PROYECTO PILOTO DE ANALÍTICA DE DATOS - DETECCIÓN DE ANOMALÍAS EN EL SISBEN Y MOVILIDAD

Marzo 2020

8

DATA-POP ALLIANCE

mit media lab

DNP Departamento Nacional de Planeación

iNNpulsa Colombia

Producido por un equipo compuesto por:

Andrés Lozano, María Antonia Bravo

Basado en documentación de DataInnova para iNNpulsa Colombia.

Coordinado por Emmanuel Letouzé

Bajo la supervisión general de Alex Pentland

Diagramación editorial por Paola Caile

Marzo de 2020
Versión revisada y ajustada

ÍNDICE

Glosario 2

Introducción 3

Contexto Social 3

El SISBEN como sistema de focalización del gasto social público 3

El SISBEN III 4

La problemática: las anomalías en el SISBEN 5

Objetivos del piloto 6

I. Inventario de datos utilizados 7

II. Descripción técnica del análisis efectuado 8

1. Arquitectura de datos 8

Limpieza de datos 8

Estructura de la base de datos tipo panel 9

Creación de identificadores individuales por persona 10

2. Modelamiento y análisis de datos 11

III. Resultados Cuantitativos y Cualitativos: Interpretación de los modelos a la luz de la problemática social 14

1. Establecer los grupos de personas que por sus características de comportamiento y condiciones socioeconómicas serían propensas a incurrir en acciones fraudulentas con respecto a su puntaje en el SISBEN. 14

Los resultados de la clasificación de anomalías 14

Prueba de Concepto: la interoperabilidad entre las fuentes de datos públicas como potencial solución a la desactualización de la base de datos del SISBEN. 16

2. Identificar los factores que motivan el comportamiento de las personas registradas en la búsqueda de beneficios mediante la manipulación de la información. 17

3. Analizar tendencias y comportamientos encaminados a promover la movilidad social. 20

4. Identificar y caracterizar la(s) combinación(es) de programa(s) sociales más efectivos para que las personas en el SISBEN superen la condición de vulnerabilidad al igual que el umbral de puntaje para recibir ayuda de programas sociales. 22

IV. Retos y Recomendaciones 25

Recomendaciones Específicas: Proyecto Piloto Detección de Anomalías en SISBEN 26

Recomendaciones Generales: Proyectos de Analítica de Datos del Estado 27

Anexo A. Creación de Panel de Datos 28

Anexo B. Detalle de modelos de Aprendizaje de Máquinas 30

Anexo C. Etiquetas de anomalías en base de datos SISBEN 38

Glosario

Arquitectura de datos: arquitectura que describe cómo se recolecta, almacena, transforma, distribuye y consume los datos dentro de una organización o empresa. Incluye reglas que conciernen la estructuración de formatos, incluyendo bases de datos y sistemas de archivos, además de incluir reglas sobre cómo se vinculan datos a través de sistemas para el uso de estos a través del proceso comercial (DalleMule y Davenport 2017). Entre los elementos de una arquitectura de datos se encuentra el acceso a los datos, almacenamiento en la nube, procesamiento de datos, redes, encriptación, hosting entre otros (Knight 2018).

Servicios de AWS: Amazon Web Services (AWS) es una plataforma de servicios basados en la nube que incluye recursos para el cómputo, almacenamiento, y análisis de datos, incluyendo herramientas para desarrolladores y aplicaciones empresariales ("AWS" s/f).

Clasificación binaria de datos: modelo para predecir la pertenencia de un dato a una de dos clases. Por ejemplo, para un dato que puede tomar el valor 1 cuando presenta una característica y el valor 0 en ausencia de la misma, se evalúa la capacidad de un modelo en predecir si un dato presenta la característica o no. Los resultados de las predicciones generan cuatro posibilidades:

- Verdaderos positivos: el modelo predice que un dato presenta el valor (1) y efectivamente el dato tenía el valor (1).
- Verdaderos negativos: el modelo predice que un dato no presenta el valor (0) y efectivamente el dato no tenía el valor (0).
- Falsos positivos: el modelo predice que un dato presenta el valor (1) y el dato no tiene el valor (0).
- Falsos negativos: el modelo predice que el dato no presenta el valor (0) y el dato sí tiene el valor (1).

Razón de verdaderos positivos (Sensibilidad): proporción de verdaderos positivos sobre el total de valores positivos (verdaderos positivos y falsos negativos).

Razón de falsos positivos (Especificidad): proporción de verdaderos negativos sobre el total de valores negativos (verdaderos negativos y falsos positivos).

Área bajo la curva (AUC) ROC: métrica para evaluar la capacidad predictiva de modelos para clasificación binaria de datos. Se mide como la sensibilidad sobre la especificidad de un modelo. La predicción es mejor cuando el AUC se acerca a 1, 0.5 equivale a predicciones hechas al azar.

Introducción

Este documento presenta el trabajo realizado por la empresa Data Innovation S.A, en el marco del proyecto piloto de analítica de datos adelantado por iNNpulsa Colombia, entidad a cargo de la promoción de la innovación, el emprendimiento y desarrollo empresarial en el país. Dicho proyecto se llevó a cabo en el marco de la consultoría para la elaboración de una estrategia de Big Data para el Gobierno de Colombia, suscrita entre iNNpulsa Colombia y el Massachusetts Institute of Technology (MIT). A lo largo del desarrollo del documento se plasman los resultados del proyecto piloto, los datos y metodologías usadas, así como los retos u obstáculos encontrados en la realización del mismo. La información aquí presente es un resúmen de los documentos y entregables elaborados por Data Innovation S.A y entregados a iNNpulsa, conforme el avance del piloto.

El proyecto de analítica se llevó a cabo en 4 etapas, en las cuales está estructurado el presente documento. Primero, se empezó con el análisis del problema. Allí se hizo un análisis de las anomalías en su contexto social para generar una serie de hipótesis, las cuales están reflejadas en la sección de **contexto y objetivos del piloto**. Luego se pasó a la etapa de exploración de los datos. En esta etapa se establecieron cuáles eran las fuentes de datos más pertinentes para responder a los objetivos y se creó una base de datos tipo panel para observar los datos a través del tiempo; el detalle de dichas fuentes de datos se encuentra en la sección de **inventario y arquitectura de datos utilizados**. En la tercera etapa, se implementaron técnicas de aprendizaje de máquinas con cuatro modelos (Gradient Boosting Machine, Isolation Forest, Long-Short Term Memory y Autoencoder con Long-Short Term memory) y se evaluaron sus resultados, los cuáles se encuentran en la sección de **modelamiento y análisis de datos.** Finalmente, se llevó a cabo la etapa de evaluación de los modelos. En esta etapa se analizaron y visualizaron los resultados para responder a los objetivos del piloto. El detalle de los resultados se encuentra en la sección de **interpretación de resultados a la luz de la problemática social y recomendaciones.** Vale recalcar que a lo largo del proyecto se tuvo en cuenta el enfoque de derechos con el fin de tomar en cuenta las diferencias socioeconómicas, demográficas y geográficas de la población para el análisis. Dicho enfoque se basó en las estratificaciones que maneja el Departamento Nacional de Planeación (DNP) para el análisis de la base de datos del SISBEN. Es decir que para el análisis se dividió la población según su residencia en poblaciones urbanas de las 14 principales ciudades del país, otras áreas urbanas y las áreas rurales de Colombia.

Contexto Social

<u>El SISBEN como sistema de focalización del gasto social público</u>

A través del Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales (SISBEN), el Gobierno de Colombia focaliza recursos para la población más necesitada del país, buscando mejorar la calidad de vida y asegurar el bienestar de la población. Desde 1995, y a pesar de que el sistema ha tenido modificaciones en su metodología de identificación de

potenciales beneficiarios, el SISBEN ha utilizado encuestas socioeconómicas para asignar un puntaje de 0 a 100 a los hogares de potenciales beneficiarios. Entre mayor es el puntaje otorgado a los hogares, mejor se considera que es el nivel socioeconómico de las personas en dicho hogar.

El Departamento Nacional de Planeación (DNP) es quien define los criterios para acceder como beneficiario a los programas sociales del país, y también quien se encarga de consolidar la información socioeconómica de los ciudadanos y efectuar el control de calidad de esta información. A través de los años, el enfoque de la medición de pobreza ha cambiado de manera que la metodología del SISBEN I (la primera versión del sistema), difiere de aquella usada en sus siguientes versiones, el SISBEN II y III. Mientras el primero se enfocó en la medición de la pobreza a través de la exclusión productiva (falta de acceso a la generación de ingresos), las versiones que le siguieron (II y III) se enfocaron en medir la exclusión social, además de introducir estratificaciones según el área geográfica donde habita la población objetivo. En el SISBEN I, se asignaron puntajes a nivel nacional, mientras que en el SISBEN II se separaron por puntajes de población urbana y rural. Finalmente, el SISBEN III considera tres grupos: 1) las 14 ciudades principales del país, 2) el resto de ciudades urbanas y, 3) las áreas rurales.

El SISBEN III

El objeto de estudio de este proyecto piloto es la tercera y más reciente versión del SISBEN. Esta se caracteriza por introducir dos nociones en la medición del puntaje: primero, la noción de pobreza multidimensional, y segundo el enfoque de capacidades de Amartya Sen. La pobreza multidimensional enfoca la medición de la pobreza en términos de los niveles de acceso a la salud, educación y el estándar de vida de las personas. El enfoque de capacidades establece que no basta con medir la pobreza a través del acceso a bienes o servicios, sino que se deben incluir mediciones sobre la capacidad que tienen las personas de hacer uso de los mismos. Para medir la pobreza multidimensional, la encuesta del SISBEN recoge información sobre el acceso de los individuos a saneamiento, agua potable, electricidad, vivienda y activos. Por otro lado, para el enfoque de capacidades se recoge información sobre variables como la nutrición, la mortalidad infantil, los años de escolaridad y la asistencia escolar de las personas. En total, para asignar un puntaje, el SISBEN III recoge 90 preguntas sobre la condición socioeconómica de los colombianos.

Una vez se establece el puntaje a través de la encuesta del SISBEN, y según el presupuesto disponible para los diferentes programas sociales, se establecen puntos de corte que determinan la elegibilidad de potenciales beneficiarios para acceder a los diferentes programas sociales. Es importante aclarar que aunque los puntos de corte determinan la elegibilidad de los beneficiarios a los respectivos programas, el puntaje del SISBEN no es el único medio para ser elegible[1].

[1] Algunos de los programas sociales que brinda el gobierno Colombiano que hacen parte del presente estudio son Más Familias en Acción, Jóvenes en Acción, Estrategia Red Unidos y la Red de Seguridad Alimentaria, todos estos con una población objetivo y puntos de corte diferentes.

La problemática: las anomalías en el SISBEN

Sin embargo, y dada la naturaleza de recolección de información para determinar los puntajes, existe la posibilidad de que aquel otorgado a un individuo no refleje las nociones evaluadas de manera correcta. Esto se debe en gran parte a la naturaleza declarativa de las encuestas socioeconómicas, puesto que los encuestados pueden responder a las preguntas erróneamente - con o sin intención - y a la dificultad de actualizar estas respuesta con suficiente frecuencia. De hecho, el CONPES 3877 informa que "el 74% de las personas registradas en la base SISBEN III tiene información recolectada entre 2009 y 2011 que no ha sido actualizada". Una potencial consecuencia de estas limitaciones es la desviación de la inversión social de las personas que más necesidades socioeconómicas tienen.

