



**Un análisis de la pobreza en Colombia basado en aprendizaje automático**


Hermes Sabogal, Olmer García-Bedoya, Oscar M. Granados

Universidad Jorge Tadeo Lozano, Bogotá, Colombia

**Author Note**

Hermes Sabogal  <https://orcid.org/0000-0002-2790-0282>

Olmer Garcia-Bedoya  <https://orcid.org/0000-0002-6964-3034>

Oscar M. Granados  <https://orcid.org/0000-0002-4992-8972>

## Resumen

El artículo analiza la pobreza en Colombia utilizando herramientas de aprendizaje automático supervisado a partir de los datos de Hogares, Personas y Vivienda del DANE para el periodo 2016 a 2019. Se examina la percepción de factores que influyen en la pobreza teniendo en cuenta las especificidades estructurales que conforman la medición de la pobreza, como la salud, el trabajo y la educación. El aporte de esta investigación es comparar el Índice de pobreza multidimensional con los factores relevantes de la situación de pobreza mediante el uso de herramientas aprendizaje automático. Los hallazgos revelan que el algoritmo XGBoost identifica los indicadores que causan la pobreza y permite proponer un marco de trabajo para lucha contra la pobreza.

*Palabras Clave:* Aprendizaje automático, Medición y análisis de la pobreza, construcción de modelos y estimación, cambios tecnológicos.

**Abstract.** The article analyzes poverty in Colombia using supervised machine learning tools from DANE's data of Households, People, and Housing for 2016 to 2019. The article examines the perception of factors that influence poverty, considering the structural specificities that make up the poverty measurement, such as health, work, and education. The contribution of this research is to compare the Multidimensional Poverty Index with the factors that are relevant to the poverty situation using machine learning tools. The findings reveal that the XGBoost algorithm identifies indicators that cause poverty and allows proposing a framework to fight poverty.

*Keywords:* Machine Learning, Measurement and Analysis of Poverty, Computational Techniques; Simulation Modeling, Model Construction and Estimation, Technological Change: Others

*JEL Classification:* I32, C63, C51, O39

## Un análisis de la pobreza en Colombia basado en aprendizaje automático

### Introducción

La pobreza es una situación en la cual no es posible satisfacer las necesidades básicas de una persona o un grupo de personas como la alimentación, la vivienda, la educación, la salud, la vestimenta o el acceso a servicios públicos como el agua potable y, en algunos casos, a la electricidad. Esta no satisfacción se da por no contar con los medios económicos que habitualmente son el resultado de no tener acceso a un trabajo formal o a los medios para desplazarse hacia su lugar de trabajo. Adicionalmente, la pobreza también ha sido el resultado de la exclusión social, la marginación racial o religiosa, así como por conflictos armados, que afectan la posibilidad de desarrollar las actividades cotidianas o acceder a una alimentación diaria. De esta forma, la pobreza se expresa en diferentes dimensiones, al punto de identificarse como pobreza extrema cuando no se tiene acceso ni siquiera a un mínimo de productos básicos necesarios para el desarrollo físico y mental. Para 2015, el Banco Mundial estableció la línea de pobreza extrema en *USD*1,90 diarios y la línea de pobreza en *USD*3,10, con esto la pobreza se acerca a factores en los que se afecta la vida digna de cualquier ser humano.

El análisis de la pobreza ha tenido diferentes aproximaciones que establecen la complejidad para entender este fenómeno y, por ende, buscar mediciones acertadas que permitan considerar alternativas de política pública, pues no son procesos exclusivos de las herramientas de medición sino también de sus contextos. En ese sentido, la literatura ha abordado el problema de la pobreza como un tema de desarrollo social y económico (Sen, 1983, 1999), de gobernanza y política pública (Arestis & Caner, 2010; Bastiaensen y col., 2005; Collins, 2012; Epstein & Gang, 2009; Grindle, 2004), de violencia y conflictos (Gates y col., 2012; Hegre y col., 2009; Sen, 2008), así como de su perspectiva multidimensional (Alkire & Santos, 2013; Anand & Sen, 1997; Atkinson & Bourguignon, 1982) para enumerar algunos. Frente al caso colombiano, la pobreza ha sido analizada de igual forma con diferentes metodologías (Bahamón y col., 2013; Ramírez y col., 2017; Salazar y col.,

