de programas sociales de manera rápida, objetiva, uniforme y equitativa.
- Permitir la elaboración de diagnósticos socioeconómicos precisos de la población para apoyar la planeación nacional y territorial, y el diseño e implementación de programas sociales.
- Contribuir al fortalecimiento institucional, mediante la disposición de un sistema moderno de información social que permita mejorar la identificación de la población.

“Con la nueva metodología del Sisbén se desarrolló una clasificación cuyo resultado no es un índice cuantitativo. Esto significa que ya no existirá un puntaje de 0 a 100 sino una nueva clasificación que ordena la población por grupos.”

En el Sisbén IV existen cuatro grupos:

- Grupo A: pobreza extrema (población con menor capacidad de generación de ingresos)
- Grupo B: pobreza moderada (población con mayor capacidad de generar ingresos que los del grupo A)
- Grupo C: vulnerable (población en riesgo de caer en pobreza)
- Grupo D: población no pobre, no vulnerable.

Cada grupo está compuesto por subgrupos, identificados por una letra y un número que permiten clasificar más detalladamente a las personas:

- Grupo A: conformado por 5 subgrupos (desde A1 hasta A5)
- Grupo B: conformado por 7 subgrupos (desde B1 hasta B7)
- Grupo C: conformado por 18 subgrupos (desde C1 hasta C18)
- Grupo D: conformado por 21 subgrupos (desde D1 hasta D21)

Ejemplo: una persona en el nivel A1 tendrá una menor capacidad de generar ingresos y menor calidad de vida que la del A5. [19]

Es decir, A1 tiene menor puntuación, mas riesgo, o peores condiciones que A5. Así mismo para las cuatro categorías.

5.8. ANTECEDENTES

“Se han realizado varios trabajos previos relacionados con la evaluación de las condiciones de vida de los niños en diferentes regiones. Un ejemplo de ello es el estudio realizado por el Instituto Colombiano de Bienestar Familiar (ICBF) en colaboración con el Ministerio de Salud y Protección Social, titulado "Encuesta Nacional de la Situación Nutricional en Colombia – ENSIN” Un ejemplo de ello es el estudio realizado por el Instituto Colombiano de Bienestar Familiar (ICBF) en colaboración con el Ministerio de Salud y Protección Social, titulado "Encuesta Nacional de la Situación Nutricional en Colombia ENSIN [20]

Este estudio aborda la situación nutricional de los niños en Colombia y proporciona información valiosa sobre el estado de salud y nutrición de esta población.

A nivel regional el Centro de Investigación de Riqueza Completa de la Pontificia Universidad Javeriana Cali ha desarrollado análisis descriptivo y diagnóstico con variables más allá de la nutrición, como el acceso a servicios básicos, el entorno familiar y el desarrollo infantil para la población. Enmarcado en el alcance de este proyecto estos nos sirven como referencia para el desarrollo del modelo a proponer.

El uso de modelos predictivos en contextos socioeconómicos y geográficos ha demostrado ser una herramienta valiosa para la toma de decisiones informadas en diversas áreas, desde la educación hasta la salud pública. Estos modelos utilizan técnicas de minería de datos para analizar grandes volúmenes de información y extraer patrones que pueden ser utilizados para predecir futuros comportamientos o eventos.

Un ejemplo destacado de la aplicación de modelos predictivos es el desarrollo de un modelo para predecir la deserción escolar en la educación superior. Este modelo, presentado en un estudio reciente, utilizó la metodología CRISP-DM para estructurar el proceso de minería de datos. CRISP-DM[3] es una metodología estándar que incluye las fases de comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Con datos socioeconómicos y académicos de los estudiantes, el modelo identificó con precisión quienes estaban en mayor riesgo de abandonar sus estudios, permitiendo a las instituciones intervenir oportuna y efectivamente. [22]

Otro ejemplo significativo es el uso de modelos predictivos en salud pública para la predicción de brotes de enfermedades. Utilizando datos geográficos y socioeconómicos, estos modelos pueden identificar áreas de alto riesgo y permitir una asignación más eficiente de recursos médicos. Un estudio aplicado en un contexto de salud poblacional utilizó técnicas de minería de datos para analizar patrones de incidencia de enfermedades en diferentes regiones, considerando factores como la densidad de población, la calidad del agua y las condiciones de saneamiento. Los resultados permitieron a los responsables de salud pública anticipar brotes y planificar intervenciones preventivas de manera más eficaz. [22]

[3] CRISP-DM Cross-Industry Standard Process for Data Mining

La implementación de modelos predictivos en estos contextos generalmente sigue un proceso estructurado que incluye la recolección de datos, la limpieza y preprocesamiento de los datos, la selección de variables relevantes, y la construcción y validación del modelo. Una vez implementados, estos modelos ofrecen varios beneficios, incluyendo la capacidad de identificar patrones ocultos, la mejora en la precisión de las predicciones y la posibilidad de tomar decisiones proactivas basadas en datos.

A pesar de sus beneficios, los modelos predictivos también presentan desafíos y limitaciones. La calidad y disponibilidad de los datos son factores críticos que pueden afectar la precisión del modelo. Además, los modelos pueden ser sensibles a la selección de variables y a la metodología utilizada, lo que requiere un enfoque cuidadoso y bien fundamentado. Es importante también considerar aspectos éticos y de privacidad, especialmente cuando se trabaja con datos sensibles como los de salud o educación.

6. Metodología de trabajo.

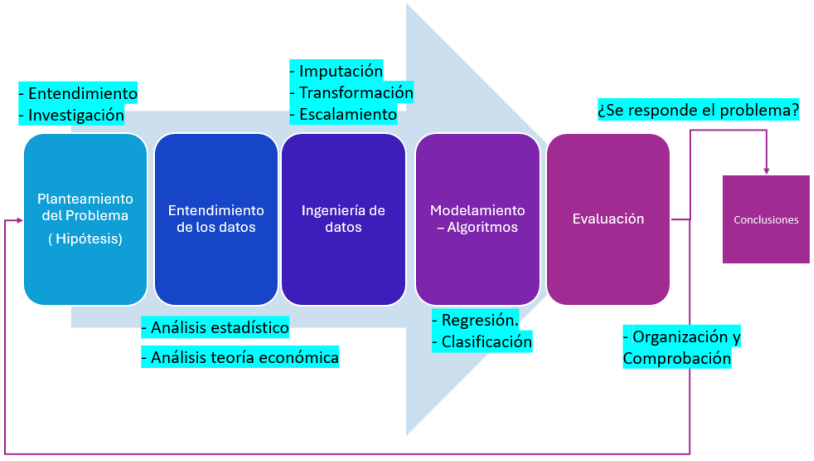
6.1. Metodología.

El presente proyecto se ejecutó utilizando la metodología “CRISP-DM”, en este se tuvieron en cuenta fases como: importación, ordenación, transformación, visualización, modelamiento y por último comunicación de los resultados obtenidos.

"CRISP-DM" es una metodología de minería de datos ampliamente utilizada. La sigla se traduce al español como "Metodología para el Modelado de Procesos en Minería de Datos". Esta metodología es elegida, puesto que utiliza un proceso estándar que permite tener una planificación y ejecución de proyecto eficiente, desde la comprensión de los datos con estadística descriptiva hasta el modelamiento con respecto predictivas, y generación de conclusiones.

A continuación, se ilustra la metodología que se utilizó en el proyecto, donde se evidencia un proceso cíclico al final de la actividad de conclusiones, debido a que cuando se buscaba aceptación de los resultados del proyecto, se evidenciaban puntos que se debería tener en cuenta y que se debía ajustar desde el planteamiento de las hipótesis.

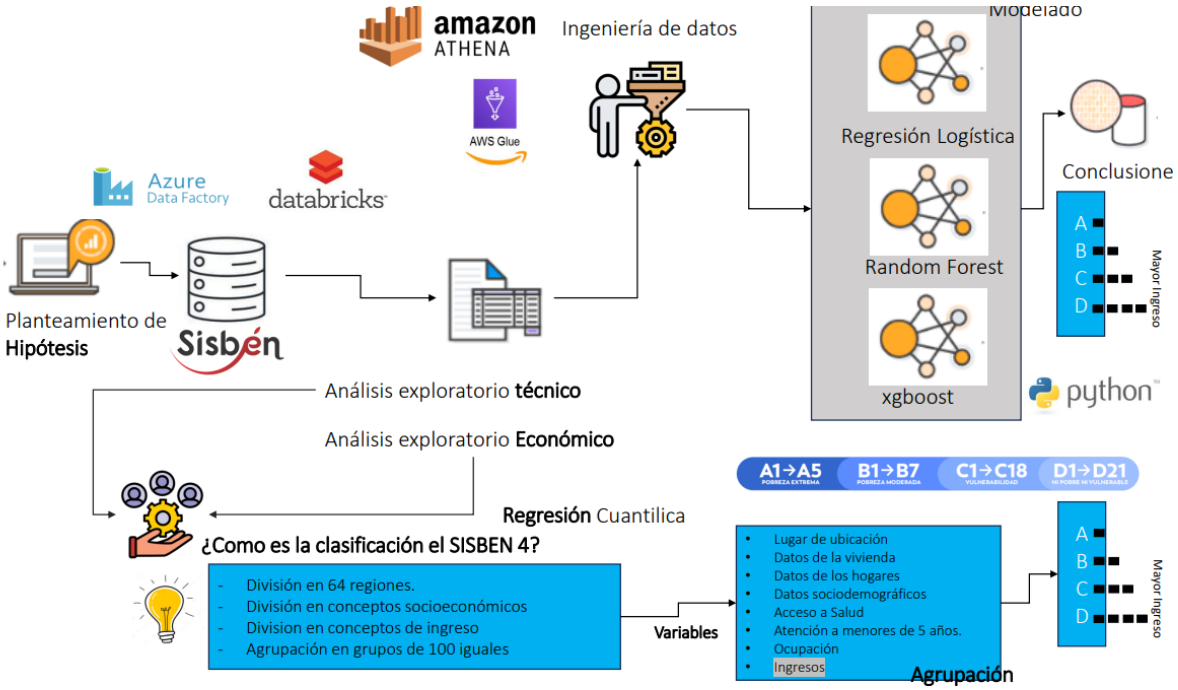
Ilustración 3Metodologia de trabajo



6.2. Generalidad del desarrollo del problema

A continuación, se detalla el paso a paso para cumplir el alcance del proyecto:

Ilustración 4Alcance del proyecto



- Paso 01: Explicación de los campos que se tienen en la base de datos del Sisbén. En esta reunión se entregó al equipo de trabajo el diccionario de datos, con una explicación general e importancia de cada columna.

- Paso 02: Explicación de la relación de los campos de la base de datos con el posible problema a resolver, y el diferencial que se podría tener con respecto a la clasificación que realiza el Sisbén.

- Paso 03: Prueba de concepto de posibles conclusiones que se podrían dar con la aplicación de la ciencia de datos, en esta reunión se plantearon posibles respuestas a las hipótesis, acompañada de una prueba de concepto para entender que posibles conclusiones y diferenciales se podría brindar con la investigación.

- Paso 04: Análisis descriptivos de las variables: utilizando técnicas de análisis descriptivo, explicando las variables desde el contexto económico y estadísticos.

- Paso 05: Ingeniería de datos: Limpieza, normalización y transformación de los datos, utilizando herramientas en la Nube

- Paso 06: Modelado: Desarrollo de los modelos, con el objetivo de probar las hipótesis y dar respuesta al planteamiento del problema.

- Paso 07: Pruebas y Conclusiones: validación de los módulos y generación de conclusiones.

6.3. Fase de trabajo.

La primera fase de la metodología se inició con el planteamiento de 4 hipótesis, estas fueron diseñadas desde el entendimiento inicial del problema en reuniones con el director del presente proyecto.

Estas hipótesis no solo buscan la predicción, también el entendimiento del contexto económico y social de las variables de la base de datos del Sisbén para tratar de responderlas desde la ciencia de los datos.

Las hipótesis son las siguientes:

- Se puede predecir la categoría de un niño menor o igual a 5 años basado en Gastos, con una precisión superior o igual del 80 % de como lo hace el Sisbén actualmente. Tanto del Valle y norte del Cauca.

  Modelo de comprobación de la hipótesis:
  Desarrollo de un modelo de Regresión lineal simple entre las variables gastos y la clasificación, diferenciando tres modelos: el primero solo para el Valle del Cauca, el segundo para el norte del Cauca y el ultimo para los departamentos de forma consolidada.

- Los niños menores o iguales de 5 años que habitan viviendas sin servicios públicos y con dificultades para obtener el agua y condiciones sanitarias son inadecuadas, tienden a estar en categorías más marginales.

  Modelo de comprobación de la hipótesis:

  Es una hipótesis que busca analizar desde la descripción de los datos y así verificar que representan. Hacer una consulta que permita contar los niños con esas características de vivienda y verificar si tienden a estar en las categorías A y B de clasificación, con una precisión igual o superior del 80 % de como lo hace el Sisbén. Diferenciando tres consultas: Primero solo para Valle del Cauca, segundo para el Norte del Cauca y por último para los departamentos de forma consolidada.

- Analizar la correlación de las variables de la población de niños menor o igual de 5 años (Primera infancia), y determinar si la variable número de personas que trabajan en el hogar probablemente influye la clasificación del infante.

  Modelo de comprobación de la hipótesis:

  Generar la correlación de las variables de la primera infancia y describir la importancia económica de las variables.

- Se puede predecir con una precisión superior o igual del 80 % las categorías de los niños menores a o iguales 5 años, si tomamos un modelo de clasificación y no un modelo de regresión como lo hace el Sisbén para categorizar los niños en A, B, C, D.

  Modelo de comprobación de la hipótesis:

  Desarrollar 3 modelos de clasificación. Estos son: el random forest, regresión logística multiclase, y xgboost, diferenciando por cada uno de los tres modelos lo siguiente: el primer análisis solo para el valle del cauca, el segundo análisis solo para el norte del Cauca y el ultimo para los departamentos de forma consolidada.

La Fase de entendimiento de los datos inicialmente se realizó un análisis con los datos obtenidos del sistema "SIG-PI" en las 5 dimensiones donde se encontró los siguientes puntos:

- Mayor del 50% de los datos para el año 2022 eran faltantes para variables claves de las dimensiones.
- Se estableció que se necesitan más datos para adicionar a los modelos matemáticos, ante esto, se definieron actividades para la obtención de datos registrados por expertos, estos datos, servirán para implementar modelos de predicción por cada dimensión.
- Datos agrupados por indicadores, sin tener un desagregado de variables importantes en los modelos de predicción.
- Para entrenar los modelos se necesitó registras y calificar variables por parte de equipos expertos, se evidenció que se necesitan mas datos pata entrenar modelos.
- Se evidencio la necesidad de clasificar los datos por tipo de ubicación geográfica. (rural, rural disperso y urbano)

Luego, en reunión con el experto del Centro de la Riqueza Completa de la Universidad Javeriana se concluyó que no era viable realiza un mayor análisis debido a que las características de los datos, se sugirió utilizar los datos de la base de datos de Sisbén utilizando la variable categoría de clasificación como etiquetado de los datos.

En una siguiente actividad se exploró con análisis descriptivo los datos del Sisbén pertenecientes a la zona de estudio. La exploración se realizó utilizando conceptos estadísticos y económicos como los siguientes: Indicadores estadísticos, gráficos univariables y multivariables, por último, tablas de frecuencia.

Este análisis se desarrolló en igual proporción desde la teoría económica, para así generar conclusiones importantes según lo encontrado en la exploración de los datos.

Las actividades realizadas en esta fase fueron los siguientes:

- Anexo 01: Diccionario de datos del Sisbén
- Anexo 02: Análisis Descriptivo
- Anexo 03: Herramientas para análisis descriptivo.

Fase de ingeniería de datos. Técnicamente la Ingeniería de datos se construyó utilizando herramientas modernas de la nube de AWS, apuntando a analizar, transformar y escalar grandes cantidades de datos.

Luego de varios análisis (Anexo 04: Colinealidad) y reprocesos se determinaron un listado de columnas, en los cuales se descartan indicadores (ind_*), nombres (nom_*), entre otros a fin de disminuir la dimensionalidad de variables que potencialmente no aportan al modelo a realizar.

Las columnas seleccionadas son:

- **data.ide_ficha_origen_hogar**: ID único del hogar, concatenando ide_ficha_origen y ide_hogar.
- **data.cod_dpto**: Código del departamento.
- **data.cod_mpio**: Código del municipio.
- **data.cod_clase**: Clase de Ubicación.
- **data.cod_barrio**: Código del barrio.
- **data.cod_vereda**: Código de la vereda.
- **data.cod_corregimiento**: Código del corregimiento.
- **data.cod_comuna**: Código de la comuna.
- **data.tot_hogares**: Total de hogares.
- **data.tip_vivienda**: Tipo de vivienda.
- **data.tip_mat_paredes**: Material de las paredes.
- **data.tip_mat_pisos**: Material de los pisos.
- **data.ind_tiene_energia**: Indicador de disponibilidad de energía eléctrica.
- **data.ind_tiene_gas**: Indicador de disponibilidad de gas.
- **data.num_cuartos_exclusivos**: Número de cuartos exclusivos.
- **data.num_cuartos_unicos_dormir**: Número de cuartos únicos para dormir.
- **data.tip_sanitario**: Tipo de sanitario.
- **data.tip_uso_sanitario**: Uso del sanitario.
- **data.tip_origen_agua**: Origen del agua.
- **data.tip_uso_agua_beber**: Uso del agua para beber.
- **data.ind_tiene_cocina**: Indicador de disponibilidad de cocina.
- **data.tip_prepara_alimentos**: Tipo de preparación de alimentos.

- **data.tip_uso_cocina**: Uso de la cocina.
- **data.tip_energia_cocina**: Tipo de energía usada en la cocina.
- **data.ind_tiene_internet**: Indicador de disponibilidad de internet.
- **data.num_habita_vivienda**: Número de habitantes en la vivienda.
- **data.ind_evento_inundacion**: Indicador de eventos de inundación.
- **data.ind_evento_avalancha**: Indicador de eventos de avalancha.
- **data.num_personas_hogar**: Número de personas en el hogar.
- **data.clasificacion**: Clasificación del hogar (A, B, C, D).

- Anexo 05: Herramientas para ingeniería de datos.

Utilizando las herramientas previamente descritas, se realizó el proceso de ingeniería de datos en la nube de AWS, cargando los datos en un Bucket S3, posteriormente con el ETL Glue que usa una adaptación de Spark se definió el esquema y se realizó el proceso de transformación de los datos.

Ilustración 5ETL con GLUE AWS

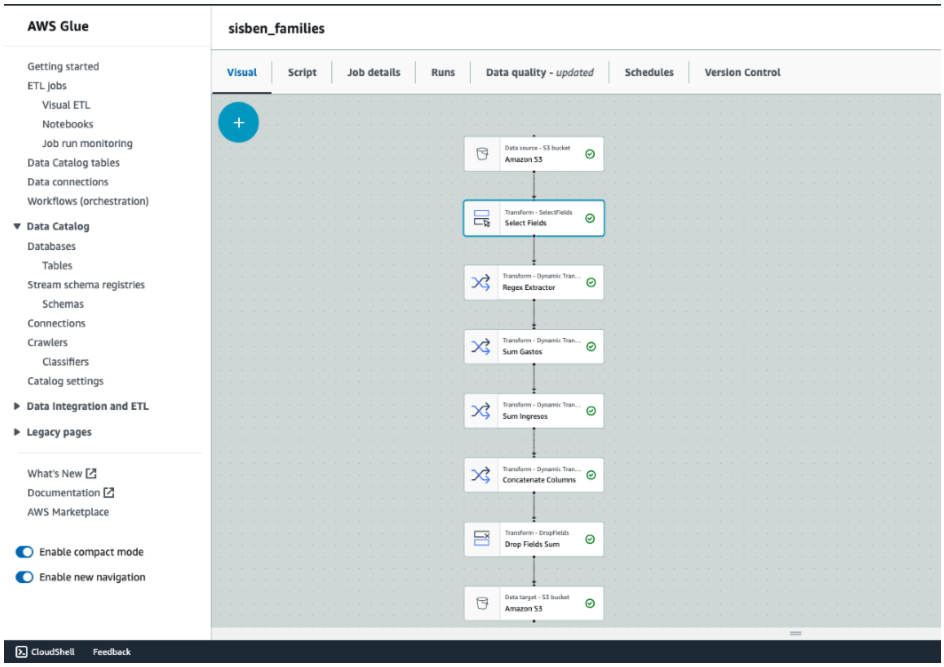
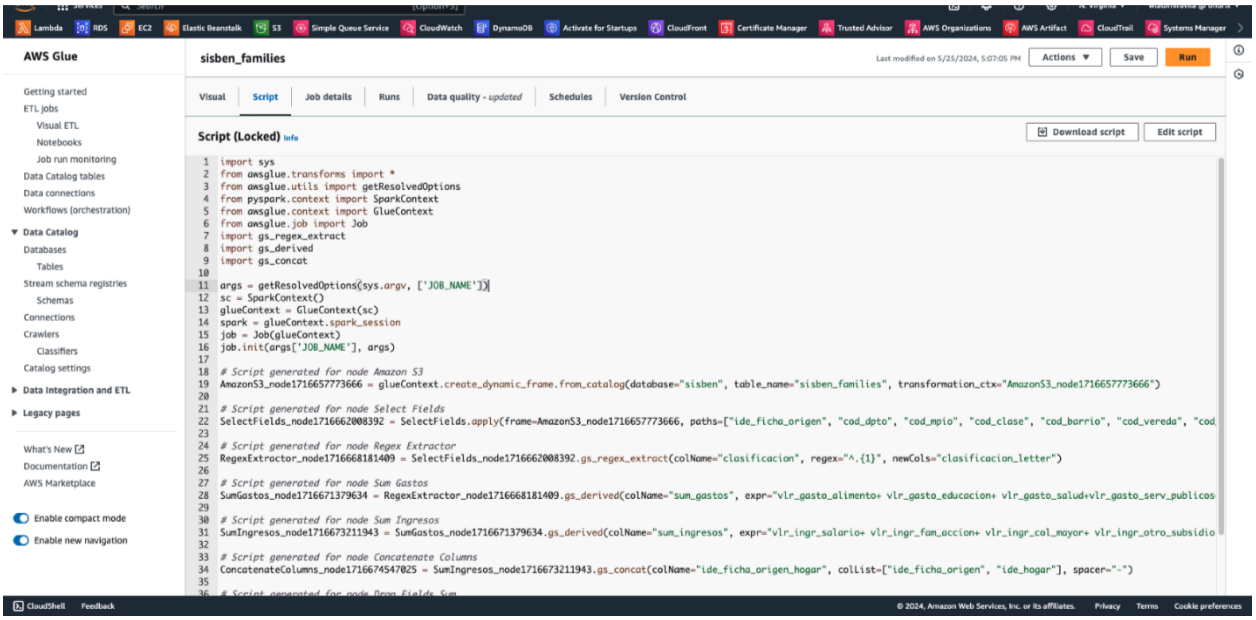


Ilustración 6 Scripts generados en Glue Aws

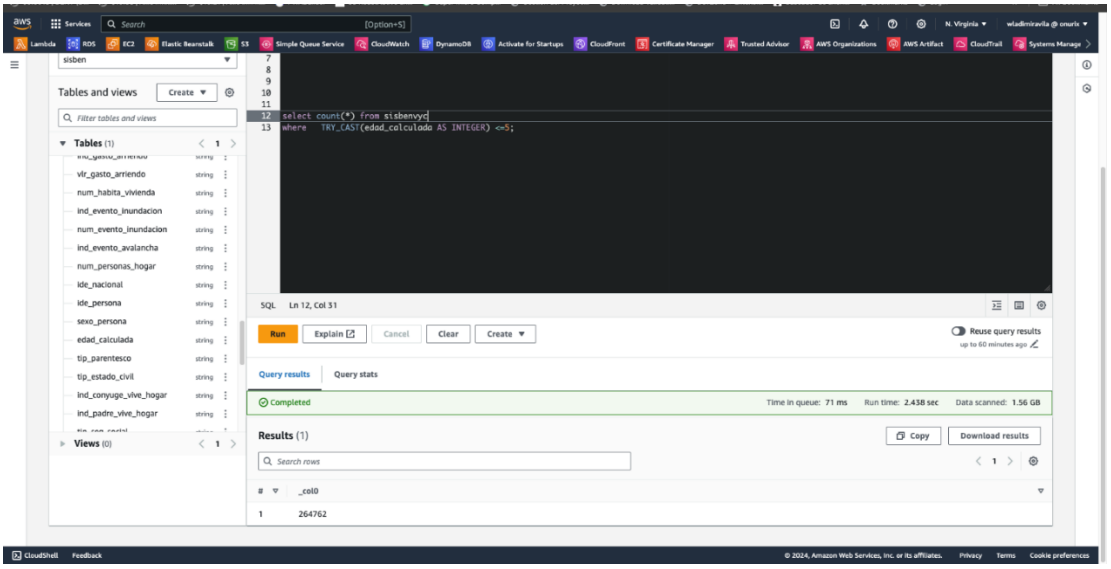


Como resultante de Glue se obtuvo un primer dataset que luego con ayuda de Athena que implementa de PrestoDB se obtuvo un segundo dataset con los datos de las familias que tienen un miembro de la familia de 0 a 5 años.