Dada esta limitación, el DNP lleva a cabo una serie de medidas para identificar las anomalías, definidas en este estudio como aquella información recogida en la base de datos del SISBEN que tiene una alta probabilidad de no reflejar la situación real de la persona que suministró la información. Por ejemplo, cuando los municipios (quienes son los encargados de realizar las encuestas) entregan sus bases de datos, esta información se cruza con datos provenientes de fuentes como ADRES[2] y UGPP[3] para verificar si hay inconsistencias en la información de la encuesta del SISBEN. Ejemplos de anomalías se presentan cuando el registro tiene una denuncia de fraude, cuando se encuentra que en la UGPP los ingresos de la persona son inconsistentes con los reportados en el SISBEN, o entre otras, cuando existe un cambio en el tipo de vivienda. Adicionalmente, se buscan duplicados o cambios de variables con el periodo anterior, para determinar estas anomalías.

A partir de este cruce de información se genera en la base de datos del SISBEN una de las variables más importantes para el piloto, la variable '*Marca'*. Esta describe si el DNP encontró anomalías en los datos de alguna persona. Por ejemplo, cuando el individuo aparece duplicado en la base de datos, se encuentra que falleció o hay inconsistencias en su información, se le da el valor de 1 a esta variable, siendo el registro excluido de la base de datos certificada. Por otro lado, existen registros que se encuentran en verificación y que el DNP identifica como anomalía, que se marcan con el número 10. Cuando el individuo no cuenta con anomalías la variable '*Marca'* es igual a 0.

Los esfuerzos del DNP para identificar las anomalías y depurar la base de datos de errores han permitido bajar la cantidad de anomalías presentes en la base. Sin embargo, a pesar de los procesos existentes para identificar anomalías, tan solo se registran 100 mil casos, equivalentes al 0.2% del total de registros en la base. La necesidad de usar modelos de analítica para la identificación de anomalías en el SISBEN viene del hecho que, a pesar de los esfuerzos ya efectuados, aún existe incertidumbre sobre la magnitud verdadera de las anomalías en la base de datos del SISBEN. Esto se debe a que la información suministrada al

[2] Previamente FOSYGA - Administradora de los Recursos del Sistema General de Seguridad Social en Salud

[3] Unidad de Gestión Pensional y Parafiscales

SISBEN es declarativa, las bases de datos con las que el SISBEN verifica la información no cubren todas las preguntas de la encuesta y que para detectar anomalías, el DNP analiza los cambios en los datos respecto al corte anterior y no respecto a largos periodos de tiempo. Adicionalmente, la necesidad de focalizar el gasto social de Estado de manera adecuada, refuerza la necesidad de identificar estas anomalías. Por ejemplo, se estima que detectar anomalías equivalentes al 0.1% del gasto social, permitirían la redistribución de alrededor de 120 mil millones de pesos.

Por otro lado, para cumplir con el objetivo de mejorar la calidad de vida y promover el bienestar de las personas que hacen parte del SISBEN, es importante entender las características de la población y las combinaciones de programas sociales que más se relacionan con incrementos en el bienestar social de la población, es decir que promueven la movilidad social de las personas.

Objetivos del piloto

Dado este contexto, el proyecto piloto de analítica de datos de Data Innovation tiene como objeto usar metodologías de la Inteligencia Artificial para analizar los datos encontrados en las bases de datos del SISBEN, para:

1. Establecer los grupos de personas que por sus características de comportamiento y condiciones socioeconómicas serían propensas a presentar anomalías en su puntaje del SISBEN.

2. Identificar los factores que motivan el comportamiento de las personas registradas en la búsqueda de beneficios mediante la manipulación de la información.

3. Analizar tendencias y comportamientos encaminados a promover la movilidad social.

4. Identificar y caracterizar la(s) combinación(es) de programa(s) sociales más efectivos para que las personas en el SISBEN superen la condición de vulnerabilidad al igual que el umbral de puntaje para recibir ayuda de programas sociales.

I. Inventario de datos utilizados

Para este estudio se utilizaron seis bases de datos (SISBEN, Programas Sociales, UGPP, RUNT, RUES y ADRES), con las siguientes características:

1) SISBEN III: La base de datos del SISBEN III tiene 2.5TB de información y contiene 169 variables y aproximadamente 40 millones de registros, distribuidos en 4 niveles: vivienda, hogar, unidad de gasto y persona. La base contiene información de cada uno de estos niveles a través de 85 cortes en el tiempo (Octubre 2011- Agosto 2019).

2) Programas Sociales: La base de datos de programas sociales tiene 3.4 GB y cuenta con información de 33 programas. En estos datos se encuentran 29'992.756 registros de beneficios, es decir momentos en los que las personas contaron con el beneficio de los programas sociales del gobierno.

3) Unidad de Gestión Pensional y Parafiscales (UGPP)*[4]: además de brindarle información al DNP sobre las personas que tienen ingresos superiores a 36 millones de pesos al año, la UGPP tiene registros del estado de aportes a pensión de las personas, ya sea que se encuentren como activos cotizantes o no.

4) Registro Único Nacional de Tránsito (RUNT)*: este sistema de información registra y mantiene actualizada información sobre los registros automotores. El uso de esta base de datos es para determinar si el hogar estudiado cuenta con un vehículo.

5) Registro Único Empresarial Social (RUES)*: contiene información del Registro Mercantil de las empresas, entre otros. Esta fuente permite determinar si una persona tiene actividades comerciales.

6) ADRES*: Esta base de datos contiene información básica de los afiliados al sistema general de seguridad social de salud, incluyendo el estado de su afiliación, régimen de afiliación, fecha de afiliación, entre otras. La finalidad de esta fuente es determinar si una persona se encuentra fallecida, según el 'estado'.

[4] Las bases de datos que tienen "*" fueron usadas a modo de prueba de concepto, para mostrarle al DNP el valor agregado que tendría la validación y actualización de los datos del SISBEN a través de cruces con bases de datos públicas del Gobierno.

II. Descripción técnica del análisis efectuado

Esta sección describe el análisis efectuado desde la recepción de los datos, la arquitectura creada para llevar a cabo el análisis, hasta el tratamiento y el procesamiento de los datos para desarrollar el proyecto piloto de analítica de datos.

1. Arquitectura de datos

A. Limpieza de datos

Los datos se recibieron en formato SQLServer, y tras filtrar por los valores inválidos, se transformaron a .TXT para facilitar su almacenamiento en AWS dado el tamaño de la información entregada. Adicionalmente, estos datos se pasaron a formato Parquet, un formato de código abierto que permite comprimir el tamaño de los archivos .txt hasta cuatro veces. Para reducir el tiempo de lectura por cada corte de información se usó Apache Spark, un framework de computación en clúster[5].

Durante la lectura de los datos recibidos, se encontraron algunas inconsistencias que podían afectar el análisis, por lo cual se incurrió en varios procesos de limpieza. Por ejemplo, se encontró que en algunos cortes aparecen variables que no están presentes en otros cortes, es decir que la base de datos no es homogénea a través de todos los cortes. Se procedió a eliminar este tipo de variables para que se pudieran analizar los cortes de la base de datos de forma homogénea.

Se identificaron los cortes con información duplicada o errores, como por ejemplo personas con más de 100 años pero con tarjeta de identidad. También, se identificaron y eliminaron datos de personas que tenían comas "," en el nombre (142 registros). Fueron removidos aquellos registros que en la variable Marca tenían el valor de 1, es decir registros de individuos que no entraron a la base de datos certificada del DNP (2'758.491 registros).

La Tabla 1 resume los registros en la variable Marca que quedaron luego de la limpieza de datos:

Tabla 1. Número de personas según marca

Marca	Número de Personas
0 - Validado OK	39'935.216
10 - En Verificación	122.401

Fuente: doc 2

[5] La computación en clúster, o *high-performance computing* es una forma de computación en donde se divide la información a procesar en diferentes partes que son analizadas separadamente por diferentes computadores, lo que permite acelerar el tiempo de procesamiento.

Al momento de unir las bases de datos del SISBEN con aquellas de los programas sociales se encontraron inconsistencias entre los números de identificación de las personas en ambas bases de datos. Por ejemplo, 6'245.489 registros de los programas sociales tienen un número de identificación 0 - lo cual impide su cruce con los datos del SISBEN. Otros individuos cuentan con hasta 383 beneficios, lo cual debe por lo menos representar que el individuo aparece como receptor de beneficios de varias personas a través de los cortes de estudio; estos individuos podrían ser o no ser parte de su hogar, sin embargo, por la duplicidad de los ID no se puede realizar una unión con la base de datos del SISBEN para confirmar.

Adicionalmente, se anonimizaron los datos referentes a los nombres, apellidos, documento de identidad con el algoritmo SHA2 (Secure Hash Algorithm 2) y un hash de 256 Bits.

B. Estructura de la base de datos tipo panel

Para lograr el objetivo de encontrar los factores que son más susceptibles de presentar anomalías en las bases de datos recibidas, al igual que el objetivo de entender los factores socioeconómicos y de participación en programas sociales que se relacionan con la movilidad social, es necesario tener una estructura de datos que permita analizar el cambio de las diferentes variables en el tiempo. Es por esto, que una de las etapas principales del desarrollo del proyecto se concentró en la creación de una base de datos tipo panel, es decir, una base de datos que permita usar los datos socioeconómicos de los diferentes individuos a través del tiempo. Este ejercicio resulta importante, puesto que la base de datos original del SISBEN tiene los datos de cada corte por separado, impidiendo el tipo de análisis en el tiempo que se hace en el presente estudio.

La información recibida del SISBEN se organizó según las siguientes dimensiones[6]:

- Geográfica: barrio, municipio, zona y área.
- Temporal: mes, año, fecha completa, corte/trimestre
- Persona: datos de identificación de individuos.
- Vivienda: datos de identificación de la estructura donde las personas viven.
- Hogar: datos de identificación de los hogares que existen dentro de las viviendas[7].
- Unidad de gasto: datos por grupo de personas que dependen económicamente de alguien más dentro de un hogar.
- Programas: programas sociales a los que pertenece la unidad de análisis (persona, vivienda u hogar) y la entidad que brinda el programa social.

[6] Para más información sobre las variables de cada dimensión y el proceso de creación del panel de datos, ver Anexo A.

[7] La vivienda se diferencia del hogar en que la vivienda es la estructura física donde varios grupos de personas pueden establecer 1 o más hogares. En la vivienda se encuentra información sobre acceso a acueducto (infraestructura), mientras que en hogar se encuentra información sobre variables como el número de baños.

Creación de identificadores individuales por persona

Una de las limitaciones encontradas en la base de datos tipo panel es la falta de un número de identificación único que se mantenga en el tiempo para cada individuo. Esto se debe a que los números de identificación de las personas en el SISBEN son una combinación entre un identificador único para el hogar y un orden que se le da a los miembros del hogar. Es decir, en un hogar de 4 personas cuya cantidad de miembros pasa a 3, no es posible identificar de forma eficiente la correspondencia del identificador antiguo con el nuevo para cada miembro del hogar. Por ejemplo, no se sabe si el miembro 1 salió del hogar y si el miembro 2 pasó a ser el 1 o si el 3 pasó a ser el 1. Adicionalmente, el identificador depende del lugar de residencia, el cual cambia si la familia cambia de municipio. A su vez, el documento de identidad también cambia en el tiempo de tarjeta de identidad a cédula de ciudadanía, lo cual añade a las dificultades para la trazabilidad de individuos.

Como estrategia de mitigación se creó un identificador único, juntando los nombres, fecha de nacimiento y documento de identidad de cada individuo. En los casos en los que la persona cambió de documento de identidad en un periodo, se tomó el documento de identidad más reciente dentro del mismo periodo. La limitante que se mantiene en esta solución es que no es posible tener una identificación única para las personas que cambian de documento al mismo tiempo que lo hacen de residencia. Sin embargo, esta solución permitió pasar de 67 millones de registros únicos a alrededor de 46.5 millones. Una segunda revisión de duplicados dentro de los 46.5 millones llevó a depurar otras 3,800 personas que actualmente no están registradas como duplicados en el SISBEN, lo cual muestra el valor agregado de la creación de identificadores únicos por persona.