2011; Sánchez Torres y col., 2020) y varias de ellas incorporando su impacto en posibles efectos en los articuladores del crecimiento económico (Sánchez Torres, 2015), aunque en ocasiones con algunos resultados de disminución de la pobreza multidimensional (Salazar y col., 2011), así como con aproximaciones locales o regionales (Espinosa-Espinosa y col., 2020). Sin embargo, la pobreza es una variable que cada vez requiere una mayor cantidad de elementos para ser verificada de una manera más amplia, donde se requiere la integración de metodologías que permitan fortalecer su medición y verificación de parámetros para ayudar a direccionar las decisiones de política pública, pues los factores que influyen van desde las condiciones educativas del hogar, las condiciones de la niñez y juventud, hasta la salud y el acceso a servicios públicos. Es decir, la política pública se debe articular desde diferentes vertientes para lograr resultados efectivos.

El propósito de este documento es incorporar algunos modelos de aprendizaje automático al análisis de la pobreza multidimensional para explicar desde otra perspectiva los factores de este fenómeno en Colombia. Adicionalmente, poder determinar los posibles factores que influyen en los principales indicadores socioeconómicos que se integran en la evaluación de la pobreza. El presente documento propone una nueva posibilidad para entender la pobreza multidimensional a través de modelos de aprendizaje automático en la que se identifique las condiciones más significativas para cada una de las dimensiones analizadas. En este caso y de manera experimental, se busca detectar comportamientos en los factores que implican las casuísticas más relevantes entre los años 2016 y 2019, que permitan mostrar resultados socio-económicos que conlleven efectos de cambio significativos en áreas rurales y urbanas, basándose en la información de la Medición de Pobreza Monetaria y Desigualdad del DANE.

Para lograr esto, el documento se divide de la siguiente forma. En una primera sección se revisan los datos y se hacen algunas aproximaciones y análisis de su estructura, adicionalmente, se plantea la metodología de aprendizaje automático usada. En una siguiente sección se presentan y discuten los resultados. Finalmente, se presenta una

sección de conclusiones y el direccionamiento del trabajo futuro.

### **Conjuntos de Datos Y Métodos**

La información utilizada fue extraída del sitio de microdatos del DANE, los cuales miden la pobreza con el Índice de Pobreza Multidimensional. Este índice está basado en la metodología Alkire-Foster (Alkire & Foster, 2011) que ha sido implementada por el DANE y un centenar de estudios e instituciones en diferentes lugares del mundo. En el caso de Colombia mide la pobreza basada en cinco dimensiones: condiciones del hogar, condiciones de la niñez y juventud, salud, trabajo y acceso a servicios públicos domiciliarios y condiciones de la vivienda. Esta parametrización definida a partir de sistemas expertos establece la caracterización de la pobreza multidimensional.

#### **Estructura de los datos**

Los conjuntos de datos para la medición de la pobreza multidimensional consisten en tres grupos definidos por el DANE que son persona, hogar y vivienda. La integración de la información se hizo desde el motor de base de datos SQL Server con el fin de unificar cada uno de los grupos, pues los conjuntos de datos se encontraban divididos. Al realizar la exploración de cada uno de los conjuntos de datos, se identificaron las siguientes estructuras. Para el caso de hogares, se cuenta con 215.019 registros y una estructura fundamentada en 29 variables predictoras (Tabla A1), de las cuales tres son numéricas, nueve Enteras categóricas, 16 binarias y una de tiempo. Para el caso del conjunto de datos de personas se identificaron 625.871 registros y una estructura fundamentada en 28 variables predictoras para clasificación (Tabla A2). De estas variables una es numérica, 25 son Enteras categóricas, una binarias y una de tiempo. Para el conjunto de datos de vivienda se cuenta con 197.490 registros, donde se definieron nueve variables predictoras para clasificación (Tabla A3). De las cuales, tres son Enteras categóricas, dos son binarias, una de tiempo, una de identificación y dos columnas para borrar.