Para más información sobre sobre el código fuente de la ETL realizada en la ingeniería de datos.

"https://github.com/wladimiravila/modelos-infancia/blob/main/glue/sisben_families.py"

Ilustración 7Presto en Athena AWS



Para obtener el dataset final se filtraron los niños de 0 a 5 años agrupando datos del hogar creando estos campos:

- **data.suma_gastos_hogar**: Suma de los gastos del hogar (arriendo, servicios públicos, salud, educación, alimentos).
- **data.suma_ingresos_hogar**: Suma de los ingresos del hogar (salario, familia en acción, Colombia Mayor, otros subsidios).
- **data.empleado_o_similar**: Indicador de si al menos un miembro del hogar es empleado o dueño de empresa.
- **data.alguien_cotiza_pension_hogar**: Indicador de si al menos un miembro del hogar cotiza a pensiones.
- **data.max_nivel_educativo_hogar**: Nivel educativo más alto en el hogar.

Finalmente, se eliminó la variable ide_ficha_origen_hogar para la realización del modelo, ya que solo servía para agrupar e identificar los hogares.

Modelado y algoritmos

En esta fase del proyecto se utilizaron algoritmos de análisis predictivo de regresión y clasificación para corroborar las hipótesis, generando conclusiones importantes tras usar métricas de evaluación de los modelos utilizados. Es importante mencionar lo siguiente:

1. No se hizo tratamiento de valores faltantes, el único valor faltante fue el municipio de Jambaló, el cual fue descartado para el análisis.
2. Anexos del proyecto.
   - Anexo 06: Categorías de niños sin servicios básicos.
   - Anexo 07: Correlación gasto y clasificación
   - Anexo 08: Repositorio con los modelos desarrollados.

Fase de Evaluación. Todos los modelos se desarrolló la validación cruzada.

Valle del Cauca

Modelo	Precisión	Recall	F1 Score	Accuracy	Hiperparametros	estimadores	Tecnica de Normalizacion	Curva ROC
Modelo de Regresión Logística	A: 0.82	A: 0.87	A: 0.84	0.76%	{'C': np.float64(7.430820019313862), 'class_weight': None, 'penalty': 'l2'}	No aplica	StandardScaler	0.95
	B: 0.69	B: 0.67	B: 0.68					0.84
	C: 0.74	C: 0.75	C: 0.75					0.94
	D: 0.67	D: 0.29	D: 0.40					0.98
Modelo Random Forest	A: 0.88	A: 0.90	A: 0.89	82.37%	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2}	450	MinMaxScaler	0.97
	B: 0.77	B: 0.79	B: 0.78					0.91
	C: 0.81	C: 0.80	C: 0.88					0.96
	D:0.75	D: 0.28	D: 0.41					0.97
Modelo Xgboost	A: 0.85	A: 0.90	A: 0.87	79.31%	{'colsample_bytree': 1.0, 'learning_rate': 0.1, 'max_depth': 6, 'subsample': 0.8}	500	MinMaxScaler	0.97
	B: 0.73	B: 0.70	B: 0.71					0.90
	C: 0.78	C: 0.78	C: 0.78					0.96
	D: 0.78	D: 0.27	D: 0.40					0.98

Tabla 2 Modelos de clasificación Valle del Cauca

Norte Del Cauca

Modelo	Precisión	Recall	F1 Score	Accuracy	Hiperparametros	estimadores	Tecnica de Normalizacion	Curva ROC
Modelo de Regresión Logística	A: 0.89	A: 0.98	A: 0.92	84%	{'C': np.float64(3.8672288491177444), 'class_weight': None, 'penalty': 'l2'}	No aplica	StandardScaler	0.94
	B: 0.64	B: 0.46	B: 0.54					0.87
	C: 0.71	C: 0.78	C: 0.74					0.97
	D: 0.52	D: 0.34	D: 0.42					0.99
Modelo Random Forest	A: 0.92	A: 0.98	A: 0.95	87.89%	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2}	250	MinMaxScaler	0.97
	B: 0.76	B: 0.62	B: 0.69					0.93
	C: 0.79	C: 0.76	C: 0.79					0.98
	D: 1	D: 0.11	D: 0.20					0.99
Modelo Xgboost	A: 0.93	A: 0.97	A: 0.95	88%	{'colsample_bytree': 0.8, 'learning_rate': 0.1, 'max_depth': 6, 'subsample': 0.8}	200	MinMaxScaler	0.97
	B: 0.73	B: 0.67	B: 0.70					0.93
	C: 0.79	C: 0.76	C: 0.78					0.98
	D: 0.75	D: 0.32	D: 0.45					0.99

Tabla 3 Modelos de clasificación norte del Cauca

Modelo	Precisión	Recall	F1 Score	Accuracy	Hiperparametros	estimadores	Tecnica de Normalizacion	Curva ROC
Modelo de Regresión Logística	A: 0.83	A: 0.88	A: 0.86	77.00%	{'C': np.float64(36.70373762529 39), 'class_weight': None, 'penalty': 'l2'}	No Aplica	StandardScaler	0.95
	B: 0.69	B: 0.65	B: 0.67					0.85
	C: 0.74	C: 0.76	C: 0.75					0.94
	D: 0.63	D: 0.27	D: 038					0.97
Modelo Random Forest	A: 0.88	A: 0.91	A: 0.90	82.76%	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2}	500	MinMaxScaler	0.97
	B: 0.78	B: 0.77	B: 0.77					0.91
	C: 0.80	C: 0.81	C: 0.81					0.96
	D: 0.79	D: 0.29	D: 0.42					0.97
Modelo Xgboost	A: 0.89	A: 0.92	A: 0.90	85.00%	{'colsample_bytree': 1.0, 'learning_rate': 0.1, 'max_depth': 6, 'subsample': 0.8}	500	MinMaxScaler	0.97
	B: 0.80	B: 0.79	B: 0.79					0.90
	C: 0.85	C: 0.84	C: 0.84					0.96
	D: 0.91	D: 0.60	D: 0.72					0.98

Tabla 4 Modelos de clasificación Valle y norte del Cauca

El ajuste de parámetros de estimadores se realizó por medio de GridSearchCV, y dependiendo del set de dato se observan variaciones en los resultados a nivel de acurracy y otras métricas.

Para el modelo de regresión logística se aplicaron pruebas adicionales, estos son las siguientes:

MAE, MSE, Curvas de precisión y recall, Log loss.

Para mayor detalle de los modelos realizados, métricas empleadas, y otras graficas se adjunta Anexo 8 "https://github.com/wladimiravila/modelos-infancia"

6.4.Conclusiones de las Hipótesis

- **Hipótesis 1:**

Tomando como referencia la clasificación de los hogares que realiza actualmente el Sisbén para valorar el desarrollo integrar, se puede predecir la categoría de un niño menor o igual a 5 años basado en la variable Gastos, con una precisión superior o igual del 80 % de como lo hace el Sisbén actualmente. Tanto del Valle y norte del Cauca.

Modelo de comprobación de la hipótesis:

Desarrollo de un modelo de Regresión lineal simple entre las variables gastos y la clasificación, diferenciando tres modelos: el primero solo para el Valle del Cauca, el segundo para el norte del Cauca y el ultimo para los departamentos de forma consolidada.

Conclusión de la hipótesis

La variable ingresos es la variable clave en el cálculo econométrico de evaluación categórica de los hogares para el Sisbén, se decide tomar la variable gastos para el presente proyecto de investigación debido a que representa mayor consistencia en el cálculo al ser una variable con menos valores faltantes.

El gasto representa las salidas económicas de un hogar, al poder compararlo y/o relacionarlo con otras variables se tienen alternativas diferentes al método tradicional, y así poder evidenciar que la variable Gastos es fundamental dentro del estudio realizado.

Con respecto a la relación gasto y clasificación se evidencia una relación positiva poco significativa, es decir: La relación entre las variables suma de gastos del hogar y clasificación estadísticamente no es significativa debido a que solo representa el valor 0.032 para el R-cuadrado. La variable clasificación solo es explicada en una pequeña proporción por la variable suma de gastos del hogar, debido a la varianza que existe entre la variable dependiente e independiente.

Posteriormente al realizar el cálculo de coeficiente de correlación de Pearson se identifica que el grado de correlación entre las dos variables es del 0.18, entendiendo que a medida que sube los gatos del hogar, incrementaría la clasificación del niño.

El coeficiente positivo para la variable clasificación sugiere que hay una tendencia general de aumento en los gastos del hogar con cambios en la categoría clasificada, pero esta tendencia es débil.

El Coeficiente de la variable clasificación esta dado por el valor 4.516e+05: Esto indica que, en promedio, un cambio unitario en la categoría clasificada se asocia con un aumento de 451,600 en los gastos del hogar. Antes esto la hipótesis se descarta.

- **Hipótesis 2:**

Tomando como base de datos la proporcionada por el Sisbén, se decide analizar los niños menores o iguales de 5 años que habitan viviendas sin servicios públicos y con dificultades para obtener el agua y condiciones sanitarias inadecuadas, y así identificar si tienden a estar en categorías más marginales.

Modelo de comprobación de la hipótesis:
Es una hipótesis que busca analizar desde la descripción de los datos y así verificar que representan.

Hacer una consulta que permita contar los niños con esas características de vivienda y verificar si tienden a estar en las categorías A y B de clasificación, con una precisión igual o superior del 80 % de como lo hace el Sisbén. Diferenciando tres consultas: Primero solo para Valle del Cauca, segundo para el Norte del Cauca y por último para los departamentos de forma consolidada.

Conclusión de la hipótesis
Para la base de datos Sisbén se evidencia lo siguiente:
Siendo la categoría A y B las que contienen mayor cantidad de distribución de datos en las regiones estudiadas:

- Categoría A: 111,744

- Categoría B:89,907

Esto representa que más del 60% de los hogares analizados están en categorías marginales.

Al analizar los datos de los niños de 0 a 5 años que no tienen los siguientes servicios:

- Alcantarillado
- Gas
- Acueducto
- Energía

Se evidencia una cantidad de datos de niños sin estos servicios básicos:

- categoría A: 9,756
- categoría B: 1,785

esto representa que los niños de estas categorías en gran medida si tienen cubierto estas necesidades básicas. Luego, e detalla un poco más el análisis, analizando los niños que tiene problemas para conseguir acceso a agua:

- categoría A: 42,176
- categoría B: 16,772

El agua representa un dato no menor debido a que los niños de las categorías A; quien son los que están en condiciones más marginales, en un porcentaje superior al 30% no cumplen esta necesidad básica.
Seguidamente, se realiza un análisis con los niños que tienen condiciones sanitarias inadecuadas:

- categoría A: 50,207
- categoría B: 17,884

Evidencian que los niños de estas categorías económicas en un alto porcentaje tienen cubierta esta necesidad básica.

Finalmente, al filtrar los niños con que cumplen con todas condiciones previamente descritas se tienen la siguiente distribución de datos

- categoría A: 8,984
- categoría B: 1,523

Si se compara este resultado con los datos de la clasificación del Sisbén, el resultado es del 8% de los niños con estas condiciones ubicados en el total de la categoría A, por consiguiente, la hipótesis es acertada. Es importante resaltar que este tipo de niños tienen mayor dificultada al no contar con estos servicios básicos, siendo niños en condiciones de marginalidad extrema.
Es importante resaltar que los datos que mas influyen es: condiciones sanitarias y agua, ya que en gran porcentaje no cumplen con estas necesidades e influyen en la medida analizada.

- **Hipótesis 3:**

Tomando de base los datos del Sisbén para categorizar lo hogares, se analiza la correlación de las variables de la población de niños menor o igual de 5 años (Primera infancia), y determinar si la variable número de personas que trabajan en el hogar probablemente influye la clasificación del infante.

Modelo de comprobación de la hipótesis:
Generar la correlación de las variables de la primera infancia y describir la importancia económica de las variables.

Conclusión de la hipótesis
Al realizar el cálculo de coeficiente de correlación de Pearson se identifica que el grado de correlación entre las dos variables es del 0.66 , entendiendo que a medida que las personas del hogar trabajan, probablemente suba la clasificación del niño.

Esto es un hallazgo importante dentro del proyecto, ya que la variable personas que trabajan en el hogar y la clasificación de los niños son las que mayor tiene grado de correlación.

Se evidencia que existe un tipo de crecimiento poblacional positivo. Debido a que no solo aumenta la población del hogar, también aumenta el consumo de bienes y servicios, derivando así en incrementos productivos, que conllevan a probablemente mejorar la clasificación de los niños de estas edades.

Se toma de base de análisis la base de dato de la categorización de los hogares del Sisbén, al ser una fuente importante de estudio debido a la base ya clasifica los hogares según datos tabulados en la encuesta.

Al identificar que la clasificación económica del Sisbén se puede influenciar por el numero de personas empleadas del hogar, es importante analizar el contexto socio económico de lo dos departamentos, esto debido a que son regiones del país con altos índices de desempleo e informalidad, dato expuestos en el análisis descriptivo del presente proyecto de investigación.

- **Hipótesis 4:**

Se puede predecir con una precisión superior o igual del 80 % las categorías de los niños menores a o iguales 5 años, si tomamos un modelo de clasificación y no un modelo de regresión como lo hace el Sisbén para categorizar los niños en A, B, C, D.

Modelo de comprobación de la hipótesis:

Desarrollar 3 modelos de clasificación. Estos son: el random forest, regresión logística multiclase, y xgboost, diferenciando por cada uno de los tres modelos lo siguiente: el primer análisis solo para el Valle del Cauca, el segundo análisis solo para el norte del Cauca y el ultimo para los departamentos de forma consolidada.

Conclusión de la hipótesis

Al aplicar el modelo de regresión logística se evidencia un nivel bajo de precisión al dividir los modelos por departamentos o consolidarlos, teniendo un valor en cada una de las categorías por debajo del 70% al revisar la métrica de precisión.

Al aplicar el modelo de xgboost se evidencia un nivel baja de precisión al dividir los modelos por departamentos o consolidarlos, teniendo un valor en cada una de las categorías por debajo del 80% al revisar la métrica de precisión.

Al aplicar el modelo de random forest se evidencia un nivel superior de precisión del 80% al dividir los modelos por departamentos o consolidarlos, teniendo un valor en cada una de las categorías por debajo del 80% al revisar la métrica de precisión. Siendo este modelo es elegido al aplicar un modelo de clasificación.

Siendo un hallazgo importante debido a que se realizó una predicción tomando menos valores de los que los toma el Sisbén, Se realizar un modelo de clasificación y no un modelo de regresión de acuerdo con el modelo econométrico del Sisbén.

El modelo de random forest es un modelo donde los supuestos son mínimos en comparación con otros algoritmos de machine learning y/o otros modelos estadísticos.

El modelo de ramdon forest es el que mayor porcentaje de precisión brinda si se compara con un modelo probado y funcional como el propuesto por el Sisbén.
Estos son debido a que el modelo de ramdon forest no asume una relación lineal entre las variables independientes y la variable dependiente.

Inicialmente se asumió que la variable gastos era las más importante para la construcción de modelos, se puedo constatar en la comprobación de la hipótesis 1 que no esta afirmación no era correcta.

Esto permite analizar otras variables, evidenciando así que para el cumplimiento de este modelo no todas las características deben ser igualmente importantes. Esto es porque el modelo selecciona de manera aleatoria un subconjunto de características para cada división de los nodos en los árboles, lo que ayuda a determinar la importancia relativa de cada característica.

7. CONCLUSIONES

- Al analizar las diferentes variables contenidas dentro de la encuesta del Sisbén, se lograron identificar 32 con mayor peso de las 116 actuales. Pudiendo así simplificar la cantidad de variables que hace el cálculo actualmente del Sisbén. Encontrando mejoras de rendimiento al utilizar menos variables para los cálculos de clasificación de categoría de los hogares que contienen niños entre 0 y 5 años.

- El desarrollo de tres modelos de clasificación diferentes con la finalidad de predecir la categoría de los hogares, logro evidenciar otro modelo al utilizado actualmente por el Sisbén. proponiendo así un método diferente al actual.

- Al conocer el modelo de clasificación del Sisbén 4 y los datos actuales, se pudo analizar descriptivamente la situación actual del Valle y norte del Cauca, evidenciando problemas y factores que en muchos casos no contribuyen al desarrollo de los niños de la primera infancia.

- El desarrollo integral de un niño es influenciado por una variedad de factores socioeconómicos que abarcan aspectos físicos, emocionales, sociales y cognitivos. Estos factores pueden ser positivos y/o negativos, pueden tener efectos duraderos en el crecimiento y bienestar del niño. En un país neoliberal como Colombia, que históricamente ha buscado la inversión de capital privado para fomentar el crecimiento, la infancia toma un papel crucial e importante para el desarrollo. nunca podrá haber desarrollo sin crecimiento económico.

- La desigualdad económica en esta región del país es muy clara si se analizan las cifras del presente proyecto de investigación, claramente los niños que viven en condiciones de ruralidad son más propensos a tener condiciones menos favorables respecto a la calidad de vivienda, de igual forma se identifica que a medida que las personas de un hogar se emplean, la probabilidad de que un hogar de un niño cambie de categoría es alta, pero en este tipo de regiones el desempleo es una cifra preocupante, lo que da a entender que muchos hogares de estos niños están destinados a estar en condiciones de marginalidad, si no se trabaja con políticas económicas en estos territorios colombianos.

- Se logró predecir en el mejor de los casos una precisión al 80% la clasificación del Sisbén. Esto permite concluir que los modelos tienen aplicación real al generar eficiencias en la recopilación de los datos y que el modelo econométrico desarrollado por el Sisbén puede ser simplificado por un modelo de clasificación.

- En el proceso para cumplir este este proyecto, se determinó que el mejor etiquetado de los datos son las categorizaciones generadas por el Sisbén, es decir A, B, C y D. De esta manera se identifican los grupos de personas con situaciones más vulnerables, siendo la categoría A la de mayor riesgo, y a su vez D, la de menor riesgo. Permitiendo conocer no solo la estadística de los datos del Sisbén también el contexto socioeconómico de las variables y por qué son tenidas en cuenta por este sistema de información.

- Durante el desarrollo del proyecto se utilizaron varias herramientas de ciencia de datos para llevar a cabo los análisis, ejemplo de ellas fueron S3 de AWS para cargar los datos, AWS Glue para realizar transformaciones, AWS Athena para realizar querys y consultar los datos. Así como Presto y Apache Spark, Azure Data Factory, Azure Data databricks. Permitiendo reforzar conceptos importantes aprendidos durante la maestría.

- Se llevó a cabo un análisis descriptivo, geográfico y socioeconómico de los datos, donde se logró determinar un grupo de 32 variables dentro de la encuesta completa que contienen el mayor peso en la categorización. De esta manera, se desarrollaron tres modelos con la finalidad de predecir la clasificación en el Sisbén de las personas. Utilizando un número reducido de variables – preguntas en la encuesta, se generan todo tipo de posibles eficiencias en tiempo y recursos para las entidades encargadas de llevar a cabo estas encuestas.

- Algunas de eficiencias que se encontraron son: la posibilidad de disminuir la cantidad de recursos para llevar a cabo las encuestas, la posibilidad de mantener la base de datos más actualizada, puedan tener un mejor resultado para las personas que están en situaciones de vulnerabilidad.

- En la identificación de las variables que más peso tienen dentro de los modelos, se puede inferir cuales son los factores más determinantes en el bienestar o desarrollo de las personas en las regiones estudiadas. Siendo el empleo formal de una de las personas dentro del hogar la variable más determinante para las cuatro categorías. De esta manera, lo encontrado en el proyecto permiten sugerir prioridades para las políticas públicas que estén encaminadas a generar empleo formal, siendo este el factor que más beneficios tiene para los hogares en las regiones estudiadas.

El norte del Cauca ha enfrentado diversos desafíos económicos a lo largo de su historia. Esta área del departamento ha estado marcada por una combinación de factores que incluyen la presencia de conflictos armados, la falta de infraestructura adecuada, la limitada diversificación económica y la exclusión social.

Algunas de sus actividades económicas y/o puntos a tener en cuenta se centran en:

- Agricultura: La economía del norte del Cauca ha dependido tradicionalmente de la agricultura, con cultivos como la caña de azúcar, el café, el maíz y el plátano.

- Conflictos armados: Durante décadas, la región ha sido afectada por la presencia de grupos armados y el conflicto interno en Colombia. Históricamente se han presentado conflictos entre la guerrilla de las FARC y la fuerza pública en municipios ubicados sobre la cordillera Central como Caloto, Corinto, Miranda, Toribio y Florida.

- La Economía en su mayoría es informal: Muchos habitantes del norte del Cauca se han visto obligados a recurrir a la economía informal debido a la falta de oportunidades formales de empleo. Para el año 2023, la tasa de desocupación de Cauca fue 7,5% según datos del DANE, donde el País a corte 2024 la tasa de desempleo es del 11.3%.

- Programas de desarrollo: A pesar de los desafíos, se han implementado programas de desarrollo económico y social en la región, tanto por parte del gobierno como de organizaciones no gubernamentales y agencias internacionales. Algunos programas y/o políticas son:

    - Confacauca con el programa de atención integral de la niñez.
    - Consejos gremiales orientados por el ICBF [9]

Si bien la económica del municipio no está entre las más grandes del país, si es importante ya que aporta el 1.76 del PIB del País.

El Valle del Cauca, ubicado en el suroeste de Colombia, es una de las regiones económicamente más importantes del país. Su economía es diversificada y abarca varios sectores clave que contribuyen al desarrollo regional y nacional.

Algunos puntos importantes del valle del cauca.

- Agricultura y agroindustria: es una región agrícola productiva. Cultivos como la caña de azúcar, el café, el arroz, los plátanos, las frutas tropicales y los cultivos industriales son importantes para la economía local. estando la actividad agrícola dentro de las 3 principales actividades económicas de la región.
- Industria manufacturera: La región alberga una variedad de empresas en sectores como textiles, prendas de vestir, calzado, productos químicos, plásticos, metalurgia, muebles y productos farmacéuticos. La ciudad de Cali, la capital del departamento es un importante centro industrial y comercial, algunas empresas manufactureras son: Carvajal y Colombina, Chocorramo entre muchas otras.
- Infraestructura y transporte: El Valle del Cauca cuenta con una infraestructura de transporte desarrollada que incluye carreteras, ferrocarriles, puertos y aeropuertos. La región alberga el puerto de Buenaventura, el principal puerto marítimo de Colombia en la costa del Pacífico, que es crucial para el comercio internacional y la logística.
- En términos económicos el Valle del cauca es la tercera economía más grande del país, después de Bogotá y Antioquia. Representa el 9,7 por ciento del Producto Interno Bruto (PIB). Para el año 2022, la tasa de desempleo de Valle del Cauca fue 13,2% muy alto si lo comparamos contra el índice nacional para ese año del 10.3% [10]

Si bien la económica del departamento no está entre las más grandes del país, si esta entre las 3 grandes, aportando entre Bogotá, Antioquia y valle del Cauca el 50 % del PIB del País.

8. REFERENCIAS BIBLIOGRÁFICAS

[1] Departamento Nacional de Planeación, "Preguntas comunidad," Sisben, [En línea]. Disponible: https://www.sisben.gov.co/Paginas/preguntas-comunidad.html. [Accedido: 06-mayo-2024].

[2] Departamento Nacional de Planeación, "Gobierno extiende un año el plazo para solicitar la actualización de la encuesta Sisben IV," [En línea]. Disponible: https://www.dnp.gov.co/Prensa_/Noticias/Paginas/gobierno-extiende-un-ano-el-plazo-para-solicitar-la-actualizacion-de-la-encuesta-sisben-iv.aspx. [Accedido: 03-mayo-2024].

[3] Administradora de los Recursos del Sistema General de Seguridad Social en Salud, "El sistema de salud recibió $65,19 billones en el 2021," [En línea]. Disponible: https://www.adres.gov.co/sala-de-prensa/noticias/Paginas/El-sistema-de-salud-recibi%C3%B3-$65,19-billones-en-el-2021.aspx. [Accedido: 06-jun-2024].

[4] Ministerio de Vivienda, Ciudad y Territorio, "Mi Casa Ya," [En línea]. Disponible: https://www.minvivienda.gov.co/viceministerio-de-vivienda/mi-casa-ya#:~:text=Depender%C3%A1%20de%20la%20clasificaci%C3%B3n%20que,durante%20la%20vigencia%20de%202024. [Accedido: 15- mayo -2024].

[5] Ministerio de Educación Nacional, "Gobierno del Cambio incrementará en 25% el presupuesto para el PAE en 2024," [En línea]. Disponible: https://www.mineducacion.gov.co/portal/salaprensa/Comunicados/417830:Gobierno-del-Cambio-incrementara-en-25-el-presupuesto-para-el-PAE-en-2024. [Accedido: 09- mayo -2024].

[6] ICETEX, "Créditos," [En línea]. Disponible: https://web.icetex.gov.co/creditos. [Accedido: 06-jun-2024].