Finalmente, la siguiente tabla presenta un ejemplo de la base de datos que se conformó para el estudio:

Tabla 2. Ejemplo de base de datos tipo panel

ID	Dimension geograficas				Dimension hogar	Dimension tiempo			
id_hogar	**zona**	**area**	**depto**	**...**	**variable**	**201306**	**201312**	**...**	**201906**
13|001|100002|1	1	1	13		aire	2	2		2
13|001|100002|1	1	1	13		aire_changes	1	0		0
13|001|100002|1	1	1	13		auto1	2	2		2
13|001|100002|1	1	1	13		auto1_changes	1	0		0
13|001|100002|1	1	1	13		bieraices	2	2		2
13|001|100002|1	1	1	13		bieraices_changes	1	0		0
13|001|100002|1	1	1	13		calenta	2	2		2
13|001|100002|1	1	1	13		calenta_changes	1	0		0

Datos de la variable

Cambios de la variable en el tiempo.
1: Si cambio
0: No cambio

En ella se observa la estructura mencionada anteriormente. Las columnas dan información sobre el número de identificación del hogar, datos geográficos, las variables de la encuesta del SISBEN y los diferentes cortes. Las filas presentan los datos relevantes a una pregunta de la encuesta del SISBEN aplicada a un individuo, en un espacio geográfico determinado y a través del tiempo.

2. Modelamiento y análisis de datos

Para la detección de anomalías en la base de datos del SISBEN se usaron cuatro modelos de aprendizaje de máquinas y se analizó la capacidad predictiva de cada uno de ellos. Dichos modelos son la combinación de modelos supervisados, no supervisados, secuenciales y no secuenciales. A continuación, se presenta una explicación resumida de los tipos de modelo, seguidos de las características del modelo específico que se usó para el análisis. La explicación detallada del tipo y el modelo específico se encuentran en el *Anexo B*.

Modelos Supervisados: son aquellos que buscan encontrar la función que relaciona unas variables de entrada (*input*) con una variable de salida (*output*). Estos modelos se entrenan para usar las variables del SISBEN (*input*) y crear predicciones sobre la existencia de anomalías que ya fueron registradas por el SISBEN (*output*).

Modelos No supervisados: se concentran en analizar las relaciones que hay entre los datos, lo cual se puede usar para analizar si uno o más datos pertenecen a una agrupación de datos. En los modelos no supervisados se entiende que los datos que no pertenecen a un grupo pueden representar datos con anomalías.

Modelos No secuenciales: son modelos que usan los datos de diferentes unidades de análisis (persona, hogar o vivienda) sin tomar en cuenta sus cambios en el tiempo. Estos modelos usan los datos de un solo corte para predecir anomalías en el mismo corte o en cortes posteriores.

Modelos Secuenciales: toman en cuenta los cambios en los datos de las unidades de análisis en el tiempo. Con estos modelos el algoritmo se entrena para predecir anomalías a través del tiempo, haciendo uso de la base de datos tipo panel.

A continuación, se clasifican los modelos específicos empleados según su tipo:

Tabla 3. Tipo de modelos de aprendizaje de máquinas utilizados

	Supervisado	**No supervisado**
No secuencial	Gradient Boosting Machine (GBM)	Isolation Forest (IF)
Secuencial	Long-Short Term Memory (LSTM)	LSTM con Autoencoder

<u>Gradient Boosting Machine:</u> el GBM empleado es un modelo que usa árboles de decisión para encontrar una función que presi, dice anomalías a partir de las variables del SISBEN y las variables de contexto que se crearon para el piloto. Este tipo de variables es una de las formas en las que se genera valor para el piloto, pues se crea el promedio de variables socioeconómicas de los hogares en cada barrio y se compara con el promedio de cada hogar para detectar anomalías. Por ejemplo, una persona que afirma tener un electrodoméstico en un

barrio donde las personas tienen en promedio 10 electrodomésticos puede presentar anomalías.

Isolation Forest: es un algoritmo que usa árboles de decisión para segmentar los datos de acuerdo con un gran número de variables. Los datos que requieren menos segmentaciones para separarse de un grupo son considerados como anomalías. Se resalta que este modelo no usa las etiquetas de personas con anomalías (la variable 'marca' en la base de datos), ofreciendo una alternativa para la detección de anomalías que no se utiliza actualmente.

Long-Short Term Memory: el LSTM hace uso de la base de datos tipo panel para crear predicciones de anomalías a través del tiempo. Al ser supervisado, el modelo aprende sobre las características socioeconómicas de las personas que presentan anomalías en el SISBEN para crear predicciones de anomalías en otras personas, o de una misma persona en otro momento de tiempo.

LSTM con Autoencoder: el autoencoder tiene dos etapas, ambas implementadas con LSTM. En la primera etapa se busca reducir las dimensiones de los datos, creando una versión codificada de los mismos. En la segunda se busca usar la versión codificada de los datos para recrear la versión original. Dicho proceso permite crear aproximaciones de los datos originales que contienen menos ruido y valores extremos. Se le asigna valor de anomalía a aquellos datos que tienen valores extremos cuando se compara la versión aproximada con la original.

La Tabla 4. abajo describe la capacidad predictiva resultado de cada uno de estos modelos, incluyendo el área bajo la curva de cada modelo, su facilidad de implementación al igual que sus ventajas y desventajas.

Tabla 4. Resultados de los modelos de aprendizaje de máquinas

Modelo	**AUC**	**% de anomalías***	**Ventajas**	**Desventajas**
GBM	0.91	0.71	● No requiere transformación en el formato de entrada ● Las variables de contexto se pueden guardar calculadas	
Isolation Forest	0.61	0.5	● No depende las etiquetas, y puede ser utilizada com feature para la GBM	● Tiene un mayor número de falsos positivos
LSTM	0.92	0.72	● Tiene un mejor desempeño sin feature engineering	● Complejo crear la base de datos tipo panel ● Con esos volúmenes

				el entrenamiento es lento
Autoencoder	0.65	0.54	● Obtuvo lo mejores resultados sin etiquetas	● Complejo crear la base de datos tipo panel como entrada

*en 4 millones de personas

Los modelos que tuvieron mejor predicción fueron los supervisados, es decir el Gradient Boosting Machine (GBM) y el Long-Short Term Memory (LSTM), con AUCs de 91% y 92% respectivamente. Sin embargo, los consultores resaltan que la implementación del LSTM tiene mayor complejidad. Esto se debe a que para su uso se necesita una base de datos tipo panel y debido a que el DNP almacena los datos de cada corte por separado esta opción requiere más tiempo. Por ello, se recomienda el uso del GBM, que se puede implementar de forma más simple y tiene un AUC similar.

III. Resultados Cuantitativos y Cualitativos: Interpretación de los modelos a la luz de la problemática social

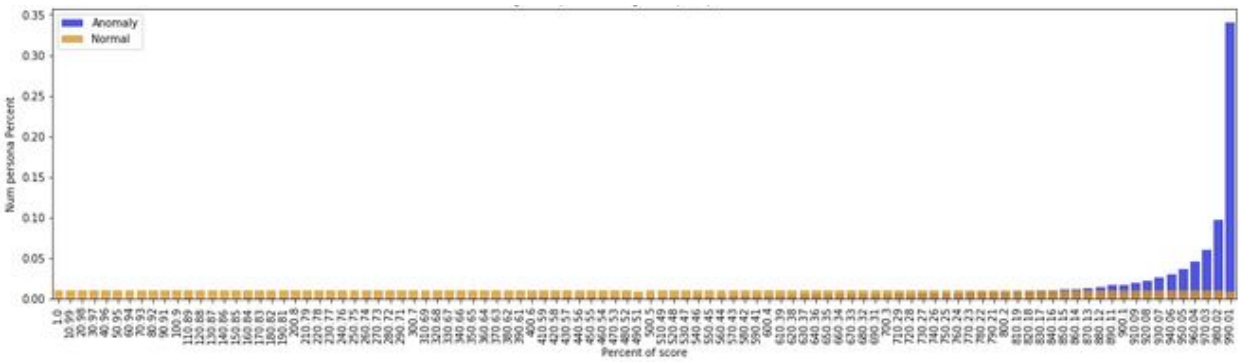
A continuación, se describe la interpretación de los resultados para cada uno de los objetivos del proyecto de analítica.

1. Establecer los grupos de personas que por sus características de comportamiento y condiciones socioeconómicas serían propensas a presentar anomalías en su puntaje del SISBEN.

A. Los resultados de la clasificación de anomalías

Los modelos de aprendizaje de máquinas permitieron crear predicciones sobre las personas con una mayor propensidad de tener anomalías. A continuación, se presentan los resultados de las predicciones provenientes del modelo GBM[8], el cual clasifica el riesgo de un individuo de presentar anomalías entre 0 y 1000; a mayor riesgo, mayor es la probabilidad de que el individuo presente anomalías. La figura 1 abajo, muestra la distribución de individuos de acuerdo con el riesgo que predijo el modelo; el eje x detalla el riesgo, y el eje y el porcentaje de personas que se encuentran por cada intervalo del eje x. En amarillo se muestra la distribución de riesgo de personas que no reportan anomalía en azul aquellas que tienen marcas de anomalía. En la gráfica se observa que las anomalías (azul) se concentran en los riesgos más altos, a partir de 850. Esto quiere decir que el modelo consigue asignarle un riesgo alto a las personas que presentan anomalías, lo cual demuestra la capacidad del modelo para clasificar correctamente a los individuos que pueden presentar anomalías.

Figura 1. Histograma - Distribución de personas según el riesgo de presentar anomalías

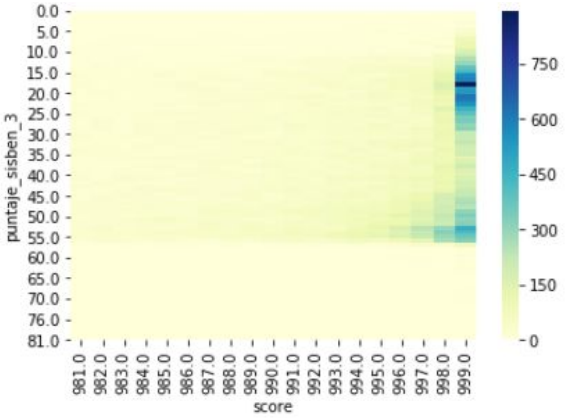


Con el objetivo de observar en qué niveles socioeconómicos se concentran las personas que presentan anomalías, se procede a caracterizar socioeconómicamente a los individuos según su puntaje en el SISBEN y sus marcas de anomalía. En la figura 2, se presenta la distribución de las anomalías, de acuerdo con el puntaje del SISBEN de las personas (eje Y) y su riesgo de

[8] Se selecciona el modelo GBM debido a que fue recomendado como el más apropiado para el DNP, ya que tiene un buen nivel de predicción (AUC de 91%), se puede usar con la estructura de datos por cortes individuales que maneja el DNP.