## **Modelación de los datos**

A partir de estos datos, se utiliza un marco de construcción para sistemas de aprendizaje automático basado en la metodología (CRISP-DM). Este marco de construcción proviene de la revisión del ciclo de vida de un proyecto de minería de datos en el que se comprende el propósito del ejercicio, seguido de la comprensión y preparación de los datos. Esto se complementa con la modelación, la cual es evaluada antes de ser desplegada (Chapman y col., 2000). Con los conjuntos de datos se lleva a cabo su preparación con la integración de 19 archivos importados desde Microsoft SQL Server Management Studio. Con este procesamiento de datos se generaron archivos independientes para hogar, persona y vivienda, al momento de cargar la información se genero una marca de tiempo de acuerdo con cada conjunto de datos con el fin de identificar los años 2016, 2017, 2018 y 2019. Después se cargaron los conjuntos de datos en Google Drive con la finalidad de realizar el procesamiento desde Google Colab. En la exploración de los datos se realizaron análisis descriptivos y se determinaron variables objetivo para el inicio del modelado. Los datos en esta fase de preparación cubre todas las actividades para construir los conjuntos finales de datos que serán alimentados para el modelado, pues, antes de realizar algoritmos de clasificación se aplica el balanceo de datos, ya que el número de observaciones no es el mismo para la clase de la variable objetivo pobre en el conjunto de hogar, no se aplica para los otros conjuntos de datos este balanceo. La fase de modelamiento es explicada en la sección de métodos.

## **Análisis de los datos**

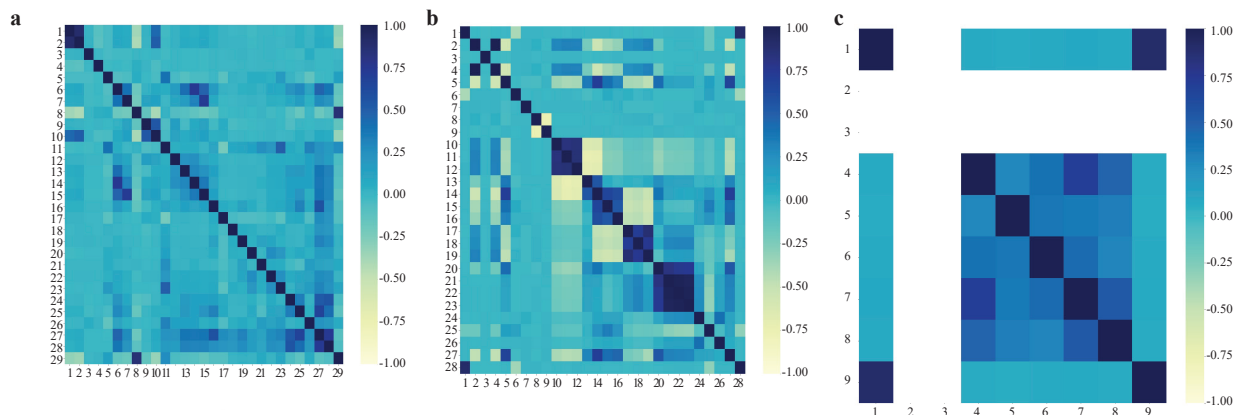
En esta fase se llevo a cabo la organización, la descripción y el análisis de los datos. Antes de analizar cada uno de los conjuntos de datos, se realizó una preparación que consistió en el mapeo y estandarización con el objetivo de asegurar su calidad y así determinar cualquier discrepancia. Posteriormente, se identificaron los datos que se relacionan entre si aplicando la correlación de Pearson para cada uno de los conjuntos de

datos, en los que se identificaron variables con correlación positiva y de esta forma se prevé la contribución a la variable objetivo de los conjuntos de datos utilizados (Figura 1). Sin embargo, una desventaja del coeficiente de correlación de Pearson es que solo es sensible a las relaciones lineales. Si la relación no es lineal, la correlación de Pearson puede ser cercana a cero incluso si hay una correspondencia de (Monroe, 1933) entre las dos variables. Por lo anterior para esta investigación no toda correlación implica causalidad en los conjuntos de datos.

### Figura 1

*Análisis de correlación de los conjuntos de datos. a. Hogar. b. Persona. c. Vivienda.*

*Fuente: elaboración propia.*



En la exploración de los conjuntos de datos de los hogares pobres, se observa una diferencia cuando el hogar es pobre y no pobre. Esta columna es clave para determinar la relación que existe con las demás variables. Al analizar las variables categóricas en este conjunto, se identifica una medición de como los hogares experimentan la pobreza en salud, educación y actividad laboral, pues al realizar la comparación durante el periodo 2016-2019 existe una similitud del *IPM*. Esto resulta en una mayor desigualdad respecto a la situación de pobreza.