[7] Defensoría del Pueblo, "Gobierno alista reforma al Sisbén por trampas que cuestan unos 364.000 millones al año," [En línea]. Disponible: https://www.defensoria.gov.co/-/gobierno-alista-reforma-al-sisb%C3%A9n-por-trampas-que-cuestan-unos-364.000-millones-al-a%C3%B1o. [Accedido: 01-

jun-2024].

[8] Departamento Nacional de Planeación, "Sisben IV: Instrumento medir capacidad pago y focalizar gasto salud," [En línea]. Disponible: https://colaboracion.dnp.gov.co/CDT/Prensa/Sisben-IV-instrumento-medir-capacidad-pago-y-focalizar-gasto-salud.pdf. [Accedido: 03-jun-2024].

[9] Amazon Web Services, "¿Qué es la regresión logística?", [Online]. Available: https://aws.amazon.com/es/what-is/logistic-regression/. [Accessed: 07-Jun-2024].

-

[10] F. Perales, "FEIR 45: Regresión logística," Apr. 4, 2019. [Online]. Available: https://gauss.inf.um.es/feir/45/. [Accessed: Jul. 16, 2024].

- [11] "¿Qué es la regresión logística?," *Amazon Web Services*, Jan. 6, 2023. [Online]. Available: https://aws.amazon.com/es/what-is/logistic-regression/. [Accessed: Jul. 16, 2024].

[12] M. Qwak, "XGBoost versus Random Forest," Dec. 19, 2022. [Online]. Available: https://www.qwak.com/post/xgboost-versus-random-forest. [Accessed: Jun. 7, 2024].

[13] S. Cole, "Random Forests in Machine Learning," *Shelf.io*, Feb. 18, 2022. [Online]. Available: https://shelf.io/blog/random-forests-in-machine-learning/. [Accessed: Jul. 16, 2024].

[14] Spiceworks, "XGBoost vs. Random Forest vs. Gradient Boosting," Feb. 21, 2024. [Online]. Available: https://www.spiceworks.com/tech/artificial-intelligence/articles/xgboost-vs-random-forest-vs-gradient-boosting/. [Accessed: Jun. 7, 2024].

[15] G. Kumar, "Gradient Boosting & Extreme Gradient Boosting (XGBoost)," *Medium*, Jan. 23, 2023. [Online]. Available: https://medium.com/@gajendra.k.s/gradient-boosting-extreme-gradient-boosting-xgboost-de865b871203. [Accessed: Jul. 16, 2024].

[16] Departamento Nacional de Planeación, "Estadísticas," [En línea]. Disponible: https://www.dnp.gov.co/LaEntidad_/subdireccion-general-prospectiva-desarrollo-nacional/direccion-desarrollo-rural-sostenible/Paginas/estadisticas.aspx. [Accedido: 06-jun-2024].

[17] Ministerio de Educación Nacional, "Recurso 1," [En línea]. Disponible: https://www.mineducacion.gov.co/1759/articles-356787_recurso_1.pdf. [Accedido: 06- mayo -2024].

[18] Instituto Nacional de Salud, "Desigualdades sociales," Informe de Salud, pág. 21. [En línea]. Disponible: https://www.ins.gov.co/direcciones/ons/informes/6.%20desigualdades%20sociales.pdf. [Accedido: 06-jun-2024].

[19] Departamento Nacional de Planeación, "Documento CONPES 3877," pág. 34. [En línea]. Disponible: https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3877.pdf. [Accedido: 20-abril-2024].

[20] "¿Qué es el Sisbén?," Sisbén, [En línea]. Disponible: https://www.sisben.gov.co/Paginas/que-es-sisben.html. [Accedido: 02- mayo -2024].

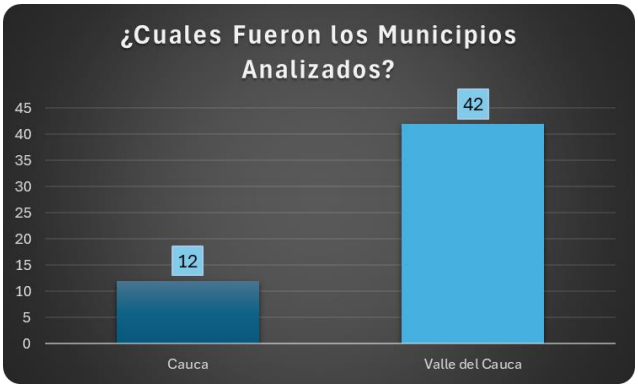
[21] Instituto Colombiano de Bienestar Familiar, "Encuesta Nacional de Situación Nutricional," [En línea]. Disponible: https://www.icbf.gov.co/bienestar/nutricion/encuesta-nacional-situacion-nutricional. [Accedido: 05-jun-2024].

[22] M. C. S. G. Cornejo, "Modelo Predictivo de la Deserción Escolar en Educación Superior: una Aproximación desde la Minería de Datos Utilizando la Metodología CRISP-DM," ResearchGate, Nov. 2023. [En línea]. Disponible: https://www.researchgate.net/publication/375807055_Modelo_Predictivo_de_la_Desercion_Escolar_en_Educacion_Superior_una_Aproximacion_desde_la_Mineria_de_Datos_Utilizando_la_Metodologia_CRISP-DM. [Accedido: 01-jun-2024].

Anexo 02: Análisis descriptivo.

Se analizo un total de 55 Municipios, 13 Municipios del norte del Cauca y 42 Municipios del Valle del Cauca, se evidencio que el Municipio Jámbalo del norte del Cauca no existe niños con la edad de 0 a 5 años, es importante aclarar que este es un municipio y a su vez es un resguardo indígena.

Los registros de la base de datos corresponden a 3.019.650 registros.



Se analizó para el norte del Cauca los niños menores o iguales de 5 años, así:

En los hogares del norte del Cauca donde hay más niños es el Municipio de Santander de Quilichao, comparado con los otros municipios del norte del Cauca, teniendo un 51 % mayor que los de Puerto Tejada; municipio que esta segundo según la cantidad de niños.

La economía del municipio de Santander de Quilichao se sustenta en el sector primario, concretamente en la agricultura como azúcar, café, piña y yuca.

Cauca según cifras del DANE, representa el 1.3 del PIB nacional a corte del año 2022, siendo Santander de Quilichao el segundo mayor Municipio del Cauca en población y en PIB del Cauca, por detrás de la capital Popayán. [1]

La niñez y la economía son dos aspectos fundamentales relacionados pero que se abordan separadamente, aspectos como pobreza, educación, costos de crianza, conflictos armados son aspectos relacionados con la gestión de las políticas económicas, variables importantes de

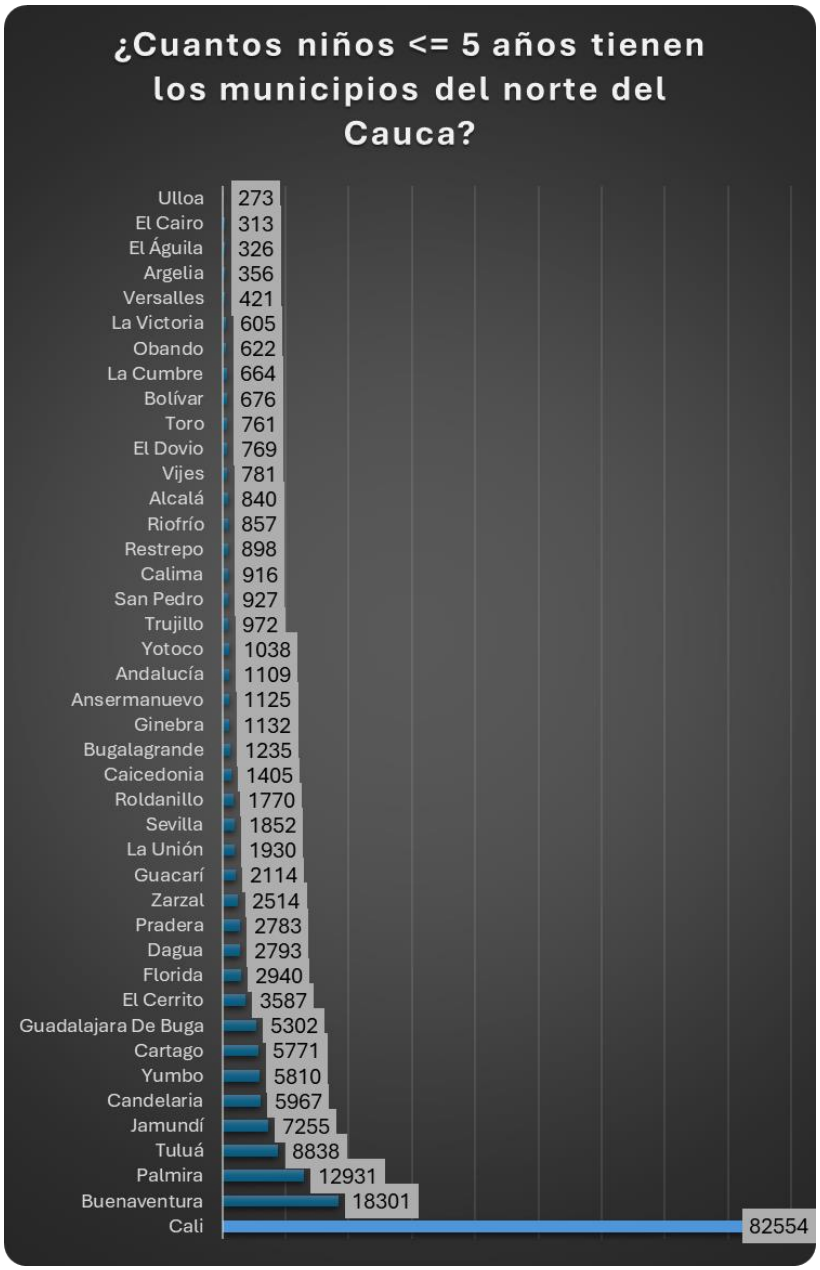
una región, fundamentales para el desarrollo, e intrínsicamente relacionadas con el crecimiento de una región.

Otro dato importante por analizar es Toribio Cauca, Municipio que ha contabilizado más de 600 ataques de parte de las guerrillas desde el año 1980, claramente esto podría ser una causa de migración de poblaciones que se ve reflejado en la densidad poblacional del municipio.

Según datos de la defensoría del pueblo; junio 2 del 2023, Disidencias de las Farc convirtieron a Toribio (Cauca) en centro de entrenamiento de niños reclutados. “Debido a que la Columnas Móviles de centros armados tiene el centro de operaciones y finanzas en el municipio, también lo convirtieron en zona de movilidad y tránsito que, gracias a su topografía, se convierte en zona de difícil acceso para la institucionalidad del Estado”. [2]

Según datos de la web de la comisión para el esclarecimiento de la verdad, los municipios más afectados del del Cauca por el conflicto armado son: Miranda, Caloto, Toribio, Suárez, Santander de Quilichao, Jambaló, Buenos Aires, Corinto, Morales, Cajibío, El Tambo, Argelia, Balboa y Patía. [3]

Para los municipios del departamento del Valle del Cauca se evidencia lo siguiente:



Se evidencia en la figura que el Municipio de Cali es el que tiene mayor representación de niños si lo comparamos con otros municipios.

Cali tiene un total de 2.172.527 de población total, en la base de datos 82554 son niños menores o iguales de 5 años. Representando así un 5% de la data del Municipio para el presente análisis.

Si se analiza Cali (Valle del Cauca) respecto a Santander de Quilichao (Norte del Cauca) se evidencia una diferencia de 76747 registros. Ante esto es importante diferenciar el análisis del Norte del Cauca y Valle del Cauca.

El municipio de Cali tiene 42%, Buenaventura 9% y Palmira el 7% de la data analizada para el Valle del Cauca de los 82554 niños menores a 5 años.

En cuanto a la distribución de la población, Cali es una ciudad habitada por gente joven según estadísticas del DANE a corte del 2022. El grueso de la población es menor de 40 años. También se observa una mayor población de mujeres en casi todos los rangos de edad, excepto entre la población más joven. [4]

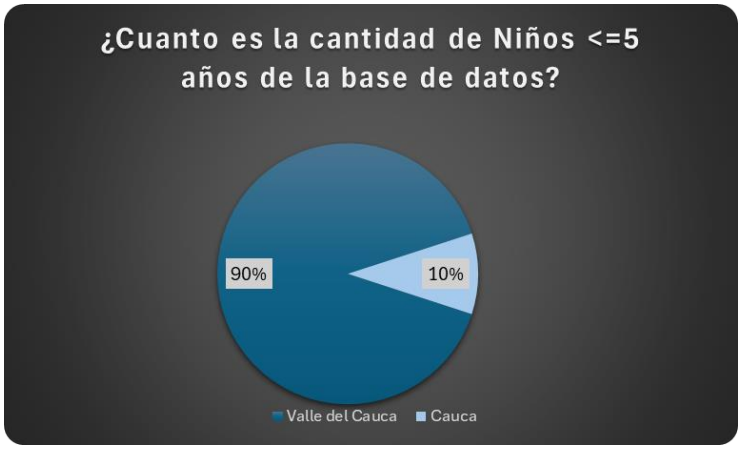
De acuerdo con las estadísticas de la secretaria de Salud Pública Municipal (2005) la tasa bruta de natalidad (TBN) de la ciudad es de 20,2 recién nacidos por cada 1.000 habitantes, un poco menor que la del país (22,0). [5]

Otro Municipio por analizar Ulloa (Valle del Cauca) con un total de 273 registros de niños menores o iguales a 5 años, es el municipio más pequeño del territorio vallecaucano, a corte del año 2022, según datos del DANE tiene un total de 5.457 densidad poblacional. Un 19% de niños en este rango de edad. [6]

En la base de datos se destaca la cantidad de hogares en el norte del Cauca y Valle del Cauca con 1.297.570 Hogares.

Estos hogares (1.297.570) tienen una representación en la base de datos total de 214372 de hogares que tiene al menos 1 niño menor o igual a 5 años: este dato corresponde el 17% de la data. Este dato es representativo ya que este 17% es la base con la que se van a hacer los modelos de predicciones.

El departamento del Valle del Cauca tiene el 90% y el Norte del Cauca el 10% respectivamente, ante lo mencionado se divide el análisis por departamento para no tener sesgo en el resultado e inconvenientes en la generación de conclusiones.



Con respecto a la cantidad de niños del norte del Cauca (10% de la data) se evidencia lo siguiente:

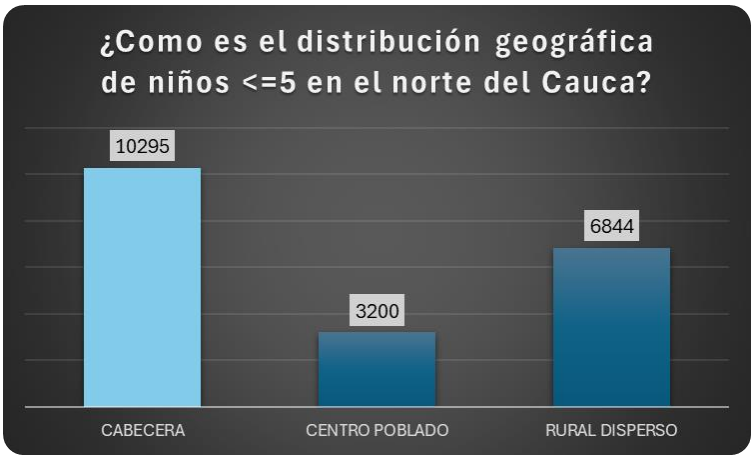
Siendo Santander De Quilichao, Puerto Tejada y Miranda los que tienen mayor concentración de niños de 0 a 5 años en el norte del Cauca. Con 29%,14%,9% respectivamente, no varía la distribución de la data luego de aplicar los filtros de los hogares en la base de datos.

Con respecto a la cantidad de niños del Valle del Cauca (90% de la data) se evidencia lo siguiente:

Siendo Cali, Buenaventura y Palmira los que tienen mayor concentración de niños de 0 a 5 años en el departamento. Con 43%,9%,7% respectivamente, no varía la distribución de la data luego de aplicar los filtros de los hogares en la base de datos.

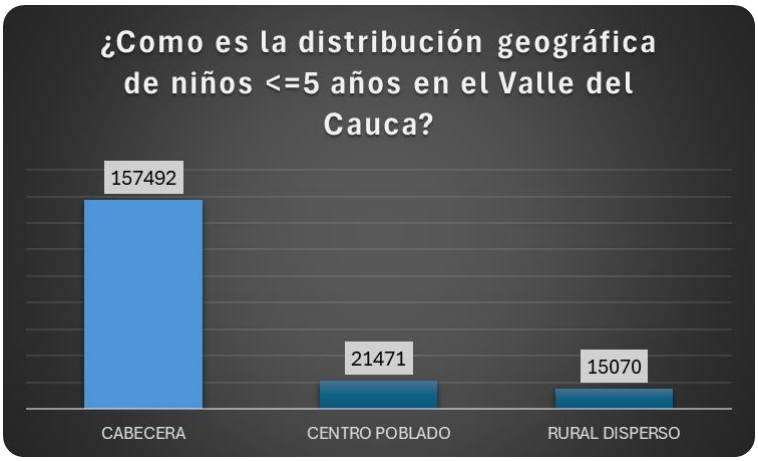
Se analiza la distribución por tipo de vivienda de los niños de los departamentos, encontrando lo siguiente:

Para el norte del Cauca.



El norte del Cauca tiene más distribución de niños menores o iguales a 5 años en las cabeceras municipales, pero también se evidencia que existen 34 % de estos niños en una distribución de rural disperso, un dato importante de aclarar ya que la población del departamento está compuesta por comunidades campesinas, afrodescendientes e indígenas que tienen sus propias formas de organización.

Para el Valle del Cauca.



El departamento del Valle del Cauca tiene más población de niños menores o iguales a 5 años en las cabeceras municipales, y se evidencia que existen una distribución similar de niños en centros poblados y en ruralidad disperso.

Según el DANE, 3,5 millones de personas ingresaron a la situación de pobreza en Colombia en 2020, siendo Bogotá & Cundinamarca (1,3 millones de personas), Valle del Cauca (518 mil personas) y Atlántico (+345 mil personas) las regiones más afectadas. En el caso de Cali, la tasa de pobreza pasó de 21,9 por ciento a 36,3 por ciento. Dato importante en la presente investigación ya que la clasificación del Sisbén se basa en los ingresos monetarios. [7]

Según datos tomados del periódico el tiempo, articulo marzo 16 de 2022 “El 65,8 % de las personas que están en el Sisbén IV son pobres”. [8]

En la siguiente tabla de frecuencias se evidencia la distribución de datos según tipo de ubicación.

Nombre Depto.	Cantidad Niños <=5 años	Distancia	Frecuencia
Cauca	3200	CENTRO POBL	1%
Cauca	6844	RURAL DISPEF	3%
Cauca	10295	CABECERA	5%
Valle del Cauc	15070	RURAL DISPEF	7%
Valle del Cauc	21471	CENTRO POBL	10%
Valle del Cauc	157492	CABECERA	73%
	214372		

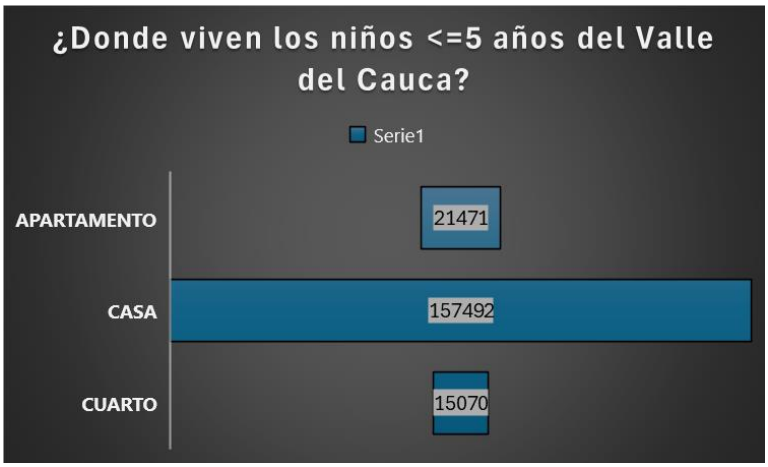
Según la data, existen una concentración de datos en las cabeceras en el departamento del Valle del Cauca, donde sus principales ciudades son Cali y Buenaventura. En el caso del norte del Cauca la distribución es muy similar, esto es principalmente porque esta región del país está compuesta por comunidades campesinas, afrodescendientes e indígenas.

Al analizar los hogares con niños menores o iguales de 5 años por tipo de vivienda se evidencian los siguientes:

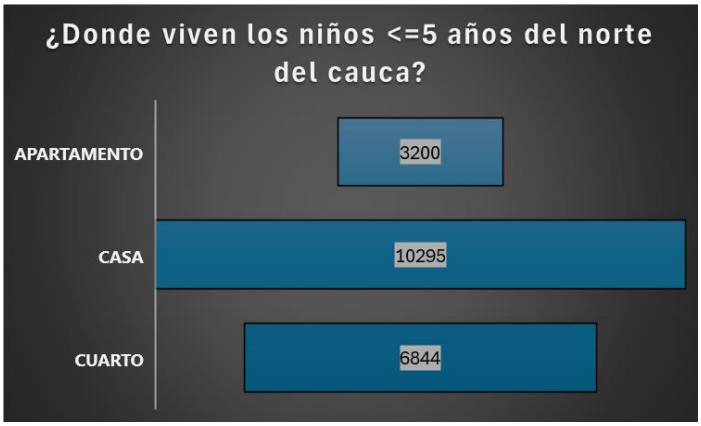
Tipo Vivienda	Cantidad Niños <=5 años	Frecuencia
CUARTO	21914	10%
APARTAMENTO	24671	12%
CASA	167787	78%
	214372	

En general los niños de este rango de edad viven en Casas si se analizan el Valle del Cauca y el norte del Cauca, luego en apartamentos y posteriormente en Cuartos.

Al detallar el Valle del Cauca se tiene un 54 % de niños que viven en casa, 34% que viven en Cuarto y 16 % que viven en apartamento. Cifra que es importante mencionar ya que claramente estos niños (16%) podrían tener condiciones de hacinamiento. En la siguiente grafica se ilustra los niños de la primera infancia en el Valle del Cauca.



Para el norte del Cauca, un 81 % de niños de la primera infancia viven en Casas, 11% viven en Apartamento y 8 % que viven en Cuartos. Cifra que refleja que los niños de esta edad viven en gran mayoría en ruralidad, pero posiblemente en la gran mayoría alojados en casas y/o apartamentos.



Luego, al revisar este tipo de vivienda por tipo de material construido se evidencia la siguiente tabla de frecuencias los tipos materiales y tipos de viviendas de los niños de la primera infancia para el Valle del Cauca y norte del Cauca.

Tipo de vivienda	Tipo de Material	Cantidad de Niños	Frecuencia
APARTAMENTO	Zinc, tela, lona, cartón, latas, desechos, plásticos	120	0,1%
APARTAMENTO	Tapia pisada, adobe	534	0,2%
APARTAMENTO	Material prefabricado	572	0,3%
APARTAMENTO	Guadua, caña, esterilla, otro vegetal	1157	0,5%
APARTAMENTO	Madera burda, tabla, tablón	2363	1,1%
APARTAMENTO	Bahareque	3090	1,4%
APARTAMENTO	Bloque, ladrillo, piedra, madera pulida	16835	7,9%
CASA	Zinc, tela, lona, cartón, latas, desechos, plásticos	793	0,4%
CASA	Material prefabricado	1829	0,9%
CASA	Tapia pisada, adobe	2059	1,0%
CASA	Guadua, caña, esterilla, otro vegetal	2626	1,2%
CASA	Bahareque	4486	2,1%
CASA	Madera burda, tabla, tablón	11643	5,4%
CASA	Bloque, ladrillo, piedra, madera pulida	144351	67,3%
CUARTO	Tapia pisada, adobe	340	0,2%
CUARTO	Zinc, tela, lona, cartón, latas, desechos, plásticos	392	0,2%
CUARTO	Material prefabricado	520	0,2%
CUARTO	Guadua, caña, esterilla, otro vegetal	2216	1,0%
CUARTO	Madera burda, tabla, tablón	3411	1,6%
CUARTO	Bahareque	6206	2,9%
CUARTO	Bloque, ladrillo, piedra, madera pulida	8829	4,1%
		214372	

Las casas con material de bloque, ladrillo, piedra con 67% son las que más prevalecen en la base de datos.

Con mayor detalle se analiza en la siguiente imagen.