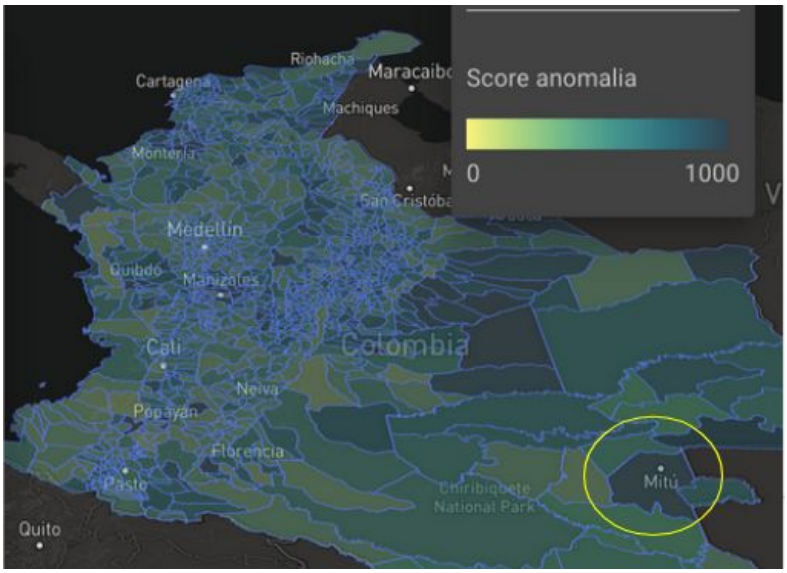
presentar anomalías (eje X). El color representa la cantidad de personas que hay en una coordenada (X,Y), entre más oscuro, más personas hay en dicha coordenada. Se observa que la mayoría de las anomalías se encuentran en el riesgo de anomalía más alto (el color azul se presenta solo al lado derecho de la gráfica), y en el rango de puntajes del SISBEN (eje Y) entre 10 y 30, seguido por puntajes entre 50 y 55 (donde el color es más oscuro). En general, la gráfica muestra que la mayoría de anomalías se presentan en los puntajes más bajos y por debajo de los puntos de corte que dan acceso a los programas sociales del gobierno[9].

Figura 2. Distribución de personas según el riesgo de presentar anomalías y puntaje del SISBEN



A continuación, se observa la distribución de las anomalías a nivel geográfico. Para ello se presenta un mapa de los municipios de Colombia, donde las áreas con mayor cantidad de anomalías tienen un color más oscuro. En el mapa, hacia el suroriente del país, se observa que Mitú (capital del Vaupés) tiene una mayor cantidad de anomalías.

Figura 3. Distribución geográfica de predicciones de anomalías



[9] Las personas con puntajes menores a 58 puntos tienen acceso a por lo menos uno de los programas sociales del gobierno.

El análisis de las distribuciones geográficas y socioeconómicas de las anomalías son útiles para que el DNP y los entes territoriales puedan verificar las anomalías a un menor costo, al priorizar este tipo de lugares. Por ejemplo, las estimaciones del presente estudio muestran que visitando el 1% de los hogares registrados en el SISBEN se podrían verificar el 70% de las anomalías encontradas.

B. Prueba de Concepto: la interoperabilidad entre las fuentes de datos públicas como potencial solución a la desactualización de la base de datos del SISBEN.

Para mostrar formas adicionales en las que la tecnología puede ayudar a resolver la problemática de identificar las anomalías en el SISBEN, y como prueba de concepto, se obtuvo la autorización del DNP para tomar una muestra de los datos del SISBEN y validarla con información de bases de datos públicas. Con este fin, se recibieron por parte del DNP registros sin anonimizar,[10] los cuales fueron anonimizados luego de haber cruzado una muestra de datos con la información de las fuentes de datos públicas[11]. Es importante mencionar que este proceso se realizó cumpliendo con los principios de proporcionalidad, recolectando la mínima información necesaria, respetando los derechos de privacidad de la información y respetando los términos y condiciones de las páginas web que se usaron para complementar los datos.

Los registros de la muestra del SISBEN se validaron con información de tres fuentes de datos: el Sistema General de Seguridad Social, que se usó para verificar si una persona se encontraba en el régimen subsidiado o contributivo de salud o si había fallecido. El Registro Único Nacional de Tránsito (RUNT), con el que se validó la posesión de vehículos por parte de los individuos seleccionados en la muestra. Finalmente, el Registro Único Empresarial y Social, con el que se verificó si una persona ejerce alguna actividad mercantil.

El realizar este tipo de validación de datos, permitió encontrar las anomalías que se describen en la tabla 5.

[10] La necesidad de la base de datos sin anonimizar se justifica por dos razones. Primero, para desarrollar esta prueba de concepto. Segundo, debido a la falta de un número de identificación de personas permanente en la base de datos, era la única forma viable de generar un registro único para poder crear la base de datos tipo panel que permitió dar respuesta a los objetivos del piloto.

[11] Adicionalmente, los datos sin anonimizar fueron fundamentales y necesarios para la creación de la base de datos tipo panel, la cual es fundamental para responder a los objetivos de encontrar la movilidad social y las anomalías en el tiempo.

Tabla 5. Resultados validación de datos para detección de anomalías

Caso	Número de personas	Porcentaje
Personas fallecidas marcadas como válidas	81 casos - muestra de 85,318	0.09%
Personas de régimen subsidiado marcadas como 'activo cotizante'	1,121 casos - muestra de 571.719	0.20%
Personas que dijeron no tener un vehículo que aparecen como propietarios de un vehículo en el RUNT	6,676 casos - muestra de 7,531	88.65%
Personas que aparecen como propietarias de un establecimiento de comercio	808 casos - muestra de 96,161	0.84%

Es de notar que el 88.65% de individuos en la muestra que dijeron no tener un vehículo en la encuesta del SISBEN, aparecen como propietarios de un vehículo en el RUNT. A pesar de que en los otros casos los porcentajes parecen bajos, es importante tener en cuenta que la identificación de pequeños porcentajes de anomalías pueden tener repercusiones importantes en la redistribución del gasto social. Como se mencionó anteriormente, si se detectan anomalías equivalentes al 0.1% del gasto social del 2018, se podría distribuir alrededor de 120 mil millones de pesos. Además, un beneficio para los ciudadanos es que no tendrían que pasar por el proceso administrativo de la actualización de datos, ni incurrir en los costos asociados (transporte por ejemplo).

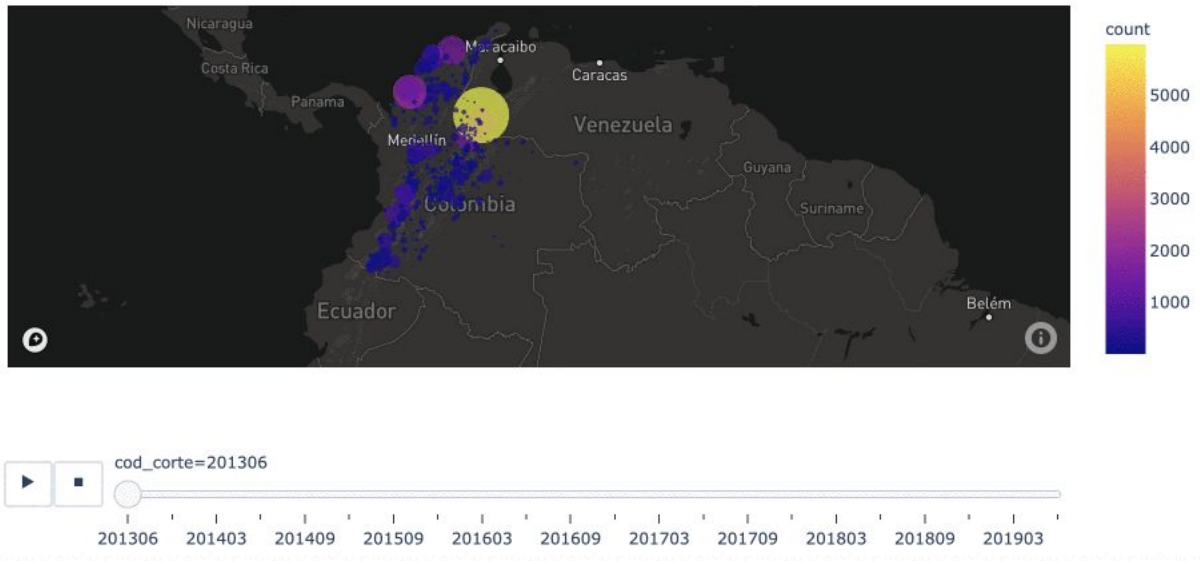
Unas de las consideraciones importantes para implementar programas que fomenten la interoperabilidad de datos del gobierno es que los cruces deben estar directamente relacionados con la información recolectada en la base de datos, para evitar el uso de aproximaciones que puedan brindar información equivocada sobre la situación socioeconómica de los miembros del SISBEN. A su vez, es importante tener en cuenta que estos registros públicos tienen sesgos, por ejemplo, la existencia de un registro mercantil no indica si una empresa está activa o si reporta utilidades. Finalmente, es crucial establecer un protocolo de manejo ético de datos para proteger la intimidad y la privacidad de los datos de las personas que suministran la información.

2. Identificar los factores que motivan el comportamiento de las personas registradas en la búsqueda de beneficios mediante la manipulación de la información.

Para este análisis se usan los datos tipo panel entre el 2013 y 2019 y se observan los cambios en las variables socioeconómicas a través del tiempo que llevan a disminuciones en el puntaje

del SISBEN, para personas marcadas con anomalías[12]. De esta manera se busca entender cuáles son los lugares y los cambios en las variables que con mayor frecuencia llevan a disminuir el puntaje del SISBEN de forma anómala. En la figura 4 abajo, los círculos más grandes y de color más claro representan una mayor cantidad de actualizaciones de datos que llevaron a disminuciones del puntaje. Es posible que los círculos de gran tamaño se den por actualizaciones sistemáticas de la base de datos que se dan en los territorios.

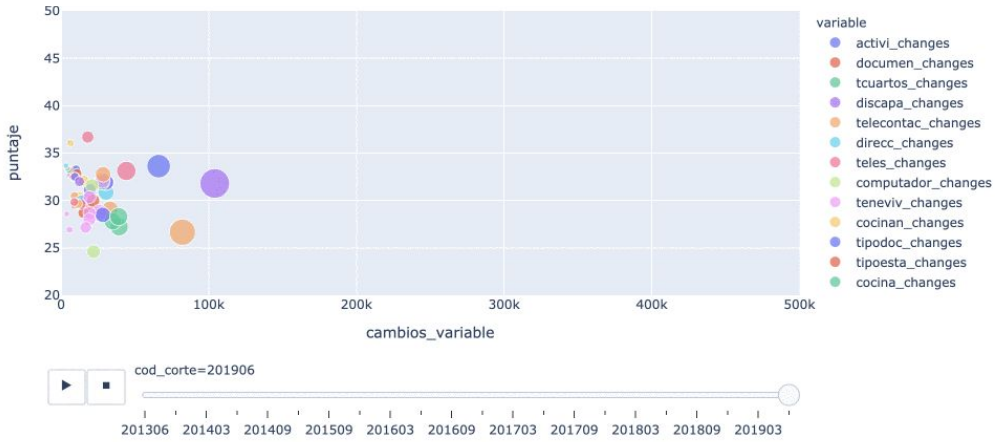
Figura 4. Cambios en las variables socioeconómicas entre el 2013 y el 2019



Similarmente, en la figura 5 se presenta la relación entre los cambios en las variables del SISBEN y el puntaje a través del tiempo para aquellos que tienen anomalías marcadas en el SISBEN. En el eje Y se observan los puntajes del SISBEN y en el eje X la cantidad de cambios que han tenido las variables desde Junio de 2013 hasta un momento dado. Este momento lo define el eje horizontal en la parte inferior de la gráfica, que muestra los cambios de los datos representados en la gráfica a través del tiempo. Conforme avanza el tiempo, los puntos de la gráfica que representan los cambios acumulados en las variables aumentan de tamaño.

[12] El análisis se dividió en las 3 áreas que maneja el DNP.