Adicionalmente, para el análisis del conjunto de datos persona, la probabilidad de leer y escribir depende de solo una respuesta de Si o No, pero más allá de esta respuesta, el analfabetismo afecta varias facetas de la vida de una persona. Se evidencia que un 85 % de

la población catalogada como pobre respondió que no sabía leer y escribir siendo 1 el valor que corresponde a ningún nivel educativo alcanzado (Figura 2a). Por otro lado, al comparar la situación laboral que en mayor parte ocupan este grupo de población considerada analfabeta (Figura 2b), el 37% de la población para la categoría 5 corresponde a incapacitado permanentemente para trabajar. Entonces, al considerar este grupo de personas discapacitadas no se puede afirmar que reciban algún tipo de ayuda del gobierno y se conecta con otra variable que no solo es relevante para esta categoría sino para la población en situación de pobreza que experimentan problemas de acceso a la salud, por consiguiente, en la mayoría de las personas que se les pregunto que hizo para tratar los problemas de la salud vemos que para la categoría 4, las respuestas se concentran en: "consultó un tigua, empírico, curandero, yerbatero, comadrona", respondieron aproximadamente un 27% (Figura 2c). Con relación a las demás variables, se evalúa la mayor parte del tiempo que ocupó entre semana ya sea laboral y si ejerció otra actividad extra de una forma u otra remunerada, para estas variables se evidencia privación al acceso a un empleo formal en el conjunto de datos de persona.

Sumado a esto, la mayoría de los hogares está afectada fundamentalmente por dos aspectos que definen el avance del hogar como son el empleo formal con 82% y el logro educativo con 57%. Adicionalmente, los hogares se ven afectados por otras variables como son el nivel de alfabetismo, el desempleo de larga duración, las barreras al acceso a la salud, el aseguramiento de la salud, el trabajo infantil, la atención integral, la inasistencia escolar, el rezago escolar, así como variables que se integran al tipo de vivienda como paredes, pisos, alcantarillado, acueducto y hacinamiento. Esta última variable presenta una asociación negativa considerable entre la pobreza y el desarrollo humano de los hogares.

Otras de las variables asociadas al nivel de pobreza, es el tipo de servicio sanitario, el cual representa el acceso de servicios sanitarios que tienen los hogares. Esta es otra forma representativa de desigualdades entre los pobres y los no pobres como se observa en la Figura 3a. Conforme a lo anterior, el no acceso al servicio sanitario de un hogar afecta