Tipo de Vivienda - Tipo de Material	Niños	Frecuencia
Cauca	**20339**	
APARTAMENTO	**3200**	
Bahareque	1003	31%
Bloque, ladrillo, piedra, madera pulida	1913	60%
Guadua, caña, esterilla, otro vegetal	149	5%
Madera burda, tabla, tablón	80	3%
Material prefabricado	20	1%
Tapia pisada, adobe	29	1%
Zinc, tela, lona, cartón, latas, desechos, plásticos	6	0%
CASA	**10295**	
Bahareque	1181	11%
Bloque, ladrillo, piedra, madera pulida	8473	82%
Guadua, caña, esterilla, otro vegetal	322	3%
Madera burda, tabla, tablón	103	1%
Material prefabricado	50	0%
Tapia pisada, adobe	141	1%
Zinc, tela, lona, cartón, latas, desechos, plásticos	25	0%
CUARTO	**6844**	
Bahareque	2903	42%
Bloque, ladrillo, piedra, madera pulida	2575	38%
Guadua, caña, esterilla, otro vegetal	589	9%
Madera burda, tabla, tablón	577	8%
Material prefabricado	34	0%
Tapia pisada, adobe	60	1%
Zinc, tela, lona, cartón, latas, desechos, plásticos	106	2%
Valle del Cauca	**194033**	
APARTAMENTO	**21471**	
Bahareque	2087	10%
Bloque, ladrillo, piedra, madera pulida	14922	69%
Guadua, caña, esterilla, otro vegetal	1008	5%
Madera burda, tabla, tablón	2283	11%
Material prefabricado	552	3%
Tapia pisada, adobe	505	2%
Zinc, tela, lona, cartón, latas, desechos, plásticos	114	1%
CASA	**157492**	
Bahareque	3305	2%
Bloque, ladrillo, piedra, madera pulida	135878	86%
Guadua, caña, esterilla, otro vegetal	2304	1%
Madera burda, tabla, tablón	11540	7%
Material prefabricado	1779	1%
Tapia pisada, adobe	1918	1%
Zinc, tela, lona, cartón, latas, desechos, plásticos	768	0%
CUARTO	**15070**	
Bahareque	3303	22%
Bloque, ladrillo, piedra, madera pulida	6254	41%
Guadua, caña, esterilla, otro vegetal	1627	11%
Madera burda, tabla, tablón	2834	19%
Material prefabricado	486	3%
Tapia pisada, adobe	280	2%
Zinc, tela, lona, cartón, latas, desechos, plásticos	286	2%

Algunos aspectos por resaltar en las gráficas anteriores

- 31% de los apartamentos del norte del Cauca son Bahareque,
- 11% de las casas del norte del Cauca son Bahareque.
- 42 % de los cuartos del norte del Cauca son Bahareque; cifra más representativa para este departamento.
- 10% de los apartamentos del Valle del Cauca son Bahareque.
- 7% de las casas del Valle del Cauca son en madera.
- 19% de los cuartos del Valle del Cauca son en madera.
- 22% de los cuartos del Valle del Cauca son en Bahareque.

Se analizan estos tipos de vivienda según si cuenta con energía, se obtiene lo siguiente:

- El 6 % de los municipios del norte del Cauca no cuenta con energía. El municipio de Santander De Quilichao (3050 niños) es el más representativo sin este servicio.
- El 6 % de los municipios del Valle del Cauca no cuenta con energía. El municipio de Buenaventura (3715 niños) es el más representativo sin este servicio.
- Con una media de 6% para los dos departamentos, dato que es representativo para el proyecto.

Luego, al analizar la cantidad de habitaciones de los tipos de vivienda para los dos departamentos, se obtiene lo siguiente:

- 51% de los niños del Norte del Cauca viven en tipo de vivienda que solo cuenta con 1 habitación. Cifra representativa.
- El 34% de los niños del Valle del Cauca viven en tipo de vivienda que cuenta con 3 habitaciones. Cifra representativa.

Al analizar este tipo de vivienda por el tipo acueducto se evidencia los siguiente:

Para el norte del Cauca.

Tipo de Sanitario	Cantidad de Hogares	Cantidad de Hogares
Con conexión a alcantarillado	9498	47%
Con conexión a pozo séptico	8223	40%
Letrina, bajamar	1086	5%
No tiene	821	4%
Sin conexión a alcantarillado ni a pozo séptico	711	3%

Los niños del norte del Cauca según las cifras mencionadas, en su gran porcentaje viven en casas (81%) , mayor mente construidas en bareque (42%) , con conexión a alcantarillado (47%), aunque es importante mencionar que existe cifras importante como la cantidad de

niños que su principal modo de acceso a un tipo de sanitario es por medio de una conexión a pozo séptico (40%).

Para el Valle del Cauca.

Tipo de Sanitario	Cantidad de Hogares	Frecuencia
Con conexión a alcantarillado	152581	79%
Con conexión a pozo séptico	19846	10%
Letrina, bajamar	10426	5%
No tiene	1541	1%
Sin conexión a alcantarillado ni a pozo séptico	9639	5%
	194033	

Los niños del Valle del Cauca según las cifras mencionadas, en su gran porcentaje viven en casas (54%) , mayor mente construidas en bareque (22%) , con conexión a alcantarillado (79%).

Al analizar este tipo de vivienda por el tipo de origen de agua se evidencia los siguiente:

Para el Norte del Cauca.

Tipo de Origen de Agua	Cantidad de Hogares	Frecuencia
Acueducto	12061	59%
Agua lluvia	132	1%
Aguatero	157	1%
Carrotanque	8	0%
No tiene	151	1%
Pila pública	119	1%
Pozo con bomba	1450	7%
Pozo sin bomba, jagüey	2185	11%
Río, quebrada, manantial o nacimiento	4076	20%
	20339	

Para el Valle del Cauca.

Tipo de Sanitario	Cantidad de Hogares	Frecuencia
Acueducto	162818	84%
Agua lluvia	11322	6%
Aguatero	559	0%
Carrotanque	829	0%
No tiene	603	0%
Pila pública	2096	1%
Pozo con bomba	4671	2%
Pozo sin bomba, jagüey	929	0%
Río, quebrada, manantial o nacimiento	10206	5%
	194033	

De estas tablas de frecuencias se puede concluir lo siguiente:

- En el Norte del Cauca, el 59% el tipo de vivienda cuenta con acceso acueducto, y 51% no cuenta respectivamente.
- En el Valle del Cauca, el 84 % tiene tipo de vivienda con acueducto, y el 16 % no.

De igual forma al analizar este tipo de vivienda respecto al acceso a la cocina, se obtienen las siguientes conclusiones.

- En el Norte del Cauca, el 52% no cuenta con tipo de vivienda con cocina, y 48% si cuenta respectivamente.
- En el Valle del Cauca, el 83% cuenta con tipo de vivienda con cocina, y 17% no cuenta respectivamente.

Otro punto importante por analizar es el acceso internet y/o computador, servicio importante en la actualidad.

- En el Norte del Cauca, el 99% no cuenta con PC, y 1% si cuenta respectivamente.
- En el Valle del Cauca, el 95% no cuenta con PC, y 5 % si cuenta respectivamente.
- En el Norte del Cauca, el 98% no cuenta con internet, y 2% si cuenta respectivamente.
- En el Valle del Cauca, el 85% no cuenta con internet, y 15 % si cuenta respectivamente.

Luego de analizar estas variables de ubicación del primer infante para los dos departamentos, se deciden analizar variables importantes como ingresos, gastos, educación de la persona responsable del primer infante.

Estas variables son importantes porque el modelo de cálculo del SISBEN las tiene en cuenta para hacer la predicción dentro del modelo econométrico utilizado.

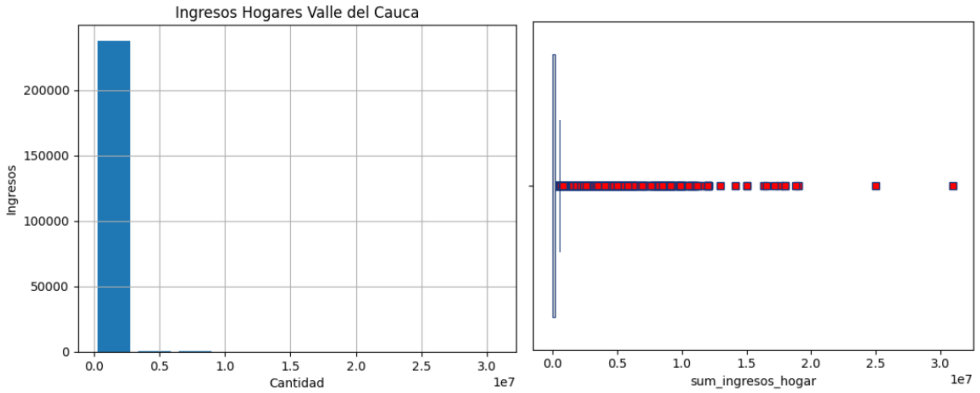
Análisis de los ingresos de los hogares

Departamento	Ingresos	Frecuencia
Valle del cauca	$ 14.412.141.821,00	96%
Norte del Cauca	$ 572.737.759,00	4%
	$ 14.984.879.580,00	

En la tabla de frecuencia se evidencia que el 96% de los ingresos hace referencia al Valle del Cauca, aunque también es importante mencionar que la variable tiene valores faltantes en un 68% de los registros, estos datos no se imputan debido a que son la mayoría de los registros de la base de datos y también debido a que es la variable clave para el Sisbén en el cálculo de las categorías de los hogares.

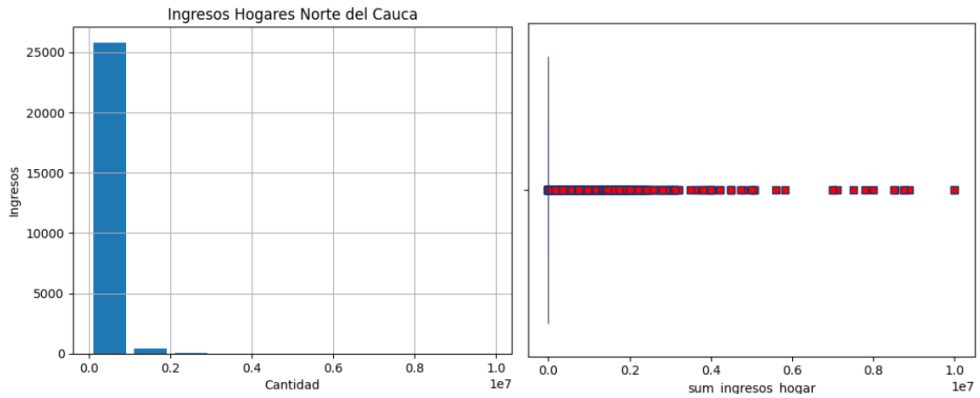
Ante esto es recomendable la utilización de un modelo donde no dependa de supuestos de normalidad.

Ingresos de los hogares de los municipios del Valle del Cauca.



Como se evidencia en la gráfica, los valores están sesgados a la izquierda de la gráfica, debido a que muchos registros están en 0, 68% de la data son datos faltantes.

Ingresos de los hogares de los municipios del norte del Cauca.



Como se evidencia en la gráfica, los valores están sesgados a la izquierda de la gráfica, debido a que muchos registros están en 0.

Análisis de los gastos de los hogares

Departamento	Gastos	Frecuencia
Valle del Cauca	$ 114.607.418.540,00	96%
Cauca	$ 5.112.820.973,00	4%
	$ 119.720.239.513,00	

En la tabla de frecuencia se evidencia que el 96% de los gastos hace referencia al Valle del Cauca, También es importante mencionar que la variable tiene valores faltantes en un 11% de los registros, muy inferior a los ingresos de cada departamento según la base datos.

En términos general hay una diferencia entre los ingresos y los gastos al analizar los dos departamentos.

Departamento	Gastos	Ingresos	Ingresos - Gastos
Valle del Cauca	$ 114.607.418.540,00	$ 14.412.141.821,00	-$ 100.195.276.719,00
Cauca	$ 5.112.820.973,00	$ 572.737.759,00	-$ 4.540.083.214,00
	$ 119.720.239.513,00	$ 14.984.879.580,00	-$ 104.735.359.933,00

El norte del Cauca ha enfrentado diversos desafíos económicos a lo largo de su historia. Esta área del departamento ha estado marcada por una combinación de factores que incluyen la presencia de conflictos armados, la falta de infraestructura adecuada, la limitada diversificación económica y la exclusión social.

Algunas de sus actividades económicas y/o puntos a tener en cuenta se centran en:

- Agricultura: La economía del norte del Cauca ha dependido tradicionalmente de la agricultura, con cultivos como la caña de azúcar, el café, el maíz y el plátano.
- Conflictos armados: Durante décadas, la región ha sido afectada por la presencia de grupos armados y el conflicto interno en Colombia. Históricamente se han presentado conflictos entre la guerrilla de las FARC y la fuerza pública en municipios ubicados sobre la cordillera Central como Caloto, Corinto, Miranda, Toribio y Florida.
- La Economía en su mayoría es informal: Muchos habitantes del norte del Cauca se han visto obligados a recurrir a la economía informal debido a la falta de oportunidades formales de empleo. Para el año 2023, la tasa de desocupación de Cauca fue 7,5% según datos del DANE, donde el País a corte 2024 la tasa de desempleo es del 11.3%.
- Programas de desarrollo: A pesar de los desafíos, se han implementado programas de desarrollo económico y social en la región, tanto por parte del gobierno como de organizaciones no gubernamentales y agencias internacionales. Algunos programas y/o políticas son:
    - Confacauca con el programa de atención integral de la niñez.
    - Consejos gremiales orientados por el ICBF [9]

Si bien la económica del municipio no está entre las más grandes del país, si es importante ya que aporta el 1.76 del PIB del País.

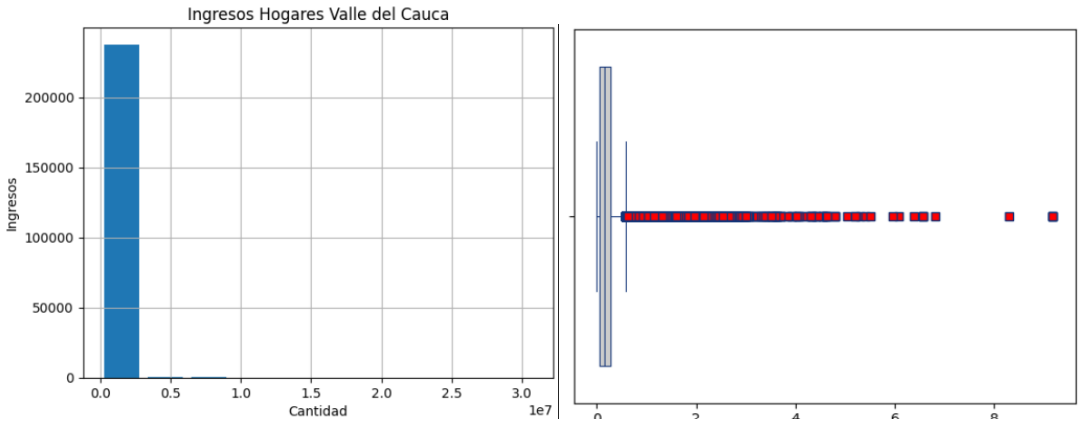
El Valle del Cauca, ubicado en el suroeste de Colombia, es una de las regiones económicamente más importantes del país. Su economía es diversificada y abarca varios sectores clave que contribuyen al desarrollo regional y nacional.

Algunos puntos importantes del valle del cauca.

- Agricultura y agroindustria: es una región agrícola productiva. Cultivos como la caña de azúcar, el café, el arroz, los plátanos, las frutas tropicales y los cultivos industriales son importantes para la economía local. estando la actividad agrícola dentro de las 3 principales actividades económicas de la región.
- Industria manufacturera: La región alberga una variedad de empresas en sectores como textiles, prendas de vestir, calzado, productos químicos, plásticos, metalurgia, muebles y productos farmacéuticos. La ciudad de Cali, la capital del departamento es un importante centro industrial y comercial, algunas empresas manufactureras son: Carvajal y Colombina, Chocorramo entre muchas otras.
- Infraestructura y transporte: El Valle del Cauca cuenta con una infraestructura de transporte desarrollada que incluye carreteras, ferrocarriles, puertos y aeropuertos. La región alberga el puerto de Buenaventura, el principal puerto marítimo de Colombia en la costa del Pacífico, que es crucial para el comercio internacional y la logística.
- En términos económicos el Valle del cauca es la tercera economía más grande del país, después de Bogotá y Antioquia. Representa el 9,7 por ciento del Producto Interno Bruto (PIB). Para el año 2022, la tasa de desempleo de Valle del Cauca fue 13,2% muy alto si lo comparamos contra el índice nacional para ese año del 10.3% [10]

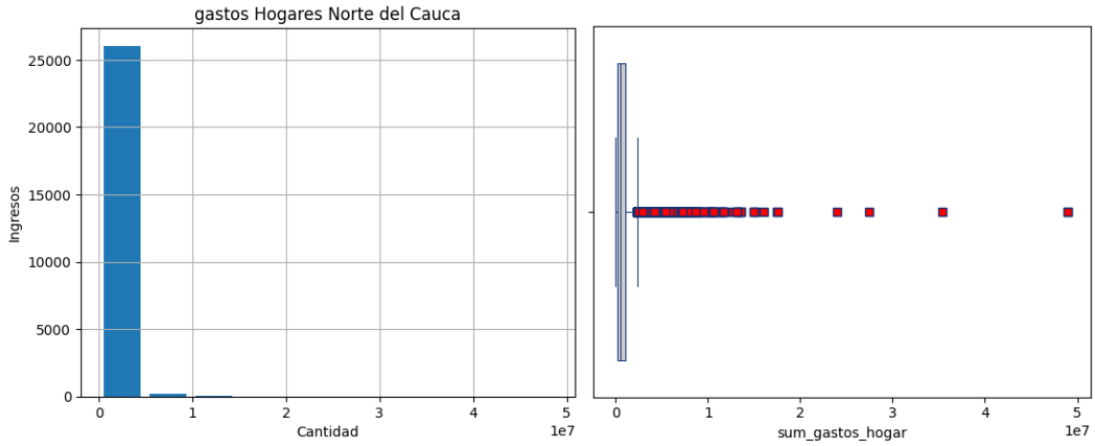
Si bien la económica del departamento no esta entre las más grandes del país, si esta entre las 3 grandes, aportando entre Bogotá, Antioquia y valle del Cauca el 50 % del PIB del País.

En términos estadísticos los gastos de los hogares de los municipios del Valle del Cauca.



Como se evidencia en la gráfica, los valores están sesgados a la izquierda de la gráfica, debido a que a la existencia de valores en 0; 11% de la data.

En términos estadísticos los gastos de los hogares de los municipios del Norte del Cauca.

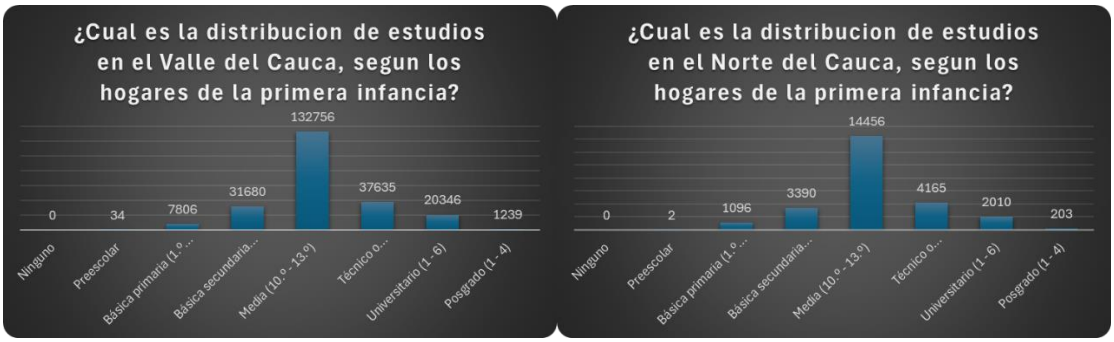


Ante este análisis se decide realiza modelos de clasificación, y comparar el modelo econométrico de regresión realizo por el Sisbén , donde la variable ingreso es fundamental para la clasificación de los hogares.

Se comparan los padres que viven en los hogares de ambos departamentos que cuenta con niños menores o iguales a 5 años.

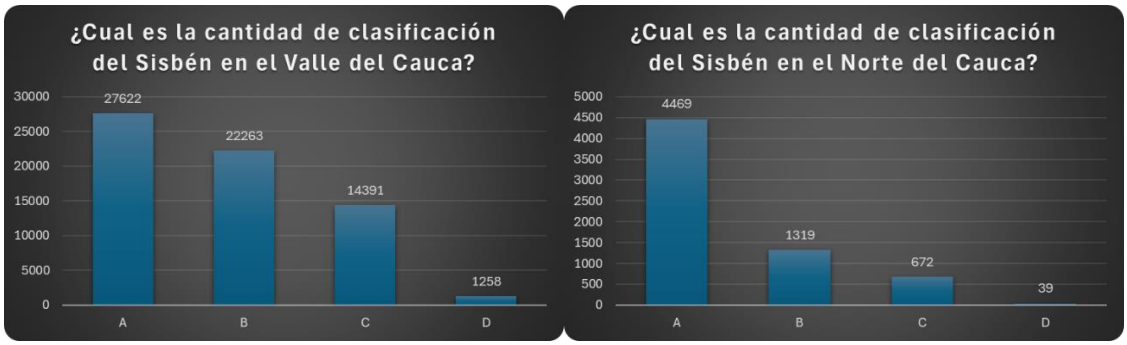
- En el Norte del Cauca, el 95% Los padres viven en los hogares con niños menores o iguales a 5 años., y 5% no respectivamente.
- En el Valle del Cauca, el 93 % Los padres viven en los hogares con niños menores o iguales a 5 años., y 7% no respectivamente.

Dato importante ya que como lo destaca la UNICEF en el artículo del 21 de Julio del 2021: un padre “aporta una mejor autoestima, más habilidades sociales, apoya con un mejor desempeño escolar, y entrega bienestar psicológico para los niños “. [11]



Como se evidencia en el histograma existen una concentración en las personas de ambos departamentos en la mitad de los datos (Media segundaria), esta variable cumple un supuesto de normalidad.

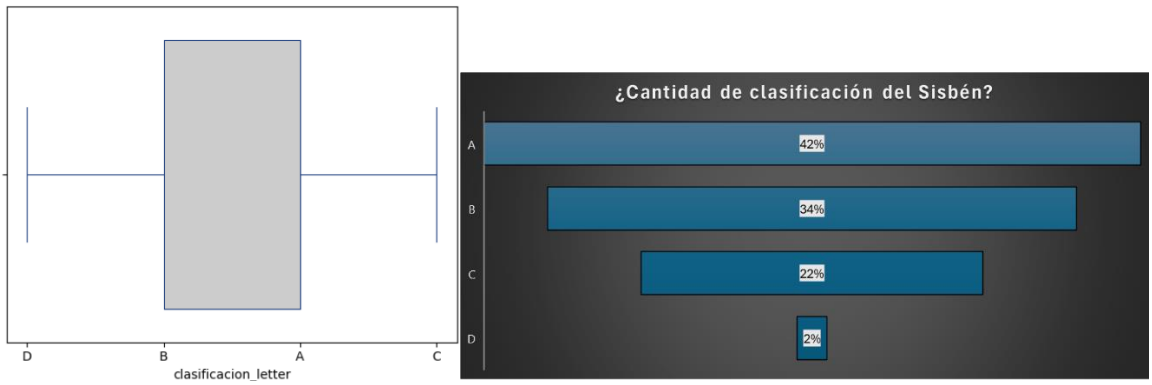
Por último se analiza la variable a predecir “clasificación”.



- 69% de los datos del Norte del Cauca tiene clasificación A.
- 20% de los datos del Norte del Cauca tiene clasificación B.
- 10% de los datos del Norte del Cauca tiene clasificación C.
- 10% de los datos del Norte del Cauca tiene clasificación D.
- 39% de los datos del Norte del Cauca tiene clasificación A.
- 35% de los datos del Norte del Cauca tiene clasificación B.
- 23% de los datos del Norte del Cauca tiene clasificación C.
- 2% de los datos del Norte del Cauca tiene clasificación D.

Esto representa que en el norte del Cauca existe mayor cantidad de hogares en pobreza monetaria que en el Valle del cauca.