Figura 5. Cambios en las variables y puntajes del SISBEN entre el 2013 y el 2019



Por medio de este análisis se encuentra que las variables que más cambian cuando se reduce el puntaje son el número de personas en el hogar, la estructura del hogar, el uso de un sanitario compartido, la edad, el uso de la cocina compartida, el número de cuartos, la vivienda y la unidad de gasto.

Adicionalmente, se encuentra que las variables que llevan a reducciones de puntaje de las personas con anomalías difieren según el tipo de anomalía que registra el DNP. La tabla 6, resume las variables que más cambian por tipo de anomalía y área del DNP[13].

Tabla 6. Variables que más cambian por tipo de anomalía en el SISBEN

Estado de descripción			Area 1	Area 2	Area 3
0	Validado				
30	Caso 3	Denuncia	Edad, Extranjero, Grado, Nivel, Dirección	Embarazo, Carnet, Buscando, aparentes, parentes	Grado, nparentes, parentes
31	Caso 3	Denuncia	Tipo de documento, documento, asistencia	Serdomes, Hijosde	Hijosde, asiste
61	Caso 2	Cambio alcantarillado	Sanitar	Sanitar	Sanitar
62	Caso 2	Cambio acueducto	Llega	Llega	Llega
63	Caso 2	Cambio edad	Hijosde, tpersona,orden	tpoersona, tipodoc	tpersona, tipodoc
70	Caso 6	Documento no válido	Embarazo, documen	Documen, sexo	Documen, sexo

[13] Se recuerda que las áreas corresponden a las 14 ciudades principales (Área 1), el resto de áreas urbanas (Área 2) y las zonas rurales (Área 3).

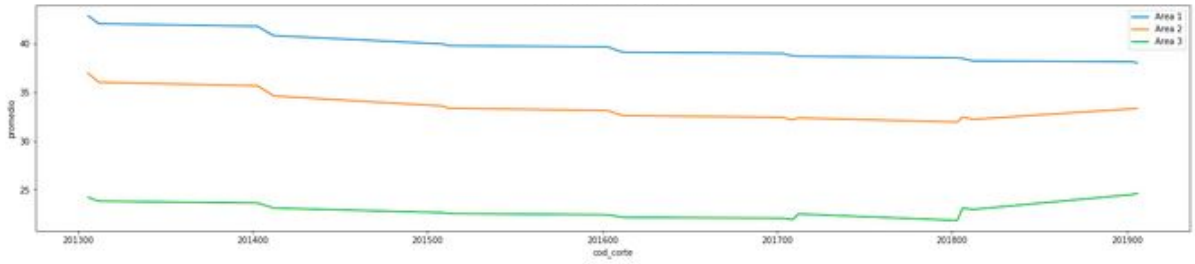
80	Caso 4	Reportado por UGPP	Nombre, apellido, fecha nacimiento	Nombre, apellido, fecha nacimiento	Buscando, tvcable, aire, equipo, horno
81	Caso 4	Cónyuge/hijo de persona reportada por UGPP	asiste, tipoesta	documen, tipodoc	Usotele, tipodoc
90	Caso 5	Cambio de tipo de vivienda último periodo	Cocina, usosani, preparan, usanitar	Usosani, tcuartos, usosani	Usosani, tcuartos
91	Caso 5	Cambio de tipo de vivienda cortes anteriores	Usosani, Cocina usanitar	usosani, cocina, tcuartos	Usosani, cocina, tcuartos

En la tabla 6 se observa que en las anomalías por denuncia con el número 30 en las 14 ciudades principales del país, la variable que más tiende a cambiar en las personas con anomalías es la de la edad. Es decir, que la edad de las personas difiere de la que se esperaría en un momento determinado según registros previos. Sin embargo, es de esperar que no todas las variables que cambian en las personas que tienen anomalías, sean las que más afectan su puntaje en el SISBEN. Una forma de priorizar la detección de anomalías es encontrar las variables que cuando cambian, tienen un mayor peso en los cambios de puntaje del SISBEN. El análisis de los modelos permite ver que las variables que más impacto tienen para reducir el puntaje son las de: sanitario, educación, carné, cómo se preparan los alimentos, tipo de sanitario y el número de electrodomésticos. Sería recomendable que el DNP pusiera medidas para detectar cambios en esas variables para mejorar la detección de anomalías.

3. Analizar tendencias y comportamientos encaminados a promover la movilidad social.

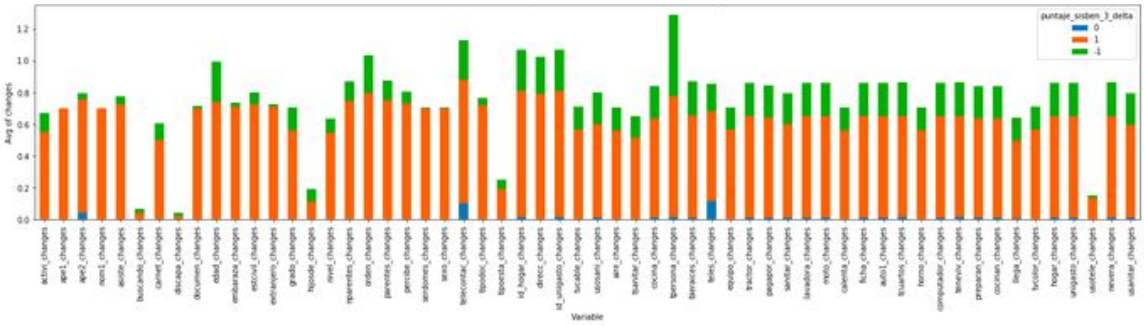
En cuanto a la movilidad social, se encontró que el puntaje promedio del SISBEN de las personas, separado por las áreas (14 ciudades, resto urbano y rural) ha disminuido en el tiempo (ver figura 6 abajo). Esto sugiere que la pobreza multidimensional ha aumentado en el tiempo, lo cual incrementa la importancia que tiene la detección de anomalías para poder priorizar el gasto público social. La siguiente gráfica presenta los cambios en la pobreza multidimensional en el tiempo, en azul para el área 1, naranja para el área 2 y verte para el área 3.

Figura 6. Promedio de puntaje del SISBEN por área entre el 2013 y el 2019



Para lograr identificar las características que se relacionan con la movilidad social, se observa la cantidad de veces que una variable de la encuesta SISBEN cambia en el tiempo y se clasifican estos cambios según se asocian con un incremento (1), una disminución (-1) o la ausencia de cambio (0) del puntaje del SISBEN del individuo. La siguiente gráfica da una representación visual de dichas frecuencias por variable. El supuesto detrás de este análisis es que las variables que más cambian en el tiempo y se relacionan con incrementos en el puntaje del SISBEN son aquellas que llevan a mejorar las condiciones de vida de las personas y por ende, llevan a la movilidad social.

Figura 7. Relationship of socioeconomic indicators and changes in SISBEN scores



A continuación se presenta una tabla que resume las combinaciones de variables que se relacionan con incrementos y disminuciones del puntaje del SISBEN. Este análisis se hizo a través de diagramas de categorías paralelas, en las que se puede identificar las variables que cambian con más frecuencia y se relacionan con los cambios en el puntaje del SISBEN. En la tabla se resumen los resultados del diagrama y se observa que la movilidad social se relaciona con cambios en las variables de educación, trabajo y salud. Por ejemplo, el aumento en el puntaje se relaciona con el paso de educación primaria a secundaria o el paso del régimen subsidiado al régimen contributivo de salud.

Tabla 7. Indicadores que más cambian en variables relacionadas con la movilidad social

Variable	Falta de movilidad social	Movilidad social
Educación (nivel)	Primaria/Secundaria	Secundaria
Salud (carnet)	Subsidiada/Ninguna	Contributiva

4. Identificar y caracterizar la(s) combinación(es) de programa(s) sociales más efectivos para que las personas en el SISBEN superen la condición de vulnerabilidad al igual que el umbral de puntaje para recibir ayuda de programas sociales.

Para analizar las combinaciones de programas que se relacionan con incrementos en la movilidad social, se unieron las bases de datos de 33 programas sociales con la base de datos del SISBEN. El trabajo realizado para crear un identificador único de las personas a través del tiempo ayudó a superar una parte importante de la limitación de las bases de datos para identificar individuos a través del tiempo. Este método permitió unir el 70% de los registros de los datos de los programas sociales con la base de datos del SISBEN.

Para entender la relación entre la participación en los programas sociales y la movilidad social, se analizaron diferentes combinaciones de programas sociales y se miró el porcentaje de personas cuyo puntaje aumentó comparado con el periodo anterior de acuerdo con su participación en programas. Los resultados se resumen en la siguiente tabla:

Tabla 8. Combinaciones de programas que más se relacionan con movilidad social

Combinación de programas	**Efectividad Área 1 (14 Ciudades)**	**Efectividad Área 2 (Otras ciudades)**	**Efectividad Área 3 (Rural)**
Mujeres ahorradoras en acción **+** Incentivo a la capacitación para el empleo**+** Más familias en acción	41%	49%	52%
Mujeres ahorradoras en acción **+** Capitalización empresarial **+** Más familias en acción	-	37%	37%
Mujeres ahorradoras en acción **+** Más familias en acción	35.43%	31%	31%
Incentivo a la capacitación para el empleo**+** Más familias en acción	30%	33%	32%
Incentivo a la capacitación para el empleo	14%	-	-

El análisis de la combinación de programas informa que en las tres áreas de análisis, la participación en los programas de 'Mujeres Ahorradoras en Acción', el 'Incentivo a la Capacitación para el Empleo' y 'Más Familias en Acción', tuvieron la mayor proporción de personas que aumentaron su puntaje del SISBEN. Vale recalcar, que los porcentajes hacen alusión al número de personas que aumentaron su puntaje en el SISBEN irrespectivo de la magnitud de este cambio.

En resumen, la tabla de conclusiones abajo resume los resultados principales encontrados en el marco de este proyecto[14].

- "Los datos de la base del SISBEN cuentan con errores en la captura y digitación de las encuestas, lo cual limita el cruce y conciliación entre los registros. Esto podría solucionarse con un identificador único para cada persona y al hacer uso de fuentes de datos públicas para validación de la información".

- "La falta de un ID único a nivel de persona hace que se pierda la trazabilidad cuando hay cambios de documento, hogar y municipio lo que limita el análisis en el tiempo y la capacidad de cruce con las bases de datos de programas".

- "Se encontraron 1.3 millones de beneficiarios en la base del DPS con puntajes mayores al punto de corte del programa en el que aparecen inscrito."

- "La creación de la base de datos tipo panel permite visualizar la trazabilidad en el tiempo de cada uno de los ciudadanos incluyendo la movilidad social. Lo cual no era posible observar ni medir a partir de la estructura de datos con tablas independientes por cortes."

- "Al analizarse los cambios de datos en el tiempo de la base de datos del SISBEN III, se pudo evidenciar que el 58.2% de las personas de la base no han tenido ningún cambio en las variables desde el inicio de la base de SISBEN III."

- "Al haberse creado una base de datos tipo panel con los cambios de hogares de una persona en el tiempo, se pudo concluir que variables como uso del sanitario, donde se preparan los alimentos el número de integrantes del hogar y las personas que lo componen son variables que presentan anomalías y cambios que no están siendo etiquetados por el modelo actual, pero que su variación tiene impacto para reducir el puntaje."

- "El modelo propuesto promete capturar el 70% de las anomalías visitando el 1% de las personas de la base, lo cual representa una relación de beneficio de 1 a 8. Esto quiere decir que invertir 1 peso en hacer la visita representa un ahorro de 8 pesos por errores de inclusión. Por lo anterior se recomienda hacer una prueba piloto controlada del modelo que permita validar el retorno esperado."