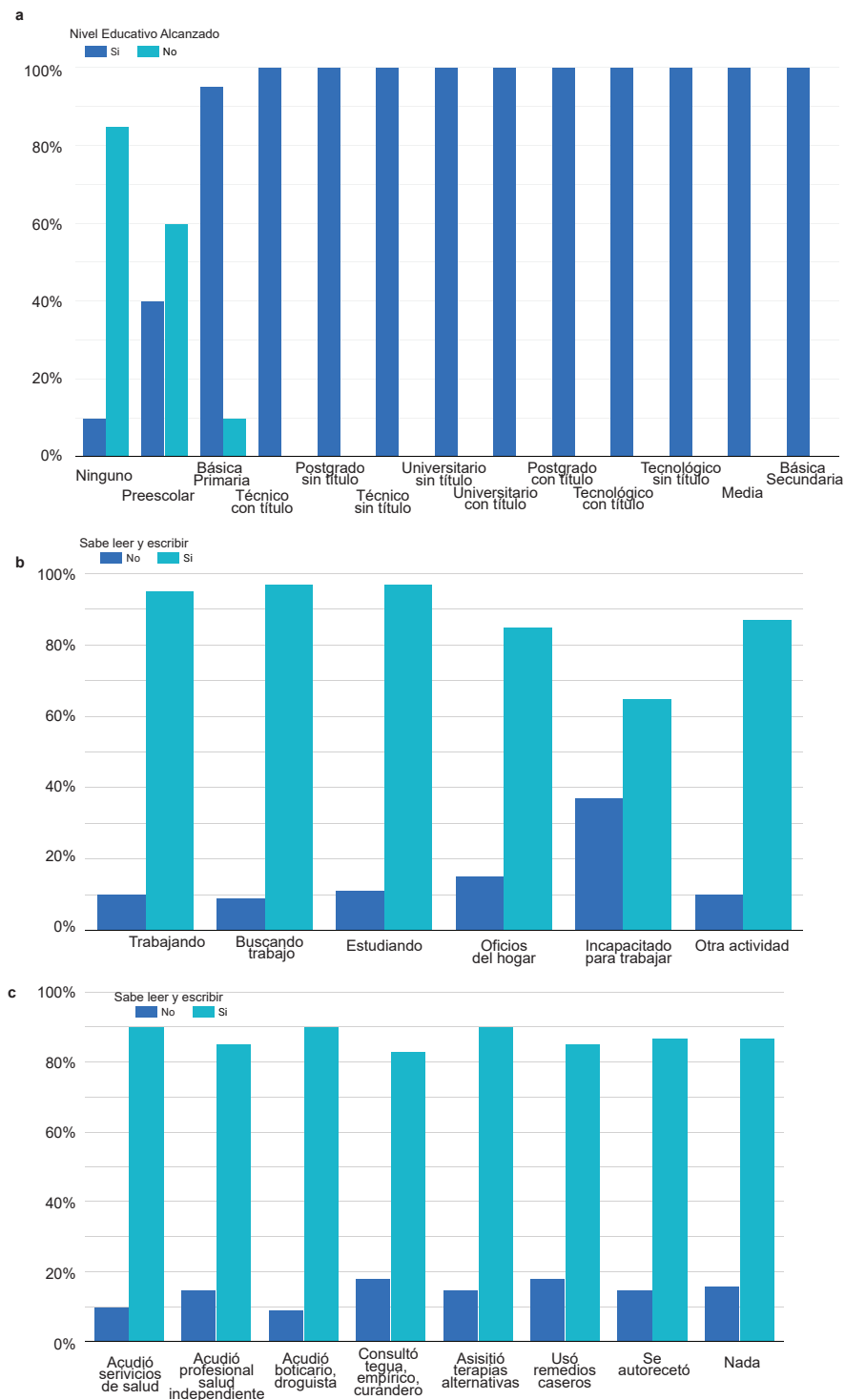


negativamente las condiciones de vida, pues afecta la salud de las familias. En el conjunto de información de datos de vivienda se analiza el acceso al alcantarillado que puede tener una vivienda. Este análisis se realiza con el fin de determinar el material predominante de los pisos y paredes, para identificar el nivel de pobreza que repercute en el déficit de vivienda. Como se observa en la Figura 3a, el derecho de tener una vivienda digna, donde 0 es si y 1 es no, se pueden determinar los hogares que residen en viviendas con deficiencias, principalmente, a partir del tipo de materiales usados en la construcción. En cuanto a los pisos existe similitud con la base de hogares con privación al pertenecer a un hogar que no cuenta con servicio público de alcantarillado. Por otro lado, se puede observar que a partir de la categoría 5 "Madera burda, tabla, tablón, otro vegetal", la categoría 6 Cemento, gravillaz la categoría 7 "Tierra, arena." existen viviendas que a pesar del material para los pisos no cuentan con servicio de alcantarillado respecto al material (Figura 3b). Finalmente, en cuanto a las paredes exteriores se resaltan la categoría 7 "Guadua, caña, esterilla, otro vegetal", la categoría 8 "Zinc, tela, carbón, latas, desechos, plásticoz la categoría 9 "Sin paredes"(Figura 3c).

**Figura 2**

*Análisis de Pobreza en Colombia. a. Alfabetismo versus nivel educativo. b. Alfabetismo versus actividad laboral. c. Alfabetismo versus problemas de salud.*

*Fuente: elaboración propia.*



**Figura 3**

*Análisis de Pobreza en Colombia. a. Servicio sanitario en porcentaje de distribución de pobreza. b. Alcantarillado versus material de pisos de la vivienda. c. Alcantarillado versus material de paredes de la vivienda.*

*Fuente: elaboración propia.*



Con estos análisis se determinan las variables objetivo para la construcción y ejecución del modelo.