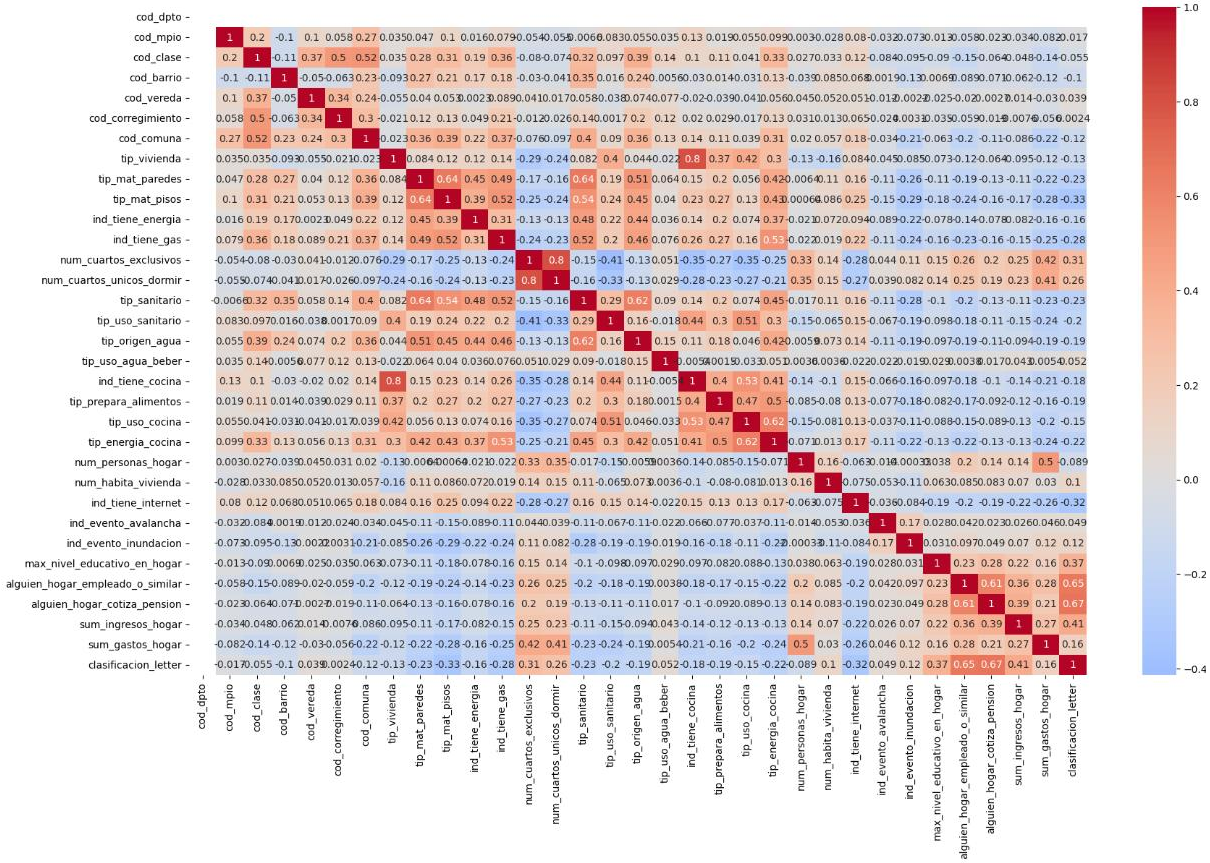
Al revisar el siguiente grafico se evidencia que la variable clasificación no tiene valores atípicos con una mayor concentración en la media de los datos, si analizamos los dos departamentos en conjunto.



Aplicando el coeficiente de curtosis, se evidencia una asimetría negativa o a la izquierda, pues la media es menor a la mediana y la mediana es menor menor que la moda. Además, vemos que tiene una cola pesada, esto nos sirve para determinar qué tipo de normalización aplicar. (Logarítmica)

- Curtosis: -0.843321

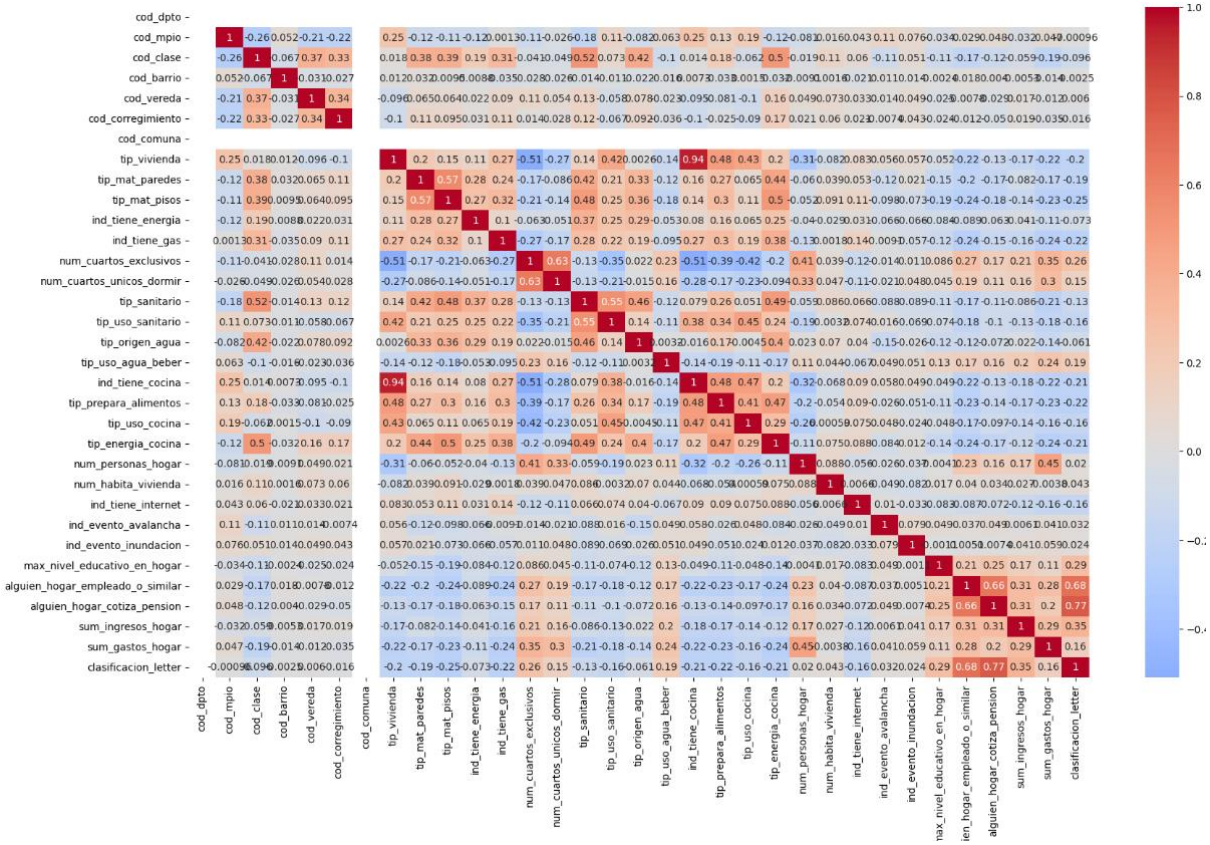
Se analiza la correlación de las variables para el Valle del Cauca.



Se analiza la correlación de las variables para el Valle del Cauca y se evidencia que tiene relación positiva en mayor medida con las variables:

- Gastos
- Ingresos.
- Nivel educativo
- Alguien en el hogar es empleado
- Numero de cuartos para dormir
- Numero de cuartos exclusivos para los niños.

Estas variables son representativas dentro del modelo ya que a medida que aumenta una aumenta podría aumentar la clasificación.



Se analiza la correlación de las variables para el norte del Cauca y se evidencia que tiene relación positiva en mayor medida con las variables:

- Gastos
- Ingresos en menor medida que para el Valle del cauca.
- Nivel educativo
- Alguien en el hogar es empleado
- Numero de cuartos para dormir
- Numero de cuartos exclusivos para los niños.
- Tipo de origen del agua, Importante para el Norte del Cauca, más que para el Valle del cauca.

Estas variables son representativas dentro del modelo ya que a medida que aumenta una aumenta podría aumentar la clasificación.

Bibliografía

1. "Repositorio Institucional ESAP," [En línea]. Disponible: https://repositoriocdim.esap.edu.co/handle/123456789/26328. [Accedido: 01-jun-2024].

2. "Defensoría del Pueblo advierte que disidencias de las FARC convirtieron a Toribío, Cauca, en centro de entrenamiento de niños reclutados," [En línea]. Disponible: https://www.defensoria.gov.co/-/defensor%C3%ADa-del-pueblo-advierte-que-disidencias-de-las-farc-convirtieron-a-torib%C3%ADo-cauca-en-centro-de-entrenamiento-de-ni%C3%B1os-reclutados. "Comisión de la Verdad," [Accedido: 01-jun-2024].

3. "Comisión de la Verdad," [En línea]. Disponible: https://web.comisiondelaverdad.co/especiales/norte-cordillera-cauca/index.html. [Accedido: 01-jun-2024].

4. "DANE," [En línea]. Disponible: https://www.dane.gov.co/files/investigaciones/planes-departamentos-ciudades/pres-ValledelCauca-18jul2023.pdf.

5. "Demografía de Cali," [En línea]. Disponible: https://www.cali.gov.co/informatica/publicaciones/106107/demografia_de_cali/. [Accedido: 01-jun-2024].

6. "DANE," [En línea]. Disponible: https://www.dane.gov.co/files/investigaciones/planes-departamentos-ciudades/pres-ValledelCauca-18jul2023.pdf.

7. "DANE," [En línea]. Disponible: https://www.dane.gov.co/files/investigaciones/planes-departamentos-ciudades/pres-ValledelCauca-18jul2023.pdf.

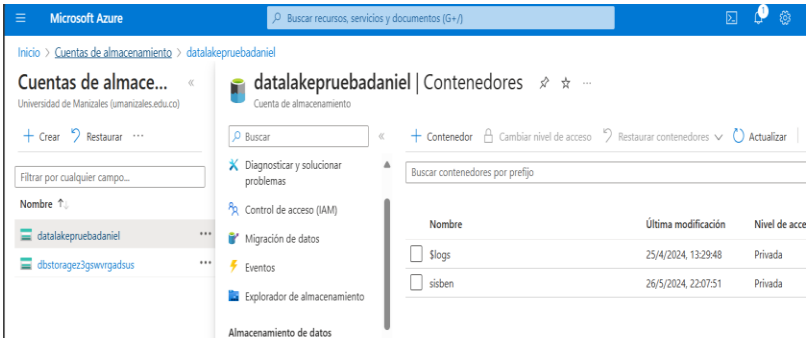
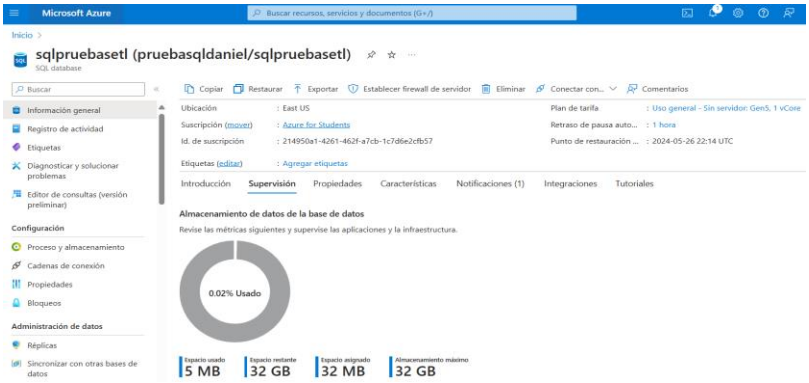
8. "Sisben IV: cuántas personas pobres hay inscritas," El Tiempo, [En línea]. Disponible: https://www.eltiempo.com/economia/sectores/sisben-iv-cuantas-personas-pobres-hay-inscritas-658684. [Accedido: 01-jun-2024].

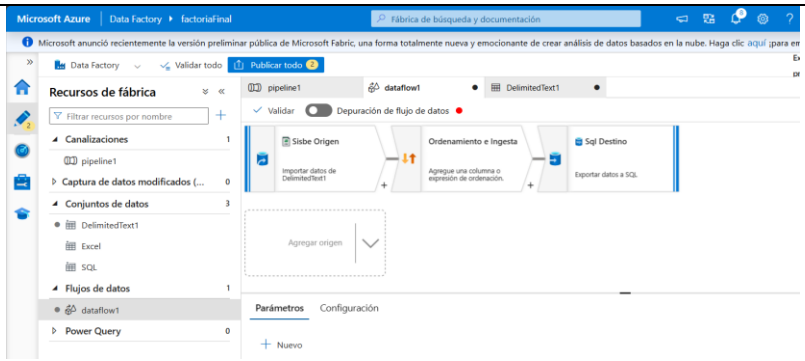
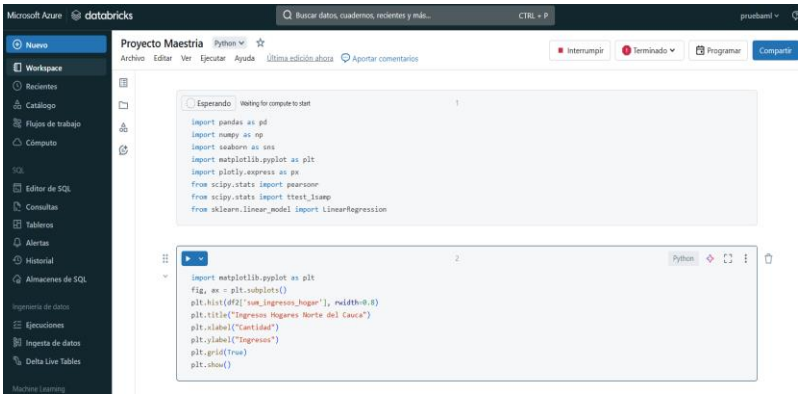
9. "DANE," [En línea]. Disponible: https://www.dane.gov.co/files/investigaciones/planes-departamentos-ciudades/pres-ValledelCauca-18jul2023.pdf.

10. "DANE," [En línea]. Disponible: https://www.dane.gov.co/files/investigaciones/planes-departamentos-ciudades/pres-ValledelCauca-18jul2023.pdf.

11. "DANE," [En línea]. Disponible: https://www.dane.gov.co/files/investigaciones/planes-departamentos-ciudades/pres-ValledelCauca-18jul2023.pdf.

Anexo 3: Herramienta de Análisis descriptivo

Nombre	Descripción
Azure- DataLake	https://azure.microsoft.com/es-es/solutions/data-lake Utilizada para almacenar la base de datos de 1.5 Gb del Sisbén y guardarla en un repositorio central. 
Azure – DataFactory	https://azure.microsoft.com/es-es/products/data-factory Utiliza para hacer la ETL entre el datalake y la base de datos Azure SQL, y así pasar de un .csv a una base de datos relacional. 
Azure- SQL	https://azure.microsoft.com/es-es/products/azure-sql/database utilizada para hacer consultas SQL, y poder ir describiendo los datos. Trabajando con SQL SERVER MANAGEMENT.

	
Azure databricks.	https://azure.microsoft.com/es-es/products/databricks Hacer un pipeline en Python para poder generar estadística descriptiva y correlación de variables, utilizando la librería numpy. 

Anexo 4

multicolinealidad

June 7, 2024

[4]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import accuracy_score, confusion_matrix,␣
 ↪classification_report
import matplotlib.pyplot as plt
```

[7]:
```
path = 'dataset/datasetFinalTesisCaliCauca.csv'
data = pd.read_csv(path)

#df = data.dropna(subset=['clasificacion_letter'])
#76 , 19
df = data.loc[data['cod_dpto'] == 76].dropna(subset=['clasificacion_letter'])
#df = data.dropna(subset=['clasificacion_letter'])

#print(df.head())

label_mapping = {'A': 0, 'B': 1, 'C': 2, 'D': 3}

df['clasificacion_letter_numeric'] = df['clasificacion_letter'].
 ↪map(label_mapping)

#'ind_evento_avalancha','ind_evento_inundacion'
#'cod_dpto',
X = df[
    ['edad_calculada', 'cod_mpio', 'cod_clase', 'cod_barrio', 'cod_vereda',␣
 ↪'cod_corregimiento', 'cod_comuna', 'tip_vivienda',
     'tip_mat_paredes', 'tip_mat_pisos', 'ind_tiene_energia', 'ind_tiene_gas',␣
 ↪'num_cuartos_exclusivos',
     'num_cuartos_unicos_dormir', 'tip_sanitario', 'tip_uso_sanitario',␣
 ↪'tip_origen_agua', 'tip_uso_agua_beber',
     'ind_tiene_cocina', 'tip_prepara_alimentos', 'tip_uso_cocina',␣
 ↪'tip_energia_cocina',
      'num_personas_hogar', 'num_habita_vivienda', 'ind_tiene_internet',␣
 ↪'ind_evento_avalancha','ind_evento_inundacion',
     'max_nivel_educativo_en_hogar', 'alguien_hogar_empleado_o_similar',␣
 ↪'alguien_hogar_cotiza_pension',
```

```
     'sum_ingresos_hogar', 'sum_gastos_hogar']]
y = df['clasificacion_letter_numeric']
```

[8]:
```
def calculateVIF(var_predictoras_df):
    var_pred_labels = list(var_predictoras_df.columns)
    num_var_pred = len(var_pred_labels)

    lr_model = LinearRegression()

    result = pd.DataFrame(index = ['VIF'], columns = var_pred_labels)
    result = result.fillna(0)

    for ite in range(num_var_pred):
        x_features = var_pred_labels[:]
        y_feature = var_pred_labels[ite]
        x_features.remove(y_feature)

        x = var_predictoras_df[x_features]
        y = var_predictoras_df[y_feature]

        lr_model.fit(var_predictoras_df[x_features],␣
 ↪var_predictoras_df[y_feature])

        result[y_feature] = 1/(1 - lr_model.
 ↪score(var_predictoras_df[x_features], var_predictoras_df[y_feature]))

    return result
```

[9]:
```
calculateVIF(X.copy(deep = True)).T
```

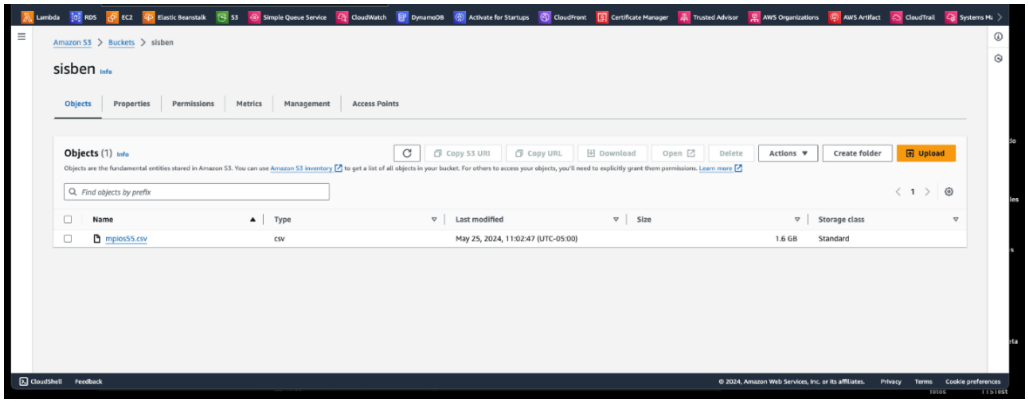
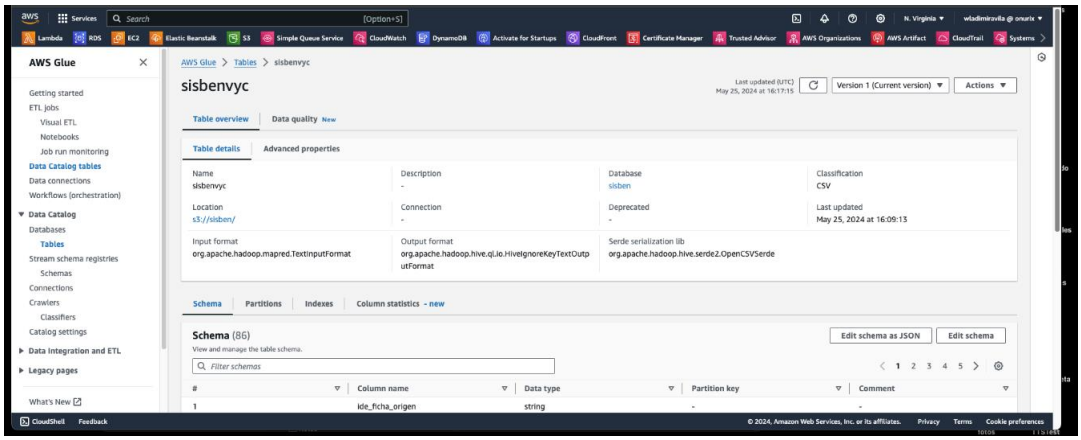
[9]:

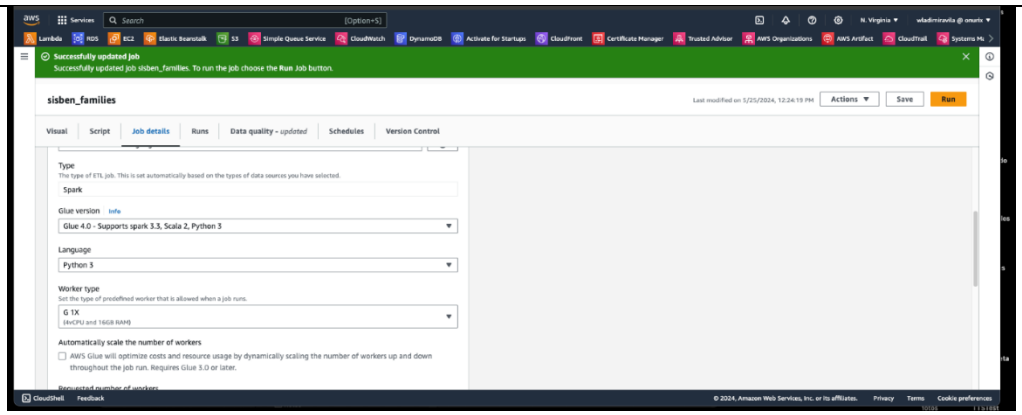
	VIF
edad_calculada	1.013821
cod_mpio	1.165248
cod_clase	2.166764
cod_barrio	1.372820
cod_vereda	1.240604
cod_corregimiento	1.418085
cod_comuna	1.884801
tip_vivienda	3.165715
tip_mat_paredes	2.286888
tip_mat_pisos	2.166847
ind_tiene_energia	1.491478
ind_tiene_gas	1.916168
num_cuartos_exclusivos	3.266729
num_cuartos_unicos_dormir	2.963700
tip_sanitario	2.655928
tip_uso_sanitario	1.776105