- "La capacidad de predicción del modelo se ve afectada cuando se entrena con todo el histórico de anomalías, ya que el modelo aprende de marcas de anomalías que luego

[14] Las conclusiones se pueden encontrar en el documento "Resúmen ejecutivo, Ramírez et al. 2020"

fueron depuradas por el SISBEN y no logra hacer predicciones sobre estas con datos futuros.”

- “Invertir en infraestructura tecnológica, incluyendo capacidad de almacenamiento y capacidad de cómputo que soporte plataformas big data resulta indispensable para migrar al uso de modelos de ML para el análisis de riesgos de anomalías. Esta inversión puede verse compensada con el incremento en la efectividad de la inversión del gasto público social.”

- “El número de registros y las variables medidas por la ficha de caracterización del SISBEN III son valiosas y suficientes para realizar modelos de predicción.”

- “Es necesario establecer un proceso de etiquetado de anomalías que informe del resultado de la verificación de la anomalía, lo cual permitiría que el modelo aprendiera sobre la validez de las etiquetas de anomalía y se ahorrarían esfuerzos en el entrenamiento de los modelos”.

- “La interoperabilidad de las bases públicas de información con la base del SISBEN reduciría la carga que tienen los ciudadanos de informar cambios a diferentes entidades públicas, así mismo mantener la información actualizada”.

IV. Retos y Recomendaciones

Con el propósito de identificar oportunidades de mejoría para proyectos de analítica de datos como este, a continuación se identifican retos y obstáculos enumerados tanto por los consultores que realizaron este proyecto piloto, así como por Data-Pop Alliance, organización que realizó el acompañamiento técnico a este proyecto.

Para empezar, se identificaron retos durante la entrega de las bases de datos, ya que los consultores recibieron una base de datos con información diferente a la que se tenía presupuestada. Adicionalmente, tras revisar los datos recibidos, se identifica que la información contenida no contenía el volumen o detalle necesario para llevar a cabo los modelos propuestos. De manera que se hace necesario realizar una segunda entrega de datos, donde se liberó el acceso a la base de datos entera del SISBEN a los consultores. Para prever este tipo de inconveniente, se recomienda especificar en los términos de referencia de proyectos como este las características específicas de los datos a compartir, o en lo posible, compartir una muestra los datos previos al inicio del proyecto. Con esto se busca asegurarse de que lo propuesto por los consultores se pueda realizar factiblemente con los datos a compartir, así como estimar las potenciales necesidades en la arquitectura de datos de los proyectos para optimizar los tiempos de procesamiento.

Por otro lado, durante el procesamiento inicial de los datos, los tiempos de procesamiento de los datos fueron más largos de lo previsto, principalmente por no contar con un identificador único por individuo u hogar en la base de datos. Realizar procesos de reingeniería de datos para lograr consolidar estos bajo un solo identificador implicó un poder computacional significativo, y tiempos de demora considerables.

De igual manera, a futuro se considera necesario pensar en la gestión del conocimiento generado a partir de proyecto pilotos y equipar a las entidades cuya información se está analizando con la infraestructura y capacidades necesarias para hacer transferencia de los conocimientos generados en el marco de estos pilotos. Durante este proceso, se hizo evidente el potencial valor agregado de utilizar modelos de aprendizaje de máquinas frecuentemente en la gestión y análisis de la información contenida en la base de datos del SISBEN.

Por último, se podría considerar incluir en los componentes de este tipo de proyecto, la necesidad de que los consultores sugieran futuros usos de los datos utilizados, puestos que son quienes se convierten a través del tiempo en expertos de negocio y datos y pueden ayudar al gobierno a identificar futuros proyectos.

Recomendaciones Específicas: Proyecto Piloto Detección de Anomalías en SISBEN

- La prueba de concepto realizada en el marco de este proyecto valida la importancia de asegurar la interoperabilidad de los sistemas y fuentes de información del estado. Para este proyecto, se comprobó que es posible validar y detectar anomalías de registros del SISBEN al cruzar diferentes registros de datos públicos.

- Respecto a este piloto en concreto, para proyectos con información como el del SISBEN IV, se recomienda implementar modelos de inteligencia artificial como la GBM aunado a la actualización de las reglas de marcas. De esta manera se podría capturar un mayor porcentaje de anomalías reduciendo el número de visitas de verificación y por ende el tiempo de respuesta. Respecto de las etiquetas, se sugiere incluir las variables a nivel de hogar identificadas en los capítulos anteriores. A medida que el SISBEN IV aumenta la desagregación territorial, es importante que esto se refleje en las marcas para no generar sesgos en cierto tipo de población tal como lo está causando la UGPP.

- Se considera valioso integrar datos de las poblaciones más vulnerables a la base de datos, tales como el registro único de víctimas para así garantizar prioridad en la asignación de beneficios a estas personas de especial protección.

- Se considera importante realizar un inventario de datos previo a iniciar con el análisis para determinar el mínimo de datos requerido para llevar a cabo los objetivos de los proyectos. Si se determinar que no hay necesidad de utilizar todos los datos recibidos, esto podría contribuir a respetar el derecho a la intimidad.

- A la luz de la base de datos de programas sociales del Departamento de Planeación Social, resulta importante priorizar la actualización de los datos, pero especialmente el documento de identidad. Esto ayudaría a evaluar los impactos de los programas en la población a través del tiempo. De igual manera, es determinante que los objetivos de cada programa puedan ser evaluados a la luz de los cambios de las variables en el tiempo, pues a la fecha, programas como familias en acción no permite evaluar la permanencia de una familia en el tiempo ni entender los cambios que esa familia ha tenido desde la fecha de ingreso hasta la fecha de última atención.

- Finalmente, es pertinente resaltar la relevancia que tiene el SISBEN no solo como herramienta de focalización de los potenciales beneficiarios de programas sociales, sino también como una herramienta de focalización de gasto público social que permita generar inversión en los lugares donde la población se encuentra en mayor nivel de vulnerabilidad por variables asociadas a la pobreza multidimensional como salubridad o educación.

Recomendaciones Generales: Proyectos de Analítica de Datos del Estado

- Para implementar este tipo de soluciones o proyectos de analítica se recomienda contar con disponibilidad de una infraestructura robusta (y un marco legal flexible que lo permita) para así poder llevar a cabo la analítica y la transferencia de los datos de una manera ágil y eficiente. El presente proyecto requirió aproximadamente 200 mil horas de procesamiento en máquinas de 250 GB de RAM y 64 procesadores en Amazon Web Services, por un costo de $10.000 USD. Aunque pudiera parecer alto, este costo es poco representativo si se tiene en cuenta los beneficios económicos de focalizar correctamente el gasto público.

- Para garantizar la interpretabilidad de los modelos, se recomienda reducir en la mayor medida posible el uso de algoritmos *black box.* Esto permite que la interpretación de los resultados sea más transparente, y también puede contribuir a reducir los sesgos en los modelos, y por consiguiente en los resultados.

- Es fundamental el trabajo conjunto y colaborativo entre las empresas con experiencia en implementación de AI y big data y las entidades públicas dueñas de la información. Esta metodología debe garantizar la transferencia de conocimiento al sector público, identificando puntos fuertes, cuellos de botella, y áreas a mejorar para que posteriormente las entidades del sector público puedan integrarlos a futuros modelos de analítica de datos que permita desarrollar proyectos generalizables y funcionales.

- Priorizar un enfoque de derechos en la implementación de este tipo de proyectos se hace absolutamente fundamental, teniendo como prioridad el respeto de los derechos de la población que está siendo analizada.

Anexo A. Creación de Panel de Datos

Los datos se recibieron en formato SQLServer y luego se transformaron a formato .TXT para facilitar su almacenamiento en Amazon S3. Para reducir el tiempo de lectura por cada corte se usó Apache Spark, que permite el procesamiento en clúster paralelo de la información, es decir, que se divide la información en partes que son analizadas separadamente por computadores y permiten acelerar el tiempo de procesamiento. Gracias al uso de Spart, el tiempo de lectura de la información pasó de 30 minutos a 1.5 minutos. A continuación, los datos se pasaron de formato .TXT a formato Parquet, un formato de código abierto que permite comprimir el tamaño del .TXT hasta 4 veces.

Los *Facts*, visibles en el centro de la gráfica encima, representan el conjunto de datos del SISBEN. Los recuadros amarillos con el nombre dimensiones (*dimensions)* son variables que permiten agrupar los datos del SISBEN presentes en la tabla de *Facts.*

Las dimensiones son:
Geográfica: barrio, municipio, zona y área.
Temporal: mes, año, fecha completa, corte/trimestre
Persona: datos de identificación de individuos anonimizados.
Hogar: datos de identificación de hogar anonimizados.
Vivienda: datos de identificación de vivienda anonimizados.
Unidad de gasto: datos de unidad de gasto anonimizados.
Programas: programas sociales a los que pertenece la unidad de análisis y su entidad.

A continuación, se menciona la información contenida en la tabla de facts agrupada según las diferentes dimensiones. Es importante tener en cuenta que dentro de cada dimensión se tiene información de las otras dimensiones. Por ejemplo, la información en la dimensión persona, también contiene datos sobre la dimensión temporal y geográfica en la que se analiza la información de la persona.

La información contenida en la tabla de *facts*, organizada por dimensión contiene:

Persona: actividad en último mes, asistencia a centro educativo, semanas buscando trabajo, carné (afiliado/a), discapacidad, embarazo o hijos, estado civil, grado, ingresos, nivel educativo, parentesco al jefe del núcleo, parentesco al jefe del hogar, puntaje sisben III, trabajo en servicio doméstico, tipo de establecimiento.

Hogar: tenencia de automóvil, tenencia de moto, bienes raíces, aire acondicionado, calentador, cocina, fuente de energía para cocinar, computador, cónyuge vive en el hogar, equipo de sonido, horno microondas, lavadora, agua, nevera, paga a otro miembro arriendo y alimentación, percibe ingresos, lugar de preparación de alimentos, sanitario, cuartos, hogar en arriendo o propio, total de personas, cantidad de sanitarios, cantidad de cuartos, tenencia de tv, tenencia de cable, sanitario y servicio telefónico.

Vivienda: acueducto, agua, alcantarillado, alumbrado, recolección de basura, cuántos días llega el agua, cuántas horas llega el agua, ducha, eliminación de basura, energía, estrato, gas, pared, piso, modo de recolección, suministro de agua, cuartos, habitaciones, cantidad de grupos de personas que cocinan por separado, tipo de vivienda.

Unidad de gasto: total de ingresos mensuales, jefe unidad de gasto.

Adicionalmente, a cada variable dentro de la tabla de *facts* se le adicionó otra columna donde se tiene el valor de la misma en el corte anterior, lo que se conoce como un *lag*. En cuanto a la forma de la base de datos tipo panel, se invirtieron las filas y columnas para que las columnas identifiquen el ID que se está manejando, por ejemplo a nivel de persona u hogar, la variable del SISBEN y enseguida las fechas respectivas a un corte. Cada fila identifica entonces a un tipo de ID, en alguna variable en específico con sus respectivos cambios en esa variable a través del tiempo.