```
tip_origen_agua                   1.953729
tip_uso_agua_beber                1.052915
ind_tiene_cocina                  3.688940
tip_prepara_alimentos             1.528710
tip_uso_cocina                    2.952534
tip_energia_cocina                3.276826
num_personas_hogar                1.520326
num_habita_vivienda               1.122757
ind_tiene_internet                1.209695
ind_evento_avalancha              1.053126
ind_evento_inundacion             1.196135
max_nivel_educativo_en_hogar      1.150728
alguien_hogar_empleado_o_similar  1.772972
alguien_hogar_cotiza_pension      1.720489
sum_ingresos_hogar                1.293861
sum_gastos_hogar                  1.732664
```

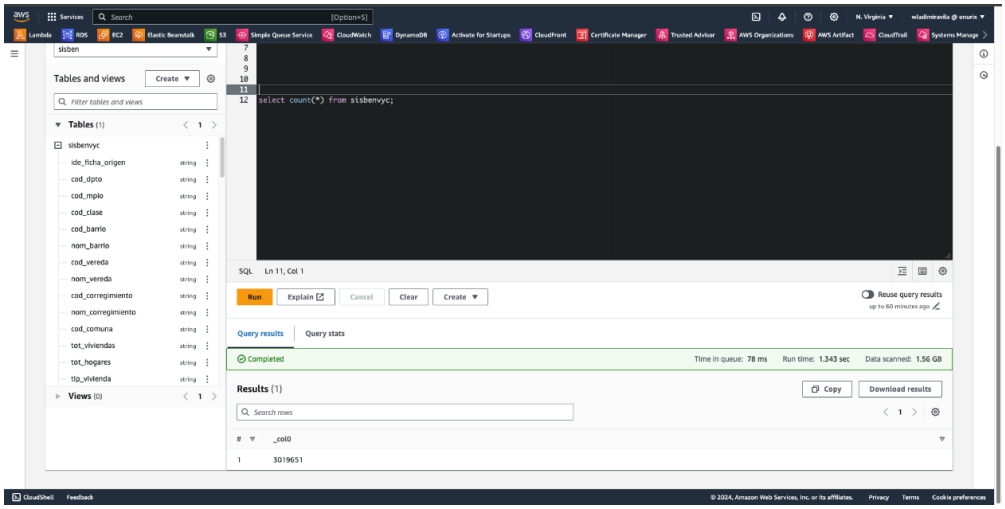
[ ]:

Anexo 5

Nombre	Descripción
Amazon S3	https://aws.amazon.com/s3/ Los datos se subieron a un bucket de Amazon S3, con un tamaño aproximado de 1,6 GB. 
AWS Glue	https://aws.amazon.com/es/glue/ Se creó el esquema de la tabla basado en el archivo CSV usando AWS Glue.  Se utilizó AWS Glue, que emplea Apache Spark para realizar transformaciones de datos personalizadas con Python, eliminando columnas irrelevantes y manteniendo solo aquellas que aportan al modelo predictivo.

<table>
<tr><td></td><td> 
</td></tr>
<tr><td>Amazon Athena</td><td>https://aws.amazon.com/es/athena/

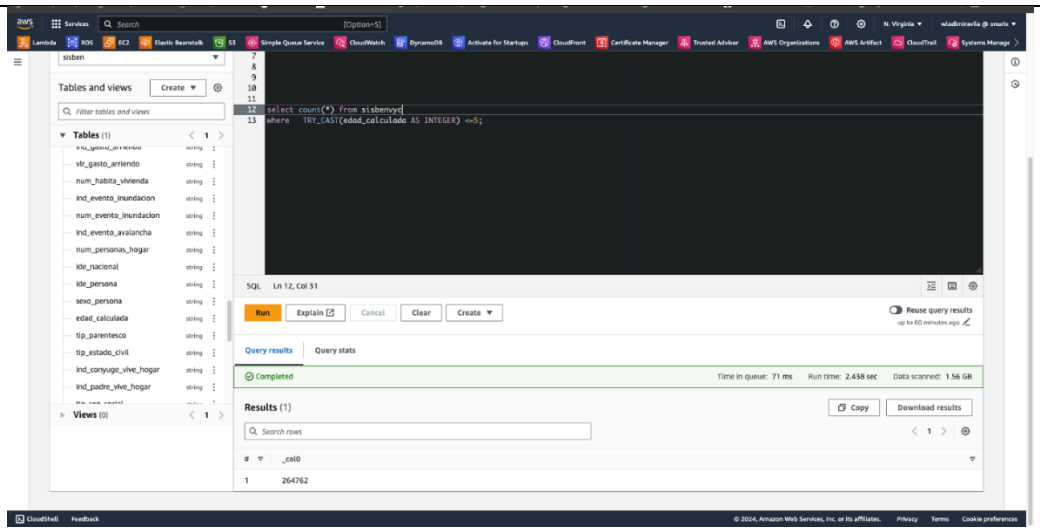
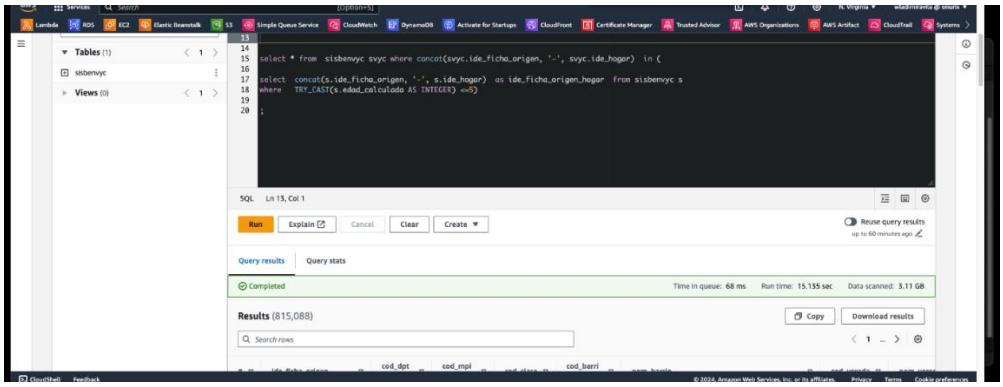
Se utilizó Amazon Athena para contar el total de registros existentes (3.019.651 registros) y filtrar los datos relevantes. Athena utiliza Presto como motor SQL para consultas sobre archivos planos.

Filtrado y Generación de Conjuntos de Datos

Se filtraron los hogares con al

menos un niño de hasta 5 años (264.764 niños).</td></tr>
</table>

	 Se generó un nuevo conjunto de datos con todos los miembros del hogar que tienen al menos un menor de 5 años, resultando en 815.088 registros. 
Presto	https://aws.amazon.com/es/emr/features/presto Presto es un motor de consultas SQL distribuidas de código abierto optimizado para el análisis de datos ad hoc de baja latencia. Es compatible con el estándar ANSI SQL, incluidas las consultas complejas, las agregaciones, las uniones y las funciones de ventana.
Apache Spark	https://aws.amazon.com/es/what-is/apache-spark/ Apache Spark es un sistema de procesamiento distribuido de código abierto que se utiliza para cargas de trabajo de macrodatos. Utiliza el almacenamiento en caché de memoria y una ejecución de consulta optimizada para consultas rápidas de análisis de cualquier tamaño.

Anexo 6

lack_public_services

June 7, 2024

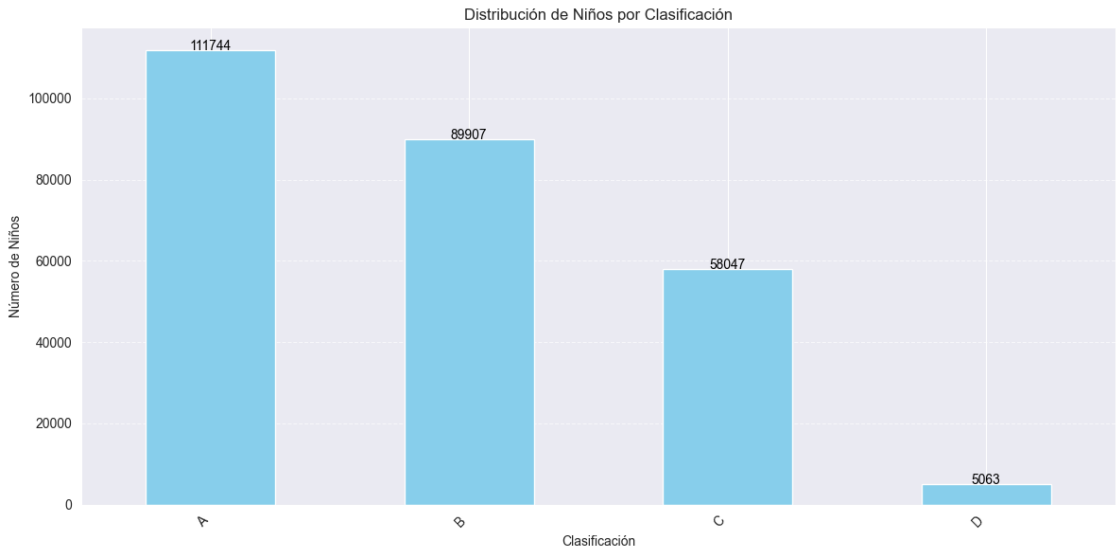
[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

[2]:
```
children_under_5 = pd.read_csv("dataset/datasetFinalTesisCaliCauca.csv")
```

[4]:
```
def plot_distribution(data, label, xxlabel, yylabel, title):
    """
    Generates a bar chart using Matplotlib.

    Parameters:
    data (list or array): Data values for the chart.
    label (str): Label for the data series.
    xlabel (str): Label for the x-axis.
    ylabel (str): Label for the y-axis.
    title (str): Title for the chart.

    Returns:
    None. (The chart is displayed directly.)
    """

    distribution_data = data #distribution_kids_inadequate_sanitary_conditions.
↪values
    category_labels = label # distribution_kids_inadequate_sanitary_conditions.
↪index.to_numpy()

    # Create a bar chart
    plt.figure(figsize=(8, 6))  # Adjust figure size as desired
    plt.bar(category_labels, distribution_data, color='skyblue')
    plt.xlabel(xxlabel)
    plt.ylabel(yylabel)
    plt.title(title)

    # Rotate x-axis labels for better readability if many categories
    plt.xticks(rotation=45, ha='right')  # Optional: Rotate x-axis labels

    plt.grid(axis='y', linestyle='--', alpha=0.7)  # Add subtle gridlines
```

```
    plt.tight_layout()  # Adjust spacing between elements
    plt.show()
```

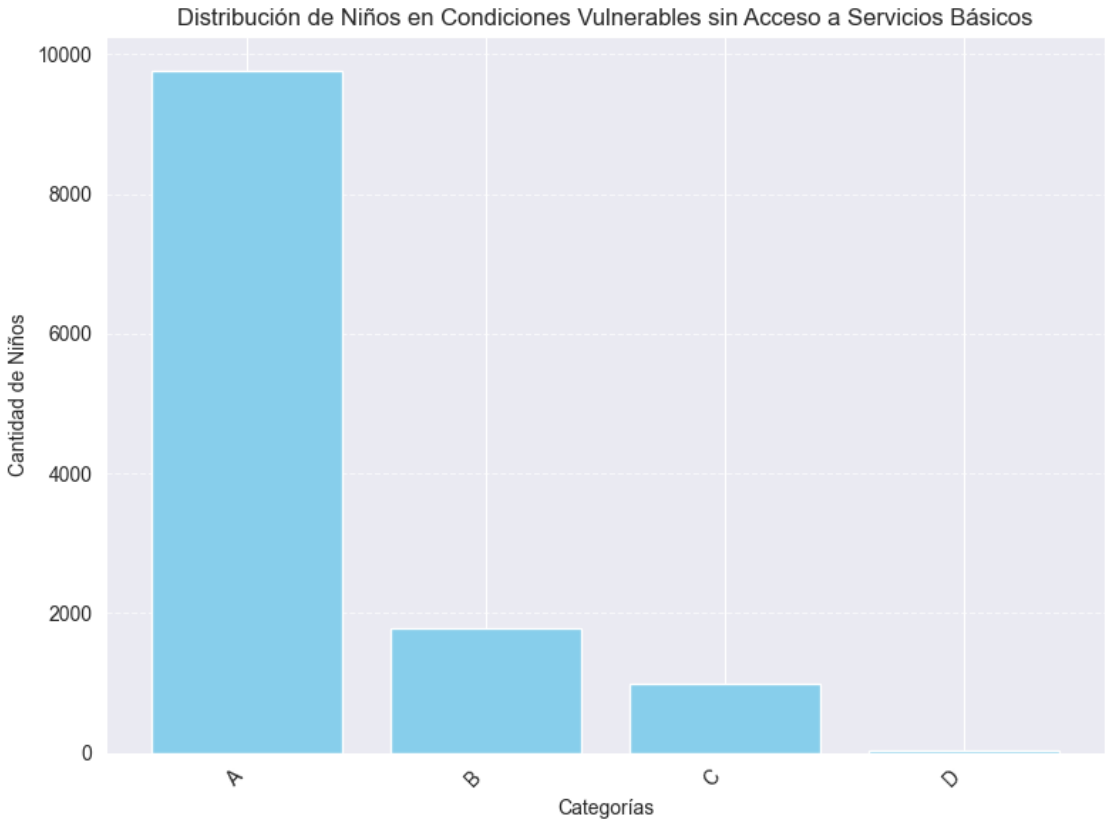
[5]:
```
plt.figure(figsize=(12, 6))  # Adjust figure size as desired
children_under_5['clasificacion_letter'].value_counts().plot(kind='bar',␣
 ↪color='skyblue')
plt.xticks(rotation=45, ha='right')  # Rotate x-axis labels for better␣
 ↪readability

# Add labels above each bar
for p in plt.gca().patches:
    plt.text(p.get_x() + p.get_width() / 2, p.get_height() + 0.1, f'{int(p.
 ↪get_height())}', ha='center', color='black')

# Customize labels and title
plt.xlabel('Clasificación')
plt.ylabel('Número de Niños')
plt.title('Distribución de Niños por Clasificación')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```



[6]:
```
###Public Services
no_services = (children_under_5['ind_tiene_energia'] == 2) & \
              (children_under_5['ind_tiene_alcantarillado'] == 2) & \
              (children_under_5['ind_tiene_gas'] == 2) & \
              (children_under_5['ind_tiene_acueducto'] == 2)
```

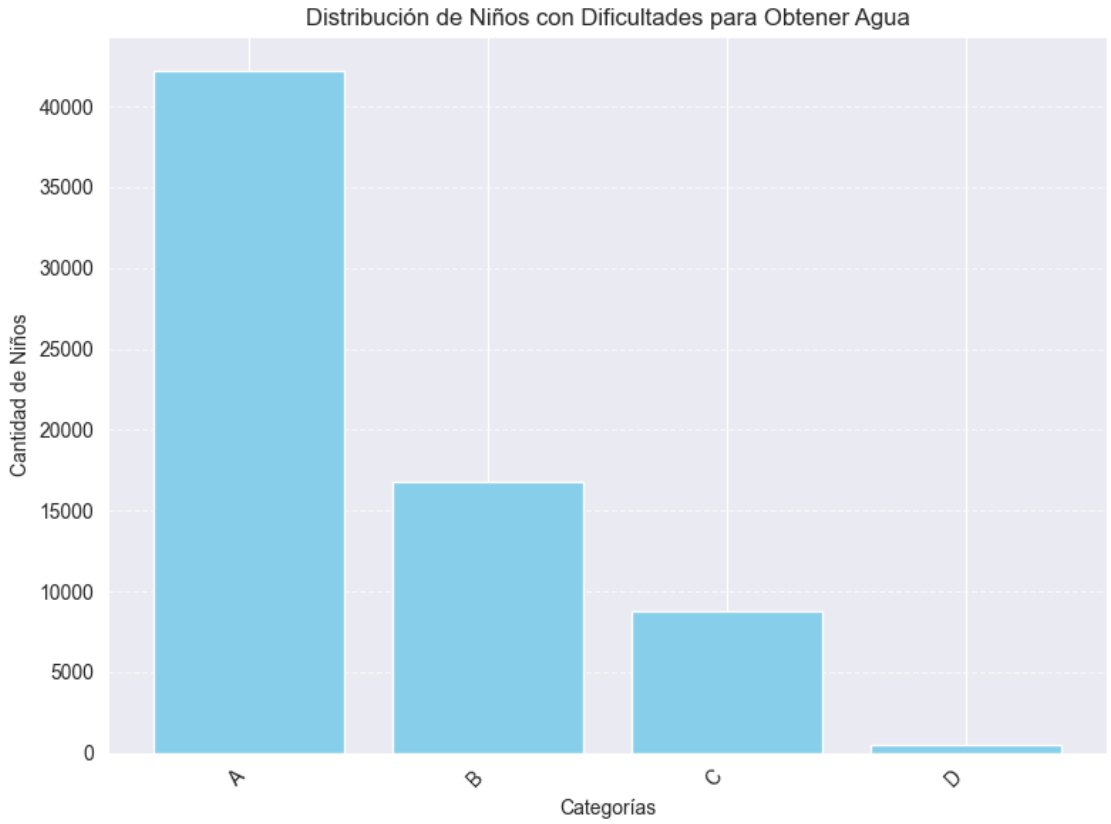
```
children_under_5["no_services"] = no_services
kids_lack_public_services = children_under_5[children_under_5['no_services']]
distribution_lack_services = kids_lack_public_services['clasificacion_letter'].
 ↪value_counts()


##Dificulties to get water

difficulties_obtaining_water = (children_under_5['ind_agua_llega_7dias'] == 2␣
 ↪)|  (children_under_5['ind_agua_llega_7dias'] == 9)

children_under_5["difficulties_obtaining_water"] = difficulties_obtaining_water
kids_difficulties_obtaining_water =␣
 ↪children_under_5[children_under_5['difficulties_obtaining_water']]

distribution_difficulties_obtaining_water=␣
 ↪kids_difficulties_obtaining_water['clasificacion_letter'].value_counts()



##Inadequate sanitary conditions

inadequate_sanitary_conditions = (children_under_5['tip_sanitario'].
 ↪isin([3,4,5])) | \
                                 (children_under_5['tip_ubi_sanitario'] == 2)

children_under_5["inadequate_sanitary_conditions"] =␣
 ↪inadequate_sanitary_conditions

kids_inadequate_sanitary_conditions =␣
 ↪children_under_5[children_under_5['inadequate_sanitary_conditions']]

distribution_kids_inadequate_sanitary_conditions=␣
 ↪kids_inadequate_sanitary_conditions['clasificacion_letter'].value_counts()
```

[7]:
```
plot_distribution(distribution_lack_services.values, distribution_lack_services.
 ↪index.to_numpy(), "Categorías", "Cantidad de Niños", "Distribución de Niños␣
 ↪en Condiciones Vulnerables sin Acceso a Servicios Básicos")


print(distribution_lack_services)
```



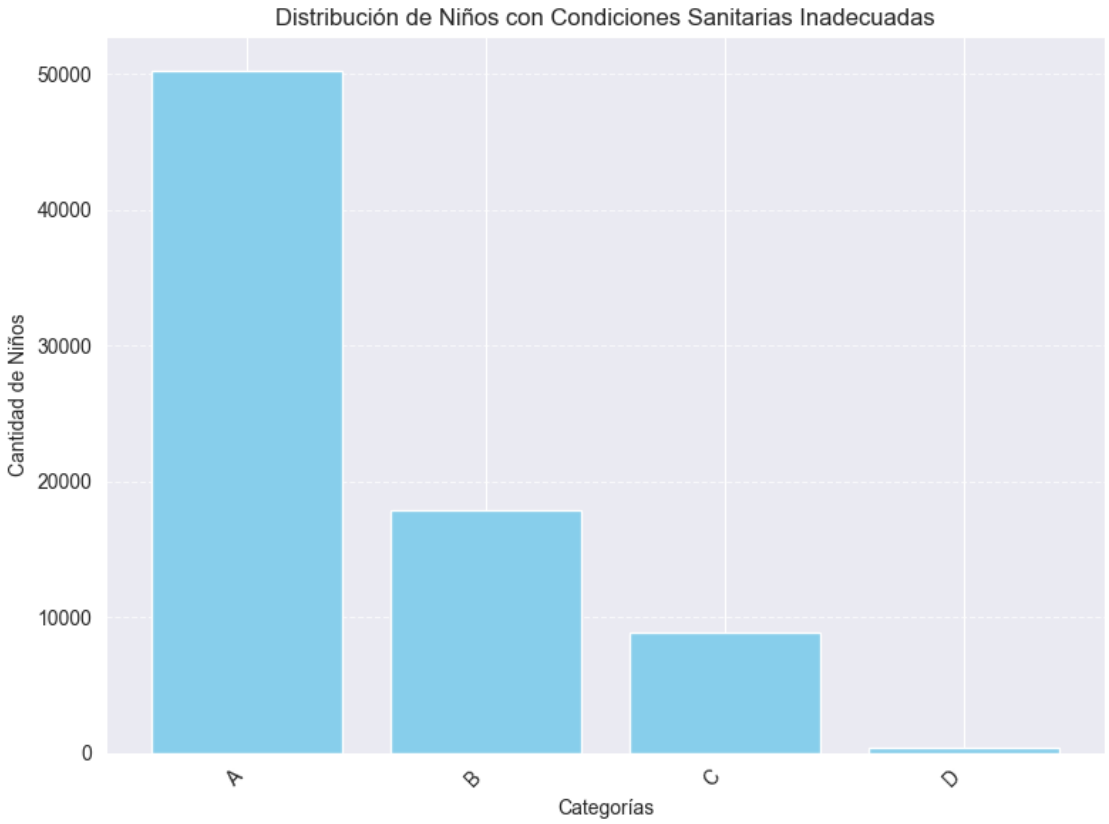

```
A    9756
B    1785
C     990
D      37
Name: clasificacion_letter, dtype: int64
```

[8]:
```
plot_distribution(distribution_difficulties_obtaining_water.values,␣
↪distribution_difficulties_obtaining_water.index.to_numpy(), "Categorías",␣
↪"Cantidad de Niños", "Distribución de Niños con Dificultades para Obtener␣
↪Agua")


print(distribution_difficulties_obtaining_water)
```




```
A    42176
B    16772
C     8740
D      534
Name: clasificacion_letter, dtype: int64
```

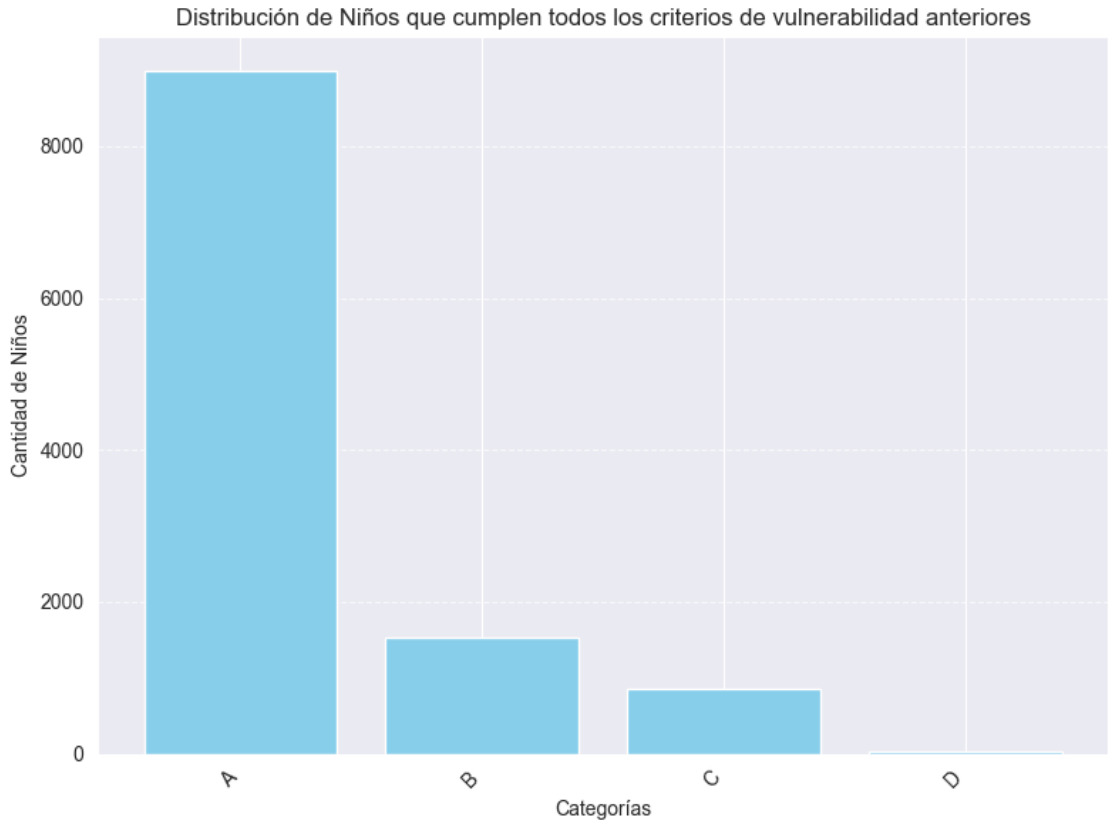
[9]:
```
plot_distribution(distribution_kids_inadequate_sanitary_conditions.values,␣
↪distribution_kids_inadequate_sanitary_conditions.index.to_numpy(),␣
↪"Categorías", "Cantidad de Niños", "Distribución de Niños con Condiciones␣
↪Sanitarias Inadecuadas")


print(distribution_kids_inadequate_sanitary_conditions)
```




```
A    50207
B    17884
C     8915
D      335
Name: clasificacion_letter, dtype: int64
```

[10]:
```
children_under_5['children_vulnerable'] = np.where(
    (no_services & difficulties_obtaining_water &␣
  ↪inadequate_sanitary_conditions),
    True,
    False
)


children_vulnerable_conditions =␣
  ↪children_under_5[children_under_5['children_vulnerable']]
distribution_children_vulnerable_conditions =␣
  ↪children_vulnerable_conditions['clasificacion_letter'].value_counts()
```

```
plot_distribution(distribution_children_vulnerable_conditions.values,␣
  ↪distribution_children_vulnerable_conditions.index.to_numpy(), "Categorías",␣
  ↪"Cantidad de Niños", "Distribución de Niños que cumplen todos los criterios␣
  ↪de vulnerabilidad anteriores")


print(distribution_children_vulnerable_conditions)
```




```
A    8984
B    1523
C     861
D      24
Name: clasificacion_letter, dtype: int64
```

[ ]:

Anexo 7

In [1]:
```
!pip install statsmodels
```

```
Requirement already satisfied: statsmodels in /usr/local/lib/python3.11/site-packages (0.14.2)
Requirement already satisfied: numpy>=1.22.3 in /usr/local/lib/python3.11/site-packages (from statsmodels) (1.26.2)
Requirement already satisfied: scipy!=1.9.2,>=1.8 in /usr/local/lib/python3.11/site-packages (from statsmodels) (1.11.4)
Requirement already satisfied: pandas!=2.1.0,>=1.4 in /usr/local/lib/python3.11/site-packages (from statsmodels) (1.5.3)
Requirement already satisfied: patsy>=0.5.6 in /usr/local/lib/python3.11/site-packages (from statsmodels) (0.5.6)
Requirement already satisfied: packaging>=21.3 in /usr/local/lib/python3.11/site-packages (from statsmodels) (23.2)
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.11/site-packages (from pandas!=2.1.0,>=1.4->statsmodels)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/site-packages (from pandas!=2.1.0,>=1.4->statsmodels) (2023.3.pos
t1)
Requirement already satisfied: six in /usr/local/lib/python3.11/site-packages (from patsy>=0.5.6->statsmodels) (1.16.0)

[notice] A new release of pip is available: 23.3.2 -> 24.0
[notice] To update, run: python3.11 -m pip install --upgrade pip
```

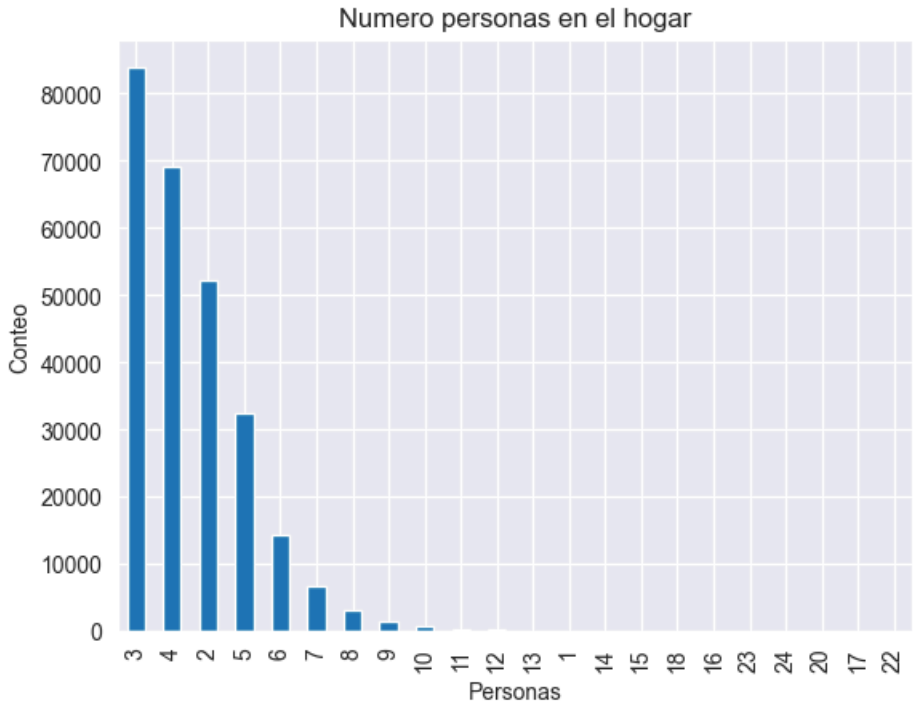
In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import mpl_toolkits
%matplotlib inline
from scipy.stats import pearsonr
import statsmodels.api as sm
```

In [6]:
```
data = pd.read_csv("dataset/datasetFinalTesisCaliCauca.csv")
```

In [1]:
```
#data.describe()
```

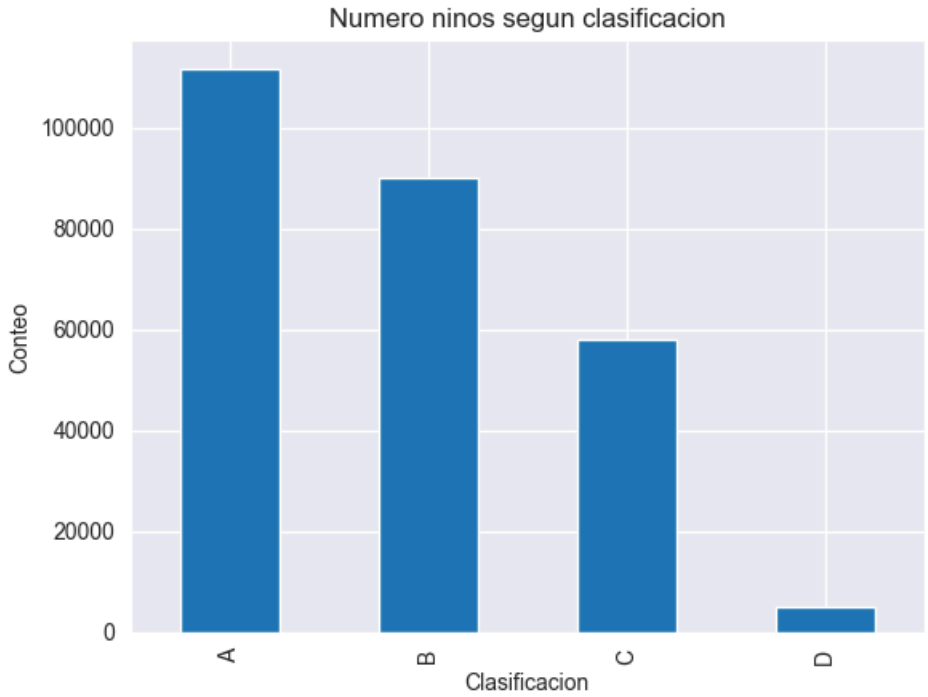
In [7]:
```
data['num_personas_hogar'].value_counts().plot(kind='bar')
plt.title('Numero personas en el hogar')
plt.xlabel('Personas')
plt.ylabel('Conteo')
sns.despine
```

Out[7]:
```
<function seaborn.utils.despine(fig=None, ax=None, top=True, right=True, left=False, bottom=False, offset=None, trim=False)>
```



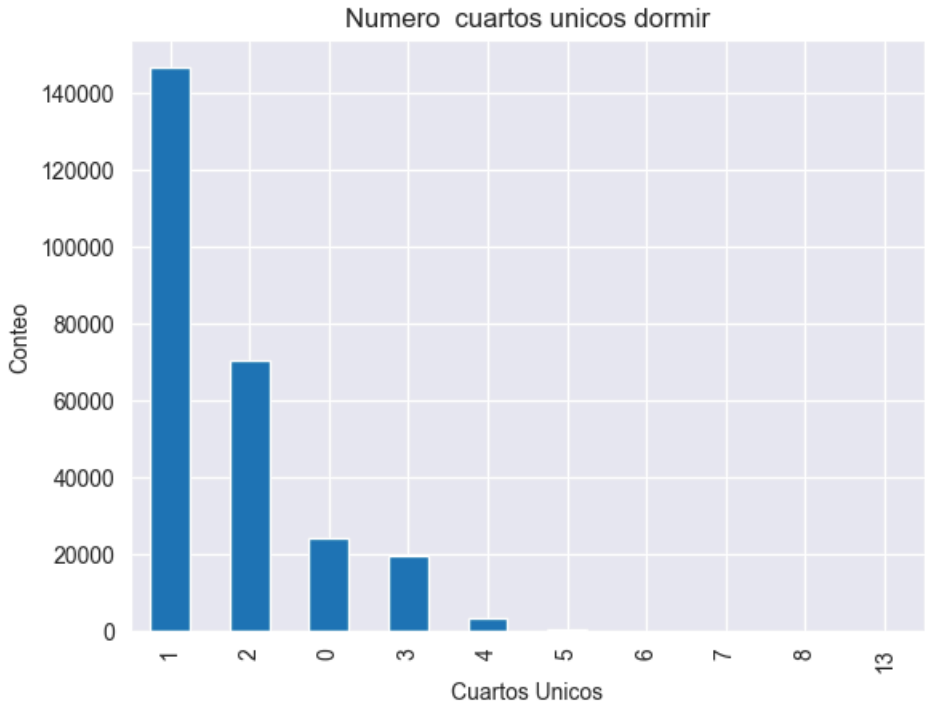
In [9]:
```
data['clasificacion_letter'].value_counts().plot(kind='bar')
plt.title('Numero ninos segun clasificacion')
plt.xlabel('Clasificacion')
plt.ylabel('Conteo')
sns.despine
```

Out[9]:
```
<function seaborn.utils.despine(fig=None, ax=None, top=True, right=True, left=False, bottom=False, offset=None, trim=False)>
```



In [10]:
```
data['num_cuartos_unicos_dormir'].value_counts().plot(kind='bar')
plt.title('Numero  cuartos unicos dormir')
plt.xlabel('Cuartos Unicos')
plt.ylabel('Conteo')
sns.despine
```

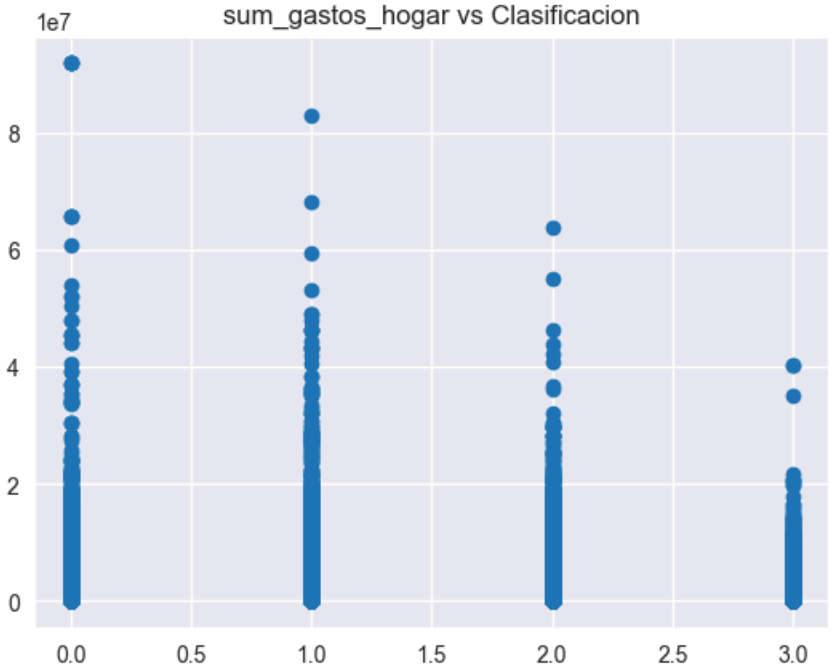
Out[10]:
```
<function seaborn.utils.despine(fig=None, ax=None, top=True, right=True, left=False, bottom=False, offset=None, trim=False)>
```



In [ ]:

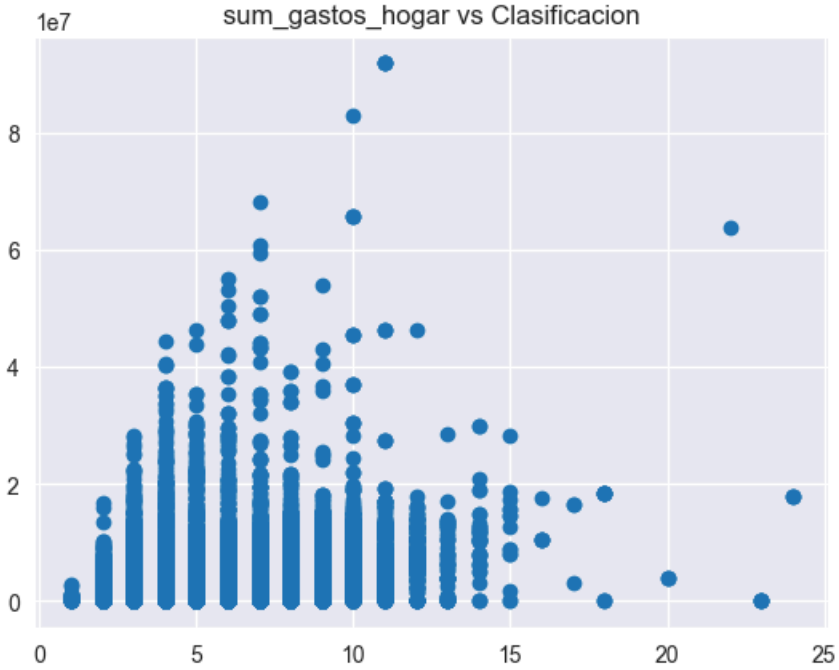
In [14]:
```
label_mapping = {'A': 0, 'B': 1, 'C': 2, 'D': 3}

data['clasificacion_letter_numeric'] = data['clasificacion_letter'].map(label_mapping)

plt.scatter(data.clasificacion_letter_numeric,data.sum_gastos_hogar)
plt.title("sum_gastos_hogar vs Clasificacion")
```

Out[14]:
```
Text(0.5, 1.0, 'sum_gastos_hogar vs Clasificacion')
```



In [15]:
```
plt.scatter(data.num_personas_hogar,data.sum_gastos_hogar)
plt.title("sum_gastos_hogar vs Clasificacion")
```

Out[15]:
```
Text(0.5, 1.0, 'sum_gastos_hogar vs Clasificacion')
```

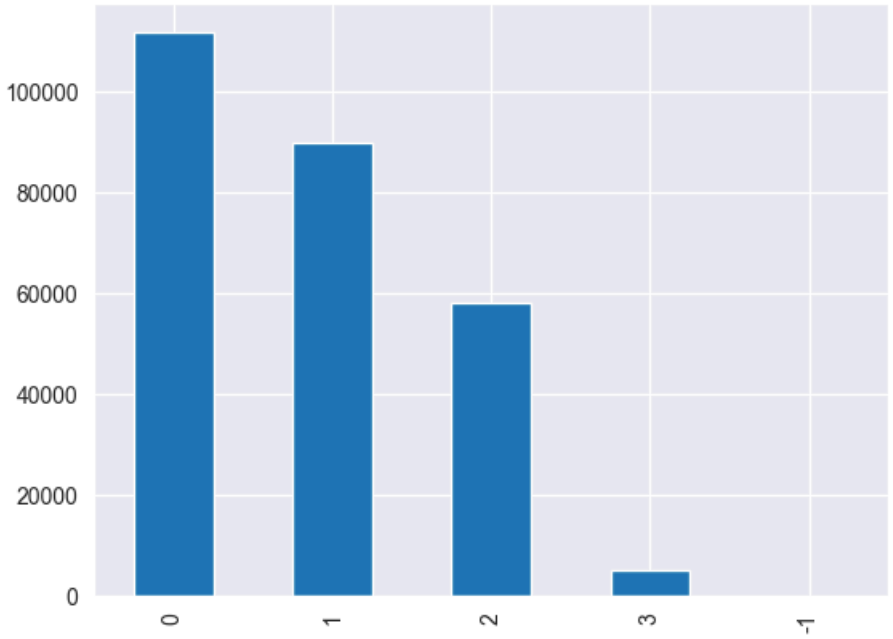


In [ ]:

In [17]:
```
if data['clasificacion_letter'].dtype == 'object':
    data['clasificacion_letter_cate'] = data['clasificacion_letter'].astype('category').cat.codes
```

In [18]:
```
data['clasificacion_letter_cate'].value_counts().plot(kind='bar')
```

Out[18]: <Axes: >



In [35]:
```
selected_features = [
    'sum_gastos_hogar', 'max_nivel_educativo_en_hogar',
    'alguien_hogar_empleado_o_similar', 'tot_hogares', 'num_cuartos_vivienda',
    'ind_tiene_energia', 'ind_tiene_alcantarillado', 'ind_tiene_gas', 'ind_tiene_acueducto',
    'clasificacion_letter_cate', 'cod_mpio' , 'num_personas_hogar'
]

data_correlation = data[selected_features]

correlation_matrix = data_correlation.corr()

print("Matriz de Correlación:")
print(correlation_matrix)

plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Matriz de Correlación')
plt.show()
```

Matriz de Correlación:

	sum_gastos_hogar	max_nivel_educativo_en_hogar	alguien_hogar_empleado_o_similar	tot_hogares	num_cuartos_vivienda	ind_tiene_energia	ind_tiene_alcantarillado	ind_tiene_gas	ind_tiene_acueducto	clasificacion_letter_cate	cod_mpio	num_personas_hogar
sum_gastos_hogar	1.000000	0.151282	0.292638	-0.098390	0.270021	-0.151840	-0.247196	-0.282686	-0.229032	0.177611	0.160707	0.498893
max_nivel_educativo_en_hogar	0.151282	1.000000	0.227704	-0.016901	0.129426	-0.078723	-0.104244	-0.147714	-0.105330	0.356183	0.003728	0.034175
alguien_hogar_empleado_o_similar	0.292638	0.227704	1.000000	0.005942	0.243927	-0.132541	-0.227250	-0.253425	-0.210644	0.660969	0.123474	0.212869
tot_hogares	-0.098390	-0.016901	0.005942	1.000000	0.421896	-0.064520	-0.093104	-0.126507				
num_cuartos_vivienda	0.270021	0.129426	0.243927	0.421896	1.000000	-0.201270	-0.270571	-0.405859				

ind_tiene_acueducto	-0.097808	-0.255069
clasificacion_letter_cate	0.003851	0.278434
cod_mpio	0.060938	0.181070
num_personas_hogar	-0.135322	0.170233

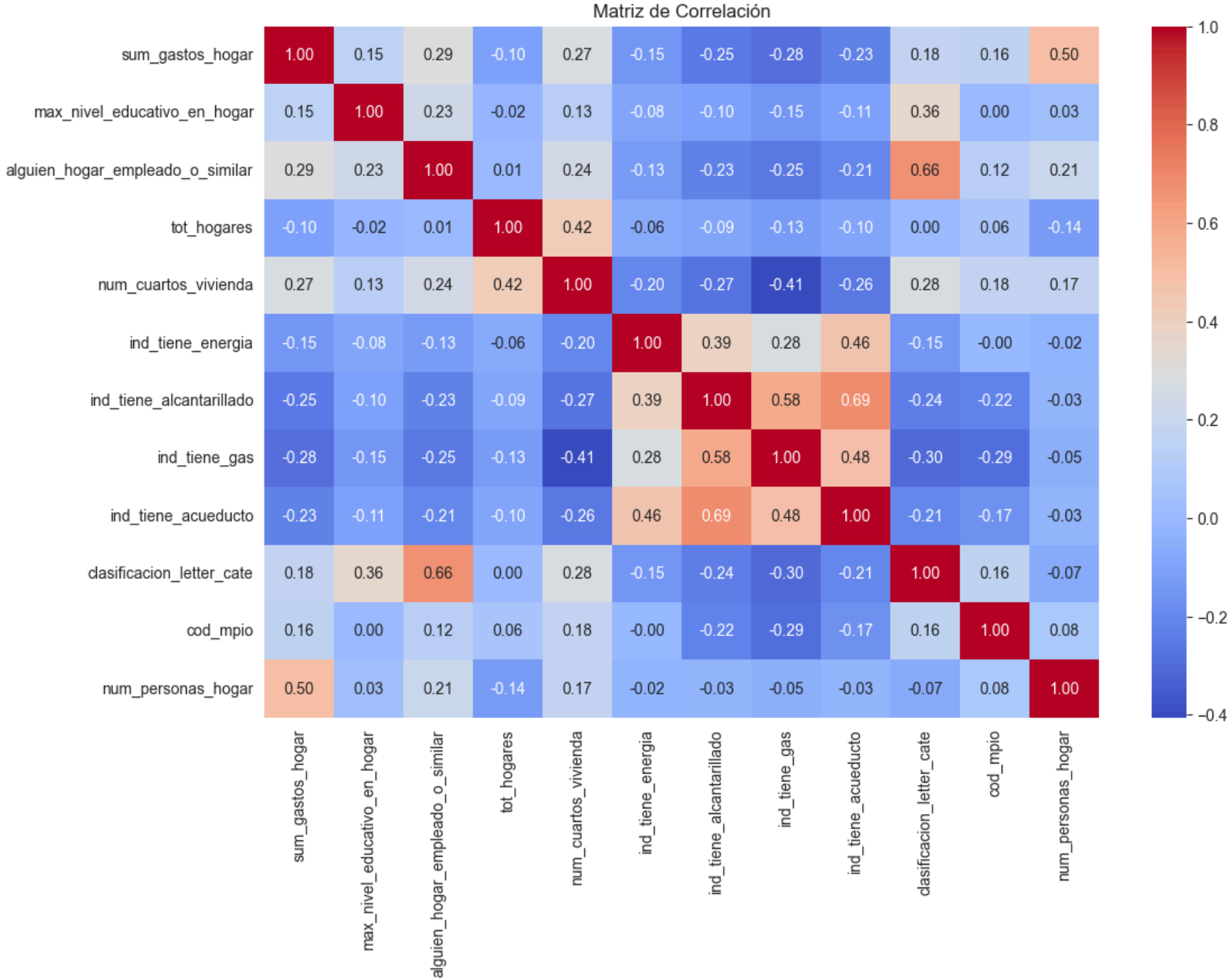
	ind_tiene_energia	ind_tiene_alcantarillado	\
sum_gastos_hogar	-0.151840	-0.247196	
max_nivel_educativo_en_hogar	-0.078723	-0.104244	
alguien_hogar_empleado_o_similar	-0.132541	-0.227250	
tot_hogares	-0.064520	-0.093104	
num_cuartos_vivienda	-0.201270	-0.270571	
ind_tiene_energia	1.000000	0.387971	
ind_tiene_alcantarillado	0.387971	1.000000	
ind_tiene_gas	0.277847	0.576752	
ind_tiene_acueducto	0.463092	0.688782	
clasificacion_letter_cate	-0.146559	-0.235712	
cod_mpio	-0.002439	-0.219460	
num_personas_hogar	-0.022592	-0.031914	

	ind_tiene_gas	ind_tiene_acueducto	\
sum_gastos_hogar	-0.282686	-0.229032	
max_nivel_educativo_en_hogar	-0.147714	-0.105330	
alguien_hogar_empleado_o_similar	-0.253425	-0.210644	
tot_hogares	-0.126507	-0.097808	
num_cuartos_vivienda	-0.405859	-0.255069	
ind_tiene_energia	0.277847	0.463092	
ind_tiene_alcantarillado	0.576752	0.688782	
ind_tiene_gas	1.000000	0.481857	
ind_tiene_acueducto	0.481857	1.000000	
clasificacion_letter_cate	-0.303711	-0.210909	
cod_mpio	-0.288022	-0.169882	
num_personas_hogar	-0.051415	-0.028502	

	clasificacion_letter_cate	cod_mpio	\
sum_gastos_hogar	0.177611	0.160707	
max_nivel_educativo_en_hogar	0.356183	0.003728	
alguien_hogar_empleado_o_similar	0.660969	0.123474	
tot_hogares	0.003851	0.060938	
num_cuartos_vivienda	0.278434	0.181070	
ind_tiene_energia	-0.146559	-0.002439	
ind_tiene_alcantarillado	-0.235712	-0.219460	
ind_tiene_gas	-0.303711	-0.288022	
ind_tiene_acueducto	-0.210909	-0.169882	
clasificacion_letter_cate	1.000000	0.160473	
cod_mpio	0.160473	1.000000	
num_personas_hogar	-0.066023	0.078940	

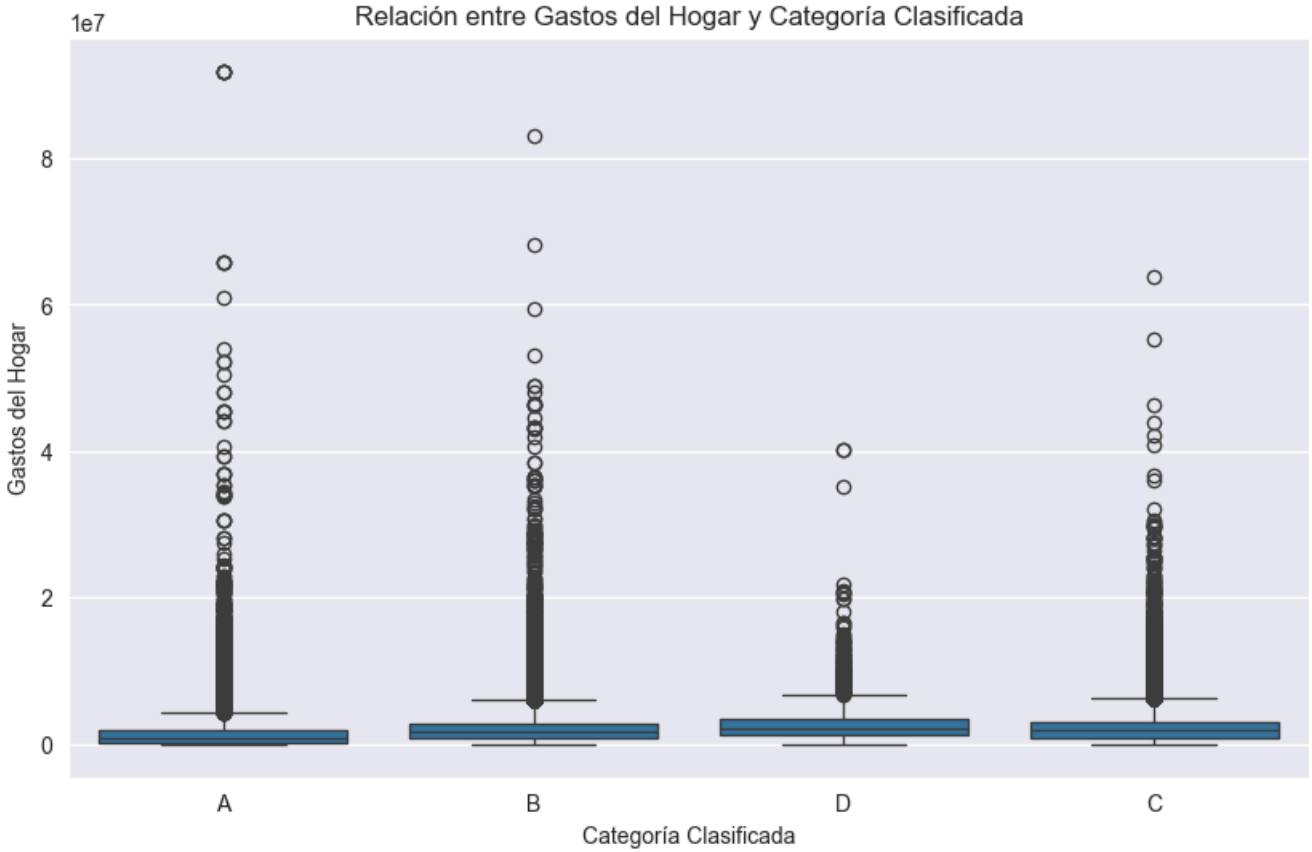
	num_personas_hogar
sum_gastos_hogar	0.498893
max_nivel_educativo_en_hogar	0.034175
alguien_hogar_empleado_o_similar	0.212869
tot_hogares	-0.135322