Anexo B. Detalle de modelos de Aprendizaje de Máquinas

Modelos supervisados no secuenciales

Los modelos de aprendizaje supervisados son aquellos usados para encontrar la función que explica la relación entre un grupo de variables de entrada y una o más variables de salida. Esto con el fin de crear predicciones de la variable de salida a partir de los datos de las variables de entrada. Para el uso de estos modelos cada observación de entrada cuenta con un valor correspondiente en la variable de salida. Debido a que se cuenta con valores para la variable de salida, se pueden comparar las predicciones que hace la función que se calcula con los valores reales de salida. La diferencia entre la predicción realizada por el modelo y el dato real permite dar cuenta de la magnitud del error en la predicción del modelo. El objetivo de estos modelos es optimizar esta predicción, es decir llevar el error de predicción a su mínimo. Esto se puede hacer calculando las derivadas parciales de la función que calcula el error de las predicciones respecto a los parámetros del modelo predictivo que se está entrenando. Las derivadas parciales del error con respecto a los diferentes parámetros permiten saber si un parámetro está haciendo crecer o disminuir el error. Con esta información se modifican los parámetros para que lleven el error a su mínimo. Si el parámetro hace que crezca el error, entonces se disminuye el valor del parámetro, en el caso contrario se aumenta el valor del parámetro hasta que el error se minimiza. Una vez se optimiza la función de error, otros datos de entrada con las mismas variables se computan con la función que se encontró para crear predicciones de la variable de salida.

Detección de anomalías con Gradient Boosting Machine (GMB)

El modelo de Gradient Boosting Machine usado, consiste en crear árboles de decisión, que clasifican los datos en los grupos de interés. En este caso, según la capacidad que tienen diferentes categorías de variables de entrada de clasificar si un individuo presenta una anomalía o no. La característica especial de este modelo es que cada árbol que se crea en las diferentes iteraciones trata de clasificar el error del anterior modelo, en lugar de clasificar la variable de salida. Es decir, se enfoca en clasificar lo que el modelo anterior no entendió, en lugar de mejorar la clasificación de lo que el modelo anterior sí entendió. Este proceso permite mejorar la capacidad de clasificación del modelo.

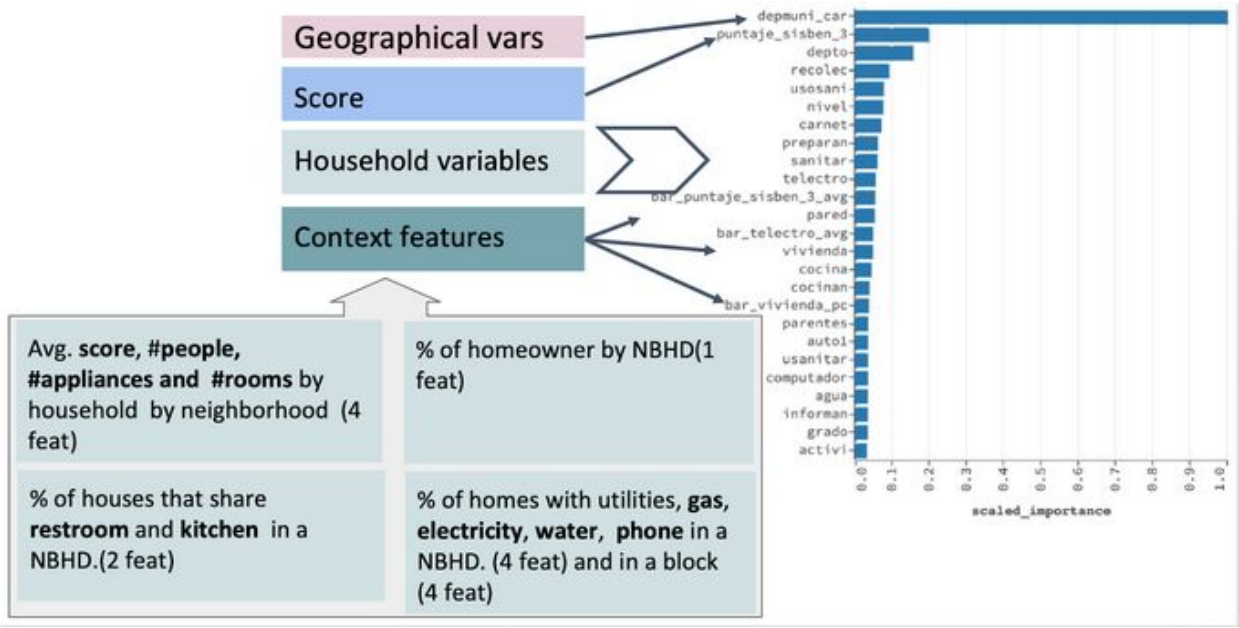
Para seleccionar las variables a usar en el modelo se hizo una regresión lineal del puntaje del SISBEN en las variables de la encuesta y se escogieron aquellas variables que tenían un mayor peso para predecir el puntaje del SISBEN. Es decir, se eligieron aquellas que, de ser modificadas, tienen un mayor efecto en el cambio del puntaje de la persona. Dichas variables se usaron de tres maneras para la detección de anomalías. Primero, se usaron en su forma original, como datos crudos de la base de datos. Por ejemplo, se usaron variables como la cantidad de baños que tiene un hogar. Segundo, se desarrollaron indicadores simples a partir de los datos crudos. Por ejemplo, a partir de variables que describen el número de baños que tiene un hogar, se crea una variable que describe si la persona tiene o no baños en el hogar. Finalmente, se crearon variables que contenían los promedios de las variables de cada hogar a nivel de barrio. A estas últimas variables se les llamó variables de contexto. La idea detrás de

estas variables es el supuesto de la similaridad de características socioeconómicas por barrio. Se espera que las características socioeconómicas de cada hogar no difieran mucho del promedio del barrio en el que se encuentran. La hipótesis es que si las variables de contexto son importantes en la predicción de anomalías, al incluirlas en el GBM se debería observar un incremento en la capacidad predictiva del modelo.

Resultados GBM

En la siguiente gráfica se describe la importancia que el GBM dio a cada variable usada, de acuerdo con su aporte a la predicción de la existencia de una anomalía para cada individuo.

Figura 8. Importancia de las variables en el modelo GBM



De acuerdo con esta información, la variable con más importancia para predecir la existencia de anomalías es el municipio, seguida del puntaje del SISBEN. Para llegar a este resultado, se realizaron 36 combinaciones de los siguientes hiperparametros:

Tabla 9. Hiperparámetros usados en el entrenamiento del GBM

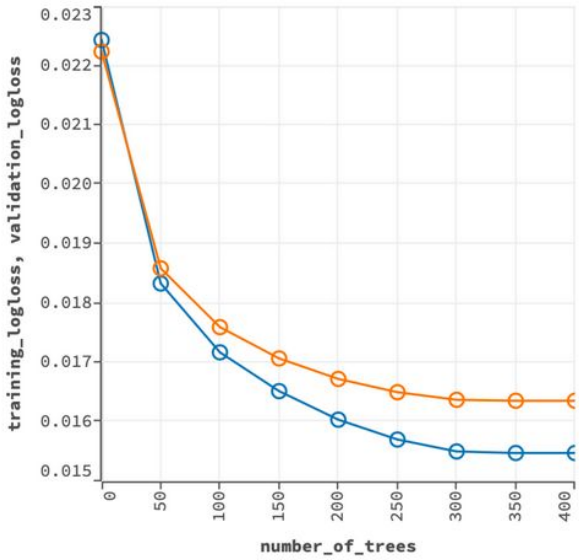
Variable	Search Value
Max_depth (profundidad máxima)	[10,8,5]
Learn_rate (tasa de aprendizaje)	[0.1, 0.02, 0.01]
Min_rows (mínimo de filas)	[15, 30]
Col_sample_rate_per_tree (tasa de muestra en columnas por árbol)	[0.7, 0.9]

Donde Max depth (Profundidad Máxima), describe la profundidad que se le permite tener a cada árbol. Es decir, el número de divisiones que el árbol puede hacer para clasificar los datos. Learn rate (Tasa de Aprendizaje), describe el cambio que se le permite tener a los parámetros de clasificación a medida que el modelo corrige sus errores. Una tasa más baja permite que el modelo cambie sus parámetros en menor proporción, lo cual puede ser conveniente ya que una tasa muy alta puede evitar que el parámetro adopte el valor que se minimiza el error[15]. Min_rows (Mínimo de filas), corresponde al mínimo de observaciones que debe tener cada hoja del árbol. Es decir, cuando el algoritmo llega a este número de observaciones en una clasificación, el árbol puede parar de clasificarlas. Finalmente, el Col_Sample_Rate_per_Tree (Tasa de muestreo por árbol), es una medida de la cantidad de ramificaciones del árbol que se usan para hacer subclasificaciones de la muestra.

Las mejores predicciones se alcanzaron con la profundidad de 10, la tasa de aprendizaje de 0.001, un mínimo de 30 observaciones por cada hoja y una muestra del 70% de las columnas de cada árbol que se consideran para hacer las clasificaciones.

En el siguiente gráfico se observa el cambio en los errores de predicción a medida que se incrementa el número de árboles. Es de notar que ambas curvas tienen un comportamiento similar, lo cual permite evidenciar la estabilidad del modelo.

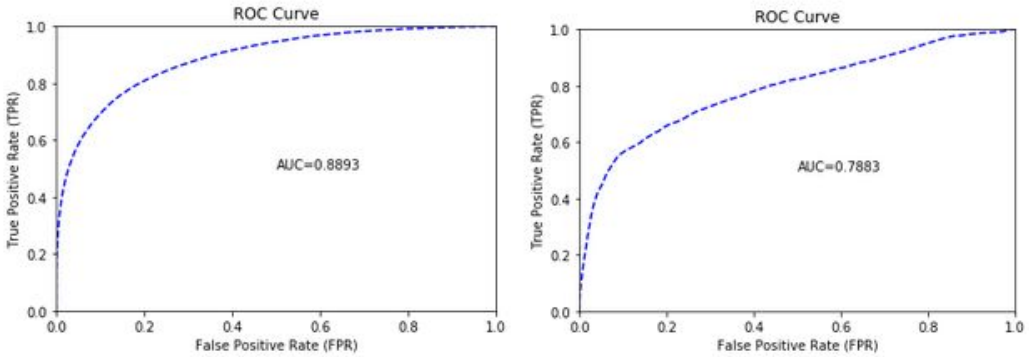
Figura 9. Errores de predicción de Isolation forest para datos de entrenamiento (azul) y de prueba (naranja)



[15] Por ejemplo, si un parámetro que tiene un valor de 2, minimiza el error cuando su valor es 4, pero la tasa de aprendizaje hace que el parámetro cambie en 5 unidades, cuando se modifica el parámetro para reducir el error este pasaría de 2 a 7. En este caso, la tasa de aprendizaje no permite llegar al punto mínimo. Por el contrario, una tasa que lleva a cambios de una unidad, permitiría que el parámetro pasará de 2 a 3 y que finalmente llegase a 4, minimizando así el error de clasificación.

Adicionalmente, cuando se entrena el modelo con datos de diciembre de 2017 y se usa para predecir las anomalías de diciembre de 2018, se logra un área bajo la curva del 89% (gráfica a la izquierda). Adicionalmente, se usa el mismo modelo para predecir las anomalías por denuncia y se logra un área bajo la curva del 79% (gráfica a la derecha). La importancia de las predicciones de anomalía por denuncia viene de que son anomalías que provienen de fuentes externas a los cambios en los datos por actualización, como es el caso de otras anomalías. Es decir, son anomalías que en cierta manera son exógenas a los cambios en la base de datos del SISBEN.

Figura 10. Área bajo la curva de Isolation Forest, predicciones de anomalías en 2019 (izquierda) y predicciones para anomalías de denuncia en 2019 (derecha)



Modelos no supervisados, no secuenciales

Los modelos de aprendizaje no supervisado buscan encontrar funciones que describen las relaciones que hay entre los datos. Estos modelos se diferencian de los supervisados en que estos últimos se enfocan en aproximar la función que existe entre unos datos de entrada y un dato de salida, mientras que los no supervisados buscan encontrar patrones en las distancias que hay entre los datos, al igual que en las densidades y/o distribuciones de los datos. Por ejemplo, estos modelos se pueden usar para encontrar agrupaciones en los datos y en consecuencia, encontrar los datos que no hacen parte de las agrupaciones de datos. Para el piloto de analítica, el enfoque está en encontrar los datos que no pertenecen a los grupos, ya que estos pueden representar anomalías que no han sido marcadas como tal por el DNP.

Isolation Forest

El isolation forest es un modelo de aprendizaje no supervisado que usa árboles de decisión para crear grupos entre los datos; este se utilizó para identificar anomalías sin hacer uso de la variable ‘marca’. El modelo toma una variable de la base de datos y hace una partición de las observaciones que contiene de forma aleatoria. Luego, examina cómo se agrupan los datos a cada lado de la partición. Posteriormente, toma otra variable, realiza una subdivisión y examina la nueva agrupación de datos. El algoritmo repite el mismo proceso y crea varios árboles que aleatoriamente dividen los datos en grupos. La parte de *Isolation* o aislamiento viene del hecho

de que el objetivo es lograr aislar las observaciones con las divisiones aleatorias. En este proceso, los datos que están más alejados de los otros, tienden -en promedio- a quedar aislados de los otros datos más rápidamente. Es decir, los árboles requieren menos ramificaciones -o segmentaciones dentro de las variables- para que un dato quede aislado de los demás. Los datos que se aíslan rápidamente, son entendidos como datos que presentan anomalías.

En el contexto del piloto, el modelo de Isolation Forest (IF) es ideal, puesto que permite analizar de manera simultánea -en un espacio multidimensional- si los datos de algún individuo son diferentes a los de otras personas; tomando en cuenta características que no son consideradas actualmente por el DNP. Además, evita entrenar un modelo con el desbalance de clases presente en la base de datos del SISBEN, en el cual solo el 0.2% de la base de datos está identificada con anomalías[16]. El modelo no supervisado permite entonces no hacer uso de la etiqueta de anomalía y observar diferentes variables simultáneamente en las cuáles los datos de una persona pueden ser anómalos, en comparación con los de otros individuos.

Resultados Isolation Forest

La siguiente tabla resume el promedio de ramificaciones que se necesitaron para separar un porcentaje de datos del resto, agrupado según la etiqueta de anomalía (1) o ausencia de anomalía (0). De acuerdo con lo esperado, el promedio de ramificaciones necesarias para aislar las observaciones con anomalía (13.46) es menor que el de las observaciones sin anomalía (13.48), a pesar de que la diferencia en magnitud no es muy grande. Sin embargo, el área bajo la curva de este modelo (61%) muestra que el modelo no tiene un nivel de predicción alto, pero sus predicciones pueden servir para mejorar el entrenamiento de modelos supervisados.

Tabla 10. Número de ramificaciones del Isolation Forest según tipo de anomalía

Label	**Avg (Tree Length)**
0	13.480757
1	13.462847

Modelos Supervisados Secuenciales

Los modelos usados previamente se concentraron en detectar anomalías a través de la comparación entre individuos, sin tomar en cuenta los cambios de un mismo individuo en el tiempo. Para tener en cuenta estos cambios, se usan los modelos secuenciales, que permiten tomar en cuenta la dimensión temporal los datos para la detección de anomalías. Esta sección

[16] Superar esta limitación es importante ya que usar modelos supervisados con tan solo el 0.2% de la población para aprender las características relacionadas con la existencia de anomalías crea el riesgo de relacionar las anomalías sistemáticamente con un segmento reducido de la población e ignorar otras anomalías que el DNP no tiene en cuenta actualmente.

del análisis permite identificar anomalías relacionadas con cambios en las respuestas en la encuesta del SISBEN.

Long-Short-Term Memory (LSTM)

El algoritmo LSTM está diseñado para encontrar la función que describe la relación entre las variables de entrada y de salida para individuos a través del tiempo. La particularidad de este algoritmo es que se entrena para encontrar un balance entre la importancia que se le da a los datos que ocurrieron en el pasado y aquellos que ocurren en el presente. Esto permite que el modelo use la información que es más pertinente para realizar predicciones y que deje de usar aquella que deja de serlo en el tiempo.

Para la implementación de este modelo se usan los datos crudos, a diferencia del modelo de GBM donde se complementan los datos crudos con indicadores simples y variables de contexto. Se incluyen 44 variables, a través de 11 cortes de tiempo. El LSTM tiene dos capas y neuronas con la forma (100, 20), es decir 100 filas por 20 columnas.

El área bajo la curva de este modelo llegó al 92%, convirtiéndolo en el modelo con mejor capacidad predictiva.

Modelos no supervisados, secuenciales

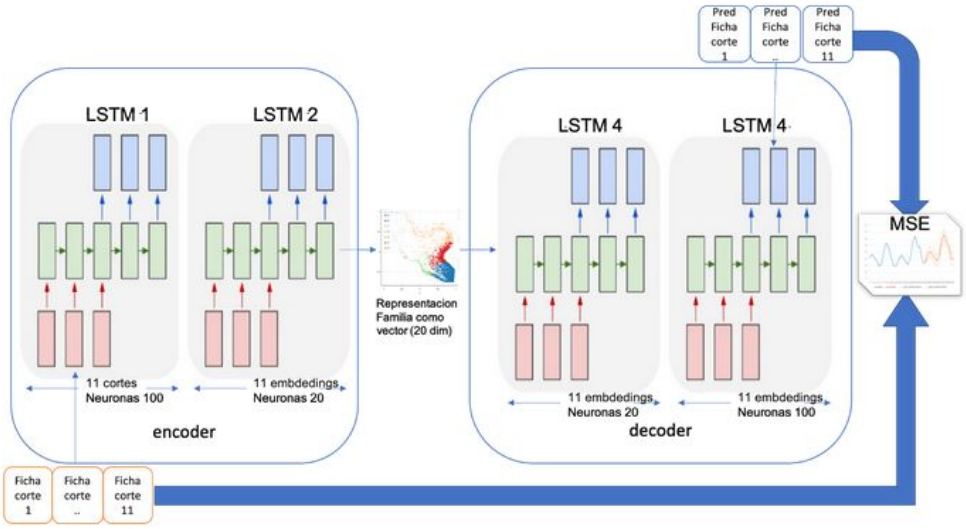
Debido a las limitaciones que ya se discutieron en los datos en cuanto a la baja cantidad de datos marcados con la existencia de una anomalía, se implementan modelos que utilizan los beneficios de los modelos secuenciales sin hacer uso de los datos etiquetados como anomalía.

Autoencoder con LSTM

El “autoencoder” es un modelo de aprendizaje no supervisado con dos etapas. En la primera etapa, el algoritmo aprende una función que permite encontrar una forma reducida de los datos que se observan. En la segunda etapa, el algoritmo se usa para encontrar una función que usa la forma reducida de los datos para reconstruirlos a su estado original. La reconstrucción de los datos a partir de la forma reducida no brinda una representación perfecta del estado original de los datos, pero permite reducir el ruido que se encuentra en los datos, o en otras palabras, ignorar los valores en los datos que se alejan de la normalidad. En el contexto de este piloto, al comparar las predicciones del autoencoder con los datos originales, se pueden encontrar aquellos valores que se alejan de lo normal y pueden representar una anomalía.

La implementación del modelo está representada en la siguiente gráfica:

Figura 11. Arquitectura del Autoencoder con LSTM



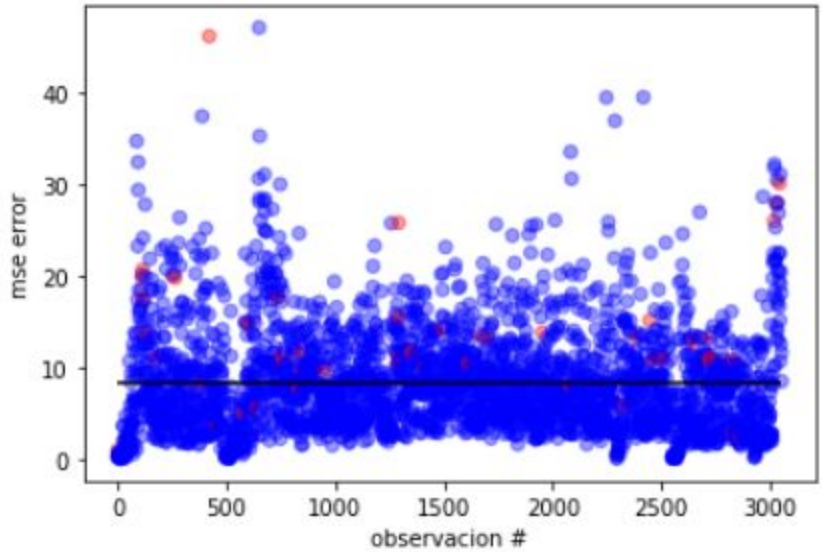
El cuadro de la izquierda representa la primera etapa del *autoencoder*, que recibe el nombre de codificador (*encoder*). El objetivo es crear una representación reducida de los datos. Para ello, el LSTM 1 toma 45 variables a nivel de familia de alrededor de 40 millones de personas, a través de 11 cortes en el tiempo y convierte los datos de cada familia en un vector de 100 elementos. El LSTM 2, usa los datos de salida del LSTM 1 y los reduce a 20 elementos por familia.

El cuadro de la derecha representa la segunda etapa del *autoencoder*, la del decodificador (*decoder*). En esta etapa el LSTM 3 usa las representaciones reducidas de las familias con 20 elementos y las convierte en 20 elementos por cada uno de los 11 cortes. Luego, el LSTM 4 usa los 20 elementos de cada corte y los convierte en 40 elementos a través de los 11 cortes.

Finalmente, se comparan las predicciones del decodificador con los datos originales y se calcula el error con la Raíz del Error Cuadrático Medio (Root Mean Squared Error, o RMSE en inglés). El error es de reconstrucción es de 2.44, lo cual significa que los datos no tienen un error estándar muy grande con respecto del promedio real y que el modelo logra aproximar de forma satisfactoria los datos originales. A continuación, para confirmar la robustez del modelo, se comparan los errores al cuadrado (Mean Squared Error) de las variables que están etiquetadas como anomalías (13.07) y las que no (8.26), lo cual indica que los datos con anomalías se alejan más de los datos que se consideran normales según el autoencoder.

En la figura 12, se representan las observaciones y su MSE. En rojo se observan aquellas etiquetadas como anomalías y en azul las que no tienen dicha etiqueta. La línea horizontal representa el MSE promedio, que es cercano a 9.

Figura 12. Errores de predicción del Autoencoder con LSTM, según marca de anomalía


mse error
observacion #
0
10
20
30
40
0
500
1000
1500
2000
2500
3000

Anexo C. Etiquetas de anomalías en base de datos SISBEN

Tabla 11. Tipos de etiqueta de anomalía en la base de datos del SISBEN

Etiqueta	Descripción
10	Excluido - Fallecido
11	Excluido por solicitud de retiro de ficha del SISBEN
30	En verificación CASO 3 - Por denuncia
31	En verificación CASO 3 - Por denuncia 2
42	Excluido por calidad
43	Excluido por calidad - fecha anterior 01/01/1890
44	Excluido por calidad - fecha posterior fecha final plataforma envío
46	Excluido por calidad - primer nombre y primer apellido inconsistente
60	En verificación CASO 2 - Cambio zona
61	En verificación CASO 2 - Cambio alcantarillado
62	En verificación CASO 2 - Cambio acueducto
63	En verificación CASO 2 - Cambio edad y calidad en datos
70	En verificación CASO 6 - Documento no válido
80	En verificación CASO 6 - Documento no válido
81	En verificación CASO 4 - Reportado por UGPP
90	En verificación CASO 5 - cambio de tipo de viviendo último periodo
91	En verificación CASE 5 - cambio de tipo de vivienda cortes anteriores