```
num_cuartos_vivienda                0.170233
ind_tiene_energia                  -0.022592
ind_tiene_alcantarillado           -0.031914
ind_tiene_gas                      -0.051415
ind_tiene_acueducto                -0.028502
clasificacion_letter_cate          -0.066023
cod_mpio                            0.078940
num_personas_hogar                  1.000000
```


Matriz de Correlación
sum_gastos_hogar 1.00 0.15 0.29 -0.10 0.27 -0.15 -0.25 -0.28 -0.23 0.18 0.16 0.50
max_nivel_educativo_en_hogar 0.15 1.00 0.23 -0.02 0.13 -0.08 -0.10 -0.15 -0.11 0.36 0.00 0.03
alguien_hogar_empleado_o_similar 0.29 0.23 1.00 0.01 0.24 -0.13 -0.23 -0.25 -0.21 0.66 0.12 0.21
tot_hogares -0.10 -0.02 0.01 1.00 0.42 -0.06 -0.09 -0.13 -0.10 0.00 0.06 -0.14
num_cuartos_vivienda 0.27 0.13 0.24 0.42 1.00 -0.20 -0.27 -0.41 -0.26 0.28 0.18 0.17
ind_tiene_energia -0.15 -0.08 -0.13 -0.06 -0.20 1.00 0.39 0.28 0.46 -0.15 -0.00 -0.02
ind_tiene_alcantarillado -0.25 -0.10 -0.23 -0.09 -0.27 0.39 1.00 0.58 0.69 -0.24 -0.22 -0.03
ind_tiene_gas -0.28 -0.15 -0.25 -0.13 -0.41 0.28 0.58 1.00 0.48 -0.30 -0.29 -0.05
ind_tiene_acueducto -0.23 -0.11 -0.21 -0.10 -0.26 0.46 0.69 0.48 1.00 -0.21 -0.17 -0.03
clasificacion_letter_cate 0.18 0.36 0.66 0.00 0.28 -0.15 -0.24 -0.30 -0.21 1.00 0.16 -0.07
cod_mpio 0.16 0.00 0.12 0.06 0.18 -0.00 -0.22 -0.29 -0.17 0.16 1.00 0.08
num_personas_hogar 0.50 0.03 0.21 -0.14 0.17 -0.02 -0.03 -0.05 -0.03 -0.07 0.08 1.00
sum_gastos_hogar
max_nivel_educativo_en_hogar
alguien_hogar_empleado_o_similar
tot_hogares
num_cuartos_vivienda
ind_tiene_energia
ind_tiene_alcantarillado
ind_tiene_gas
ind_tiene_acueducto
clasificacion_letter_cate
cod_mpio
num_personas_hogar
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

In [24]:
```
corr, p_value = pearsonr(data_correlation['sum_gastos_hogar'], data_correlation['clasificacion_letter_cate'])
print(f"Coeficiente de correlación entre 'sum_gastos_hogar' y 'clasificacion_letter': {corr:.2f}")
print(f"Valor p: {p_value:.4f}")

# Visualizar la relación entre 'sum_gastos_hogar' y 'clasificacion_letter'
plt.figure(figsize=(10, 6))
sns.boxplot(x='clasificacion_letter', y='sum_gastos_hogar', data=data)
plt.title('Relación entre Gastos del Hogar y Categoría Clasificada')
plt.xlabel('Categoría Clasificada')
plt.ylabel('Gastos del Hogar')
plt.show()
```

```
Coeficiente de correlación entre 'sum_gastos_hogar' y 'clasificacion_letter': 0.18
Valor p: 0.0000
```



In [28]:
```
X = data_correlation[['clasificacion_letter_cate']]
y = data_correlation['sum_gastos_hogar']

X = sm.add_constant(X)
```

```
model = sm.OLS(y, X).fit()

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       sum_gastos_hogar   R-squared:                       0.032
Model:                           OLS   Adj. R-squared:                  0.032
Method:                Least Squares   F-statistic:                     8624.
Date:               Tue, 04 Jun 2024   Prob (F-statistic):               0.00
Time:                       08:05:29   Log-Likelihood:            -4.2273e+06
No. Observations:             264762   AIC:                         8.455e+06
Df Residuals:                 264760   BIC:                         8.455e+06
Df Model:                          1
Covariance Type:           nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                        1.452e+06   5730.419    253.370      0.000    1.44e+06    1.46e+06
clasificacion_letter_cate    4.516e+05   4862.831     92.866      0.000    4.42e+05    4.61e+05
==============================================================================
Omnibus:                   335042.881   Durbin-Watson:                   1.887
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        207122267.509
Skew:                           6.488   Prob(JB):                         0.00
Kurtosis:                     139.406   Cond. No.                         2.47
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [ ]:

El modelo de regresión lineal aplicado para analizar la relación entre los gastos del hogar (sum_gastos_hogar) y la categoría clasificada (clasificacion_letter_cate) proporciona varios resultados importantes que debemos considerar.

Resumen de Resultados Coeficiente de Determinación (R-cuadrado):

R-cuadrado: 0.032: Este valor indica que solo el 3.2% de la variabilidad en los gastos del hogar puede ser explicada por la categoría clasificada. Esto sugiere que la relación entre estas variables es débil y que otros factores no incluidos en este modelo probablemente tienen una influencia mayor en los gastos del hogar.

Significancia Estadística:

Prob (F-statistic): 0.00: El valor p extremadamente bajo sugiere que el modelo en su conjunto es significativo, es decir, hay una relación estadísticamente significativa entre los gastos del hogar y la categoría clasificada. P-valor del coeficiente de clasificacion_letter_cate: 0.000: El coeficiente de la variable clasificacion_letter_cate es altamente significativo, lo que confirma que la categoría clasificada tiene un efecto sobre los gastos del hogar.

Coeficientes del Modelo:

Intercepto (const): 1.452e+06: Este es el valor esperado de los gastos del hogar cuando la categoría clasificada es cero (considerando cómo se codifican las categorías). Coeficiente de clasificacion_letter_cate: 4.516e+05: Esto indica que, en promedio, un cambio unitario en la categoría clasificada se asocia con un

aumento de 451,600 en los gastos del hogar.

Diagnóstico del Modelo:

Omnibus, Prob(Omnibus), Jarque-Bera, Prob(JB): Estos valores indican que hay problemas de normalidad en la distribución de los residuos, lo cual puede afectar la validez de los intervalos de confianza y pruebas de hipótesis del modelo. Durbin-Watson: 1.887: Este valor cercano a 2 sugiere que no hay una autocorrelación significativa de los residuos.

Conclusiones

Relación Débil pero Significativa:

Aunque la relación entre sum_gastos_hogar y clasificacion_letter_cate es estadísticamente significativa, el valor de R-cuadrado sugiere que esta relación es débil. Es decir, la categoría clasificada solo explica una pequeña parte de la variabilidad en los gastos del hogar.

Impacto de la Categoría Clasificada:

El coeficiente positivo y significativo para clasificacion_letter_cate sugiere que hay una tendencia general de aumento en los gastos del hogar con cambios en la categoría clasificada, pero esta tendencia es débil y no captura toda la complejidad de los gastos del hogar.

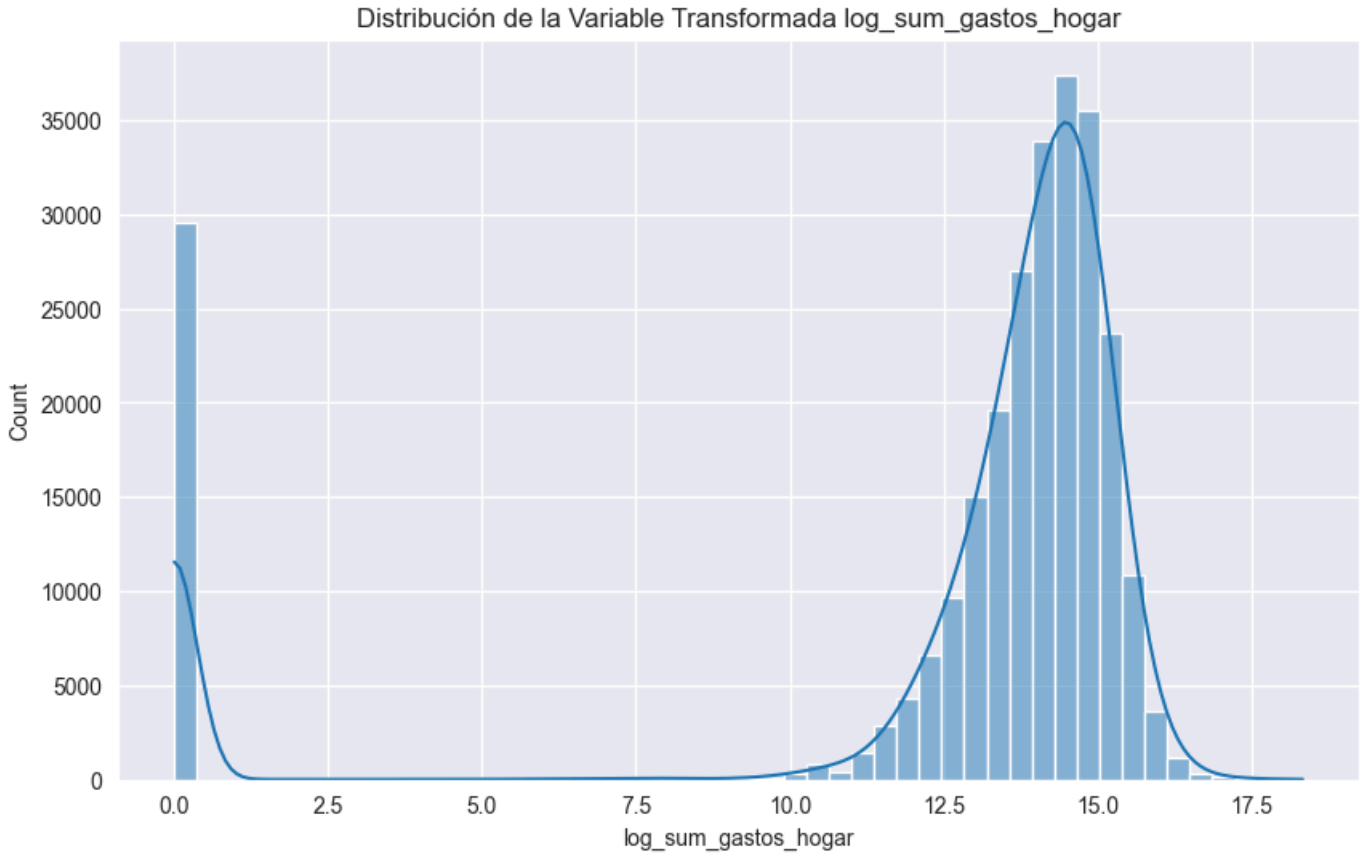
Necesidad de Considerar Otros Factores:

Dado el bajo R-cuadrado, es probable que otros factores no considerados en este modelo tengan un impacto significativo en los gastos del hogar. Sería beneficioso investigar y añadir otras variables relevantes al modelo, como ingresos, tamaño del hogar, ubicación geográfica, y otras características socioeconómicas.

In [39]:
```
data_correlation['log_sum_gastos_hogar'] = np.log(data['sum_gastos_hogar'] + 1)  # +1 para evitar log(0)

# Graficar la distribución de la variable transformada
plt.figure(figsize=(10, 6))
sns.histplot(data_correlation['log_sum_gastos_hogar'], bins=50, kde=True)
plt.title('Distribución de la Variable Transformada log_sum_gastos_hogar')
plt.show()

# Calcular la matriz de correlación con la variable transformada
correlation_matrix = data_correlation.corr()
print("Matriz de Correlación con Variable Transformada:")
print(correlation_matrix)

# Reajustar el modelo de regresión lineal con la variable transformada
X = data_correlation[['clasificacion_letter_cate']]
y = data_correlation['log_sum_gastos_hogar']

# Añadir una constante al modelo (intercepto)
X = sm.add_constant(X)
```

```python
# Ajustar el modelo de regresión lineal
model = sm.OLS(y, X).fit()

# Mostrar el resumen del modelo
print(model.summary())

# Visualizar la relación entre 'log_sum_gastos_hogar' y 'clasificacion_letter'
plt.figure(figsize=(10, 6))
sns.boxplot(x='clasificacion_letter', y='log_sum_gastos_hogar', data=data)
plt.title('Relación entre Gastos del Hogar (Log Transformado) y Categoría Clasificada')
plt.xlabel('Categoría Clasificada')
plt.ylabel('Gastos del Hogar (Log Transformado)')
plt.show()
```

```
/var/folders/vh/c65wsrm94x92d8rfk33vvpwr0000gn/T/ipykernel_89469/1699190667.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data_correlation['log_sum_gastos_hogar'] = np.log(data['sum_gastos_hogar'] + 1)  # +1 para evitar log(0)
```



Matriz de Correlación con Variable Transformada:

	sum_gastos_hogar
sum_gastos_hogar	1.000000
max_nivel_educativo_en_hogar	0.151282
alguien_hogar_empleado_o_similar	0.292638
tot_hogares	-0.098390
num_cuartos_vivienda	0.270021
ind_tiene_energia	-0.151840
ind_tiene_alcantarillado	-0.247196
ind_tiene_gas	-0.282686
ind_tiene_acueducto	-0.229032
clasificacion_letter_cate	0.177611
cod_mpio	0.160707
num_personas_hogar	0.498893
log_sum_gastos_hogar	0.453667

	max_nivel_educativo_en_hogar
sum_gastos_hogar	0.151282
max_nivel_educativo_en_hogar	1.000000
alguien_hogar_empleado_o_similar	0.227704
tot_hogares	-0.016901
num_cuartos_vivienda	0.129426
ind_tiene_energia	-0.078723
ind_tiene_alcantarillado	-0.104244
ind_tiene_gas	-0.147714
ind_tiene_acueducto	-0.105330
clasificacion_letter_cate	0.356183
cod_mpio	0.003728
num_personas_hogar	0.034175
log_sum_gastos_hogar	0.096822

	alguien_hogar_empleado_o_similar
sum_gastos_hogar	0.292638
max_nivel_educativo_en_hogar	0.227704
alguien_hogar_empleado_o_similar	1.000000
tot_hogares	0.005942
num_cuartos_vivienda	0.243927
ind_tiene_energia	-0.132541
ind_tiene_alcantarillado	-0.227250
ind_tiene_gas	-0.253425
ind_tiene_acueducto	-0.210644
clasificacion_letter_cate	0.660969
cod_mpio	0.123474
num_personas_hogar	0.212869
log_sum_gastos_hogar	0.228910

	tot_hogares	num_cuartos_vivienda
sum_gastos_hogar	-0.098390	0.270021
max_nivel_educativo_en_hogar	-0.016901	0.129426
alguien_hogar_empleado_o_similar	0.005942	0.243927
tot_hogares	1.000000	0.421896
num_cuartos_vivienda	0.421896	1.000000

ind_tiene_energia	-0.064520	-0.201270	
ind_tiene_alcantarillado	-0.093104	-0.270571	
ind_tiene_gas	-0.126507	-0.405859	
ind_tiene_acueducto	-0.097808	-0.255069	
clasificacion_letter_cate	0.003851	0.278434	
cod_mpio	0.060938	0.181070	
num_personas_hogar	-0.135322	0.170233	
log_sum_gastos_hogar	0.001116	0.244364	

	ind_tiene_energia	ind_tiene_alcantarillado	\
sum_gastos_hogar	-0.151840	-0.247196	
max_nivel_educativo_en_hogar	-0.078723	-0.104244	
alguien_hogar_empleado_o_similar	-0.132541	-0.227250	
tot_hogares	-0.064520	-0.093104	
num_cuartos_vivienda	-0.201270	-0.270571	
ind_tiene_energia	1.000000	0.387971	
ind_tiene_alcantarillado	0.387971	1.000000	
ind_tiene_gas	0.277847	0.576752	
ind_tiene_acueducto	0.463092	0.688782	
clasificacion_letter_cate	-0.146559	-0.235712	
cod_mpio	-0.002439	-0.219460	
num_personas_hogar	-0.022592	-0.031914	
log_sum_gastos_hogar	-0.205984	-0.247581	

	ind_tiene_gas	ind_tiene_acueducto	\
sum_gastos_hogar	-0.282686	-0.229032	
max_nivel_educativo_en_hogar	-0.147714	-0.105330	
alguien_hogar_empleado_o_similar	-0.253425	-0.210644	
tot_hogares	-0.126507	-0.097808	
num_cuartos_vivienda	-0.405859	-0.255069	
ind_tiene_energia	0.277847	0.463092	
ind_tiene_alcantarillado	0.576752	0.688782	
ind_tiene_gas	1.000000	0.481857	
ind_tiene_acueducto	0.481857	1.000000	
clasificacion_letter_cate	-0.303711	-0.210909	
cod_mpio	-0.288022	-0.169882	
num_personas_hogar	-0.051415	-0.028502	
log_sum_gastos_hogar	-0.270363	-0.254467	

	clasificacion_letter_cate	cod_mpio	\
sum_gastos_hogar	0.177611	0.160707	
max_nivel_educativo_en_hogar	0.356183	0.003728	
alguien_hogar_empleado_o_similar	0.660969	0.123474	
tot_hogares	0.003851	0.060938	
num_cuartos_vivienda	0.278434	0.181070	
ind_tiene_energia	-0.146559	-0.002439	
ind_tiene_alcantarillado	-0.235712	-0.219460	
ind_tiene_gas	-0.303711	-0.288022	
ind_tiene_acueducto	-0.210909	-0.169882	
clasificacion_letter_cate	1.000000	0.160473	
cod_mpio	0.160473	1.000000	
num_personas_hogar	-0.066023	0.078940	

```
log_sum_gastos_hogar                              0.169130  0.111906

                                num_personas_hogar  log_sum_gastos_hogar
sum_gastos_hogar                          0.498893              0.453667
max_nivel_educativo_en_hogar              0.034175              0.096822
alguien_hogar_empleado_o_similar          0.212869              0.228910
tot_hogares                              -0.135322              0.001116
num_cuartos_vivienda                      0.170233              0.244364
ind_tiene_energia                        -0.022592             -0.205984
ind_tiene_alcantarillado                 -0.031914             -0.247581
ind_tiene_gas                            -0.051415             -0.270363
ind_tiene_acueducto                      -0.028502             -0.254467
clasificacion_letter_cate                -0.066023              0.169130
cod_mpio                                  0.078940              0.111906
num_personas_hogar                        1.000000              0.164955
log_sum_gastos_hogar                      0.164955              1.000000
                            OLS Regression Results
==============================================================================
Dep. Variable:     log_sum_gastos_hogar   R-squared:                       0.029
Model:                              OLS   Adj. R-squared:                  0.029
Method:                   Least Squares   F-statistic:                     7796.
Date:                  Tue, 04 Jun 2024   Prob (F-statistic):               0.00
Time:                          08:36:55   Log-Likelihood:            -7.7312e+05
No. Observations:                264762   AIC:                         1.546e+06
Df Residuals:                    264760   BIC:                         1.546e+06
Df Model:                             1
Covariance Type:              nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          11.7481      0.012    950.195      0.000      11.724      11.772
clasificacion_letter_cate       0.9264      0.010     88.297      0.000       0.906       0.947
==============================================================================
Omnibus:                   107957.287   Durbin-Watson:                   1.834
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           341847.981
Skew:                          -2.207   Prob(JB):                         0.00
Kurtosis:                       6.391   Cond. No.                         2.47
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[39], line 29
     27 # Visualizar la relación entre 'log_sum_gastos_hogar' y 'clasificacion_letter'
     28 plt.figure(figsize=(10, 6))
---> 29 sns.boxplot(x='clasificacion_letter', y='log_sum_gastos_hogar', data=data)
     30 plt.title('Relación entre Gastos del Hogar (Log Transformado) y Categoría Clasificada')
     31 plt.xlabel('Categoría Clasificada')

File /usr/local/lib/python3.11/site-packages/seaborn/categorical.py:1597, in boxplot(data, x, y, hue, order, hue_order, orient, color, pal
ette, saturation, fill, dodge, width, gap, whis, linecolor, linewidth, fliersize, hue_norm, native_scale, log_scale, formatter, legend, a
x, **kwargs)
   1589 def boxplot(
   1590     data=None, *, x=None, y=None, hue=None, order=None, hue_order=None,
   1591     orient=None, color=None, palette=None, saturation=.75, fill=True,
   (...)
   1594     legend="auto", ax=None, **kwargs
   1595 ):
-> 1597     p = _CategoricalPlotter(
   1598         data=data,
   1599         variables=dict(x=x, y=y, hue=hue),
   1600         order=order,
   1601         orient=orient,
   1602         color=color,
   1603         legend=legend,
   1604     )
   1606     if ax is None:
   1607         ax = plt.gca()

File /usr/local/lib/python3.11/site-packages/seaborn/categorical.py:67, in _CategoricalPlotter.__init__(self, data, variables, order, orie
nt, require_numeric, color, legend)
     56 def __init__(
     57     self,
     58     data=None,
   (...)
     64     legend="auto",
     65 ):
---> 67     super().__init__(data=data, variables=variables)
     69     # This method takes care of some bookkeeping that is necessary because the
     70     # original categorical plots (prior to the 2021 refactor) had some rules that
     71     # don't fit exactly into VectorPlotter logic. It may be wise to have a second
   (...)
     76     # default VectorPlotter rules. If we do decide to make orient part of the
     77     # _base variable assignment, we'll want to figure out how to express that.
     78     if self.input_format == "wide" and orient in ["h", "y"]:

File /usr/local/lib/python3.11/site-packages/seaborn/_base.py:634, in VectorPlotter.__init__(self, data, variables)
    629 # var_ordered is relevant only for categorical axis variables, and may
    630 # be better handled by an internal axis information object that tracks
    631 # such information and is set up by the scale_* methods. The analogous
    632 # information for numeric axes would be information about log scales.
    633 self._var_ordered = {"x": False, "y": False}  # alt., used DefaultDict
```

```
--> 634 self.assign_variables(data, variables)
    636 # TODO Lots of tests assume that these are called to initialize the
    637 # mappings to default values on class initialization. I'd prefer to
    638 # move away from that and only have a mapping when explicitly called.
    639 for var in ["hue", "size", "style"]:

File /usr/local/lib/python3.11/site-packages/seaborn/_base.py:679, in VectorPlotter.assign_variables(self, data, variables)
    674 else:
    675     # When dealing with long-form input, use the newer PlotData
    676     # object (internal but introduced for the objects interface)
    677     # to centralize / standardize data consumption logic.
    678     self.input_format = "long"
--> 679     plot_data = PlotData(data, variables)
    680     frame = plot_data.frame
    681     names = plot_data.names

File /usr/local/lib/python3.11/site-packages/seaborn/_core/data.py:58, in PlotData.__init__(self, data, variables)
     51 def __init__(
     52     self,
     53     data: DataSource,
     54     variables: dict[str, VariableSpec],
     55 ):
     57     data = handle_data_source(data)
---> 58     frame, names, ids = self._assign_variables(data, variables)
     60     self.frame = frame
     61     self.names = names

File /usr/local/lib/python3.11/site-packages/seaborn/_core/data.py:232, in PlotData._assign_variables(self, data, variables)
    230     else:
    231         err += "An entry with this name does not appear in `data`."
--> 232     raise ValueError(err)
    234 else:
    235
    236     # Otherwise, assume the value somehow represents data
    237
    238     # Ignore empty data structures
    239     if isinstance(val, Sized) and len(val) == 0:

ValueError: Could not interpret value `log_sum_gastos_hogar` for `y`. An entry with this name does not appear in `data`.
```

<Figure size 1000x600 with 0 Axes>

Al realizar la transformacion no existe correlacion signicativa entre las variables

In [ ]:

Anexo 8

Repositorio Github

https://github.com/wladimiravila/modelos-infancia

Valle del Cauca

- **Regresión.**
  https://github.com/wladimiravila/modelos-infancia/blob/main/ModelLogisticRegressionValleDelCauca.ipynb

- **Xgboost**
  https://github.com/wladimiravila/modelos-infancia/blob/main/ModelXGboostValleDelCauca.ipynb

- **Random Forest**
  https://github.com/wladimiravila/modelos-infancia/blob/main/ModelRandomForestValleDelCauca.ipynb

Norte del Cauca

- **Regresión.**
  https://github.com/wladimiravila/modelos-infancia/blob/main/ModelLogisticRegressionNorteDelCauca.ipynb

- **Xgboost**
  https://github.com/wladimiravila/modelos-infancia/blob/main/ModelXGboostNorteCauca.ipynb

- **Random Forest**

  https://github.com/wladimiravila/modelos-infancia/blob/main/ModelRandomForestNorteCauca.ipynb

Valle y norte del Cauca.

- **Regresión.**
  https://github.com/wladimiravila/modelos-infancia/blob/main/ModelLogisticRegressionNCaucaYVlleDelCauca.ipynb

- **Xgboost**
  https://github.com/wladimiravila/modelos-infancia/blob/main/ModelXGboostNCaucaYValleDelCauca.ipynb

- **Random Forest**
  https://github.com/wladimiravila/modelos-infancia/blob/main/ModelRandomForestNCaucaYValleDelCauca.ipynb