### **Parámetros generales de los métodos**

Las técnicas de aprendizaje automático se han incorporado con mayor frecuencia en la comprensión del desarrollo de la sociedad, ayudando a identificar la complejidad de las relaciones entre los diferentes indicadores económicos y los datos heterogéneos. Ya se han implementado estos métodos para tomar decisiones en diferentes ámbitos de la sociedad donde se han recopilado datos que permitan describir la situación de acuerdo con el estudio que se realiza (Dutt & Tsetlin, 2020). Los algoritmos de aprendizaje supervisado buscan la dependencia entre el resultado final y el inicial de la descripción de la tarea. Se utilizan para tomar decisiones relativamente sencillas. Estos algoritmos están equipados de un conjunto de datos etiquetados de entrada lo suficientemente grande como ejemplo para que un experto o supervisor le muestre al aprendiz los datos de entrenamiento al mismo tiempo que le indica cuál es la respuesta correcta en cada caso (Kelleher y col., 2015). La finalidad de este algoritmo es captar patrones en los datos y montar un conjunto general de procedimientos para conceder la entrada a la clase o evento. Al incorporar la utilización de algoritmos de aprendizaje automático supervisado, se puede predecir la pobreza basado en el nivel de los hogares, como es el caso de la India en donde se utilizó XGBoost para predecir la pobreza (Sharma y col., 2019) y aplicando técnicas de aprendizaje automático en imágenes por satélite con los datos de luz nocturna, pronostico los índices económicos de pobreza en este país (S. P. Subash & Aditya, 2018). Además en Indonesia propusieron un marco para la estimación de la tasa de pobreza basadas en e-commerce de datos utilizando algoritmos de aprendizaje automático ("Wijaya, D.R., Paramita, N.L.P.S.P., Uluwiyah, 2020).

Los algoritmos de aprendizaje automático supervisado se pueden agrupar según su tipo de variable de salida en dos grupos:

- **Regresión:** La salida que se va a pronosticar es un número continuo en relevancia con un conjunto de datos de entrada dado. Este algoritmo que predice valores continuos, por ejemplo, tratar de predecir el valor de la casa o el tiempo en la calle. Este tipo de tareas no tiene un límite definido de valores, ya que el valor puede ser cualquier número sin restricciones. De esta forma, los algoritmos de regresión son útiles para predecir productos que son continuos, es decir, las respuestas a los interrogantes se representan mediante un valor que puede ser definitivo en función de las entradas del modelo, en lugar de restringirse a un conjunto de posibles etiquetas (Witten y col., 2017).
- **Clasificación:** Este algoritmo predice valores discretos, por ejemplo, clasificar algo por categorías. La clasificación es una subcategoría de aprendizaje supervisado, cuyo objetivo consiste en predecir las etiquetas de clase categóricas (ej.: discreta o valores no ordenadas) de las nuevas instancias, basándose en observaciones históricas (Kelleher y col., 2015; Witten y col., 2017). Para esto, la construcción de modelos de aprendizaje automático supervisado tiene tres etapas: entrenamiento, prueba o validación y predicción.

El presente proyecto tiene un enfoque de aprendizaje automático para evaluar y supervisar mediante algoritmos de clasificación, en este caso se clasifica para la variable pobreza teniendo en cuenta los datos descritos en las tablas A1, A2 y A3. Posteriormente, se entrenan los modelos Naive Bayes, Decision Tree, Random Forest y XGBClassifier con la métrica F1, seleccionada teniendo en cuenta el desbalanceo de los datos. Adicionalmente, se incluyó la validación cruzada para estimar el rendimiento del modelo en particiones de cinco. Sobre el modelo seleccionado, XGBClassifier, se efectúa la optimización de los hiper-parámetros mediante la Librería Optuna que se adapta con el modelo seleccionado, a partir de un marco de optimización de hiperparámetros de búsqueda (Akiba y col., 2019). Finalmente en la evaluación del modelo, se centra exclusivamente la precisión del modelo en el conjunto de datos, evaluando los resultados obtenidos desde el punto de vista técnico

concluyendo una buena predicción y una buena percepción con las variables categóricas en los conjuntos de datos utilizados. A continuación, se explican los dos métodos utilizados.

### **Arboles de decisión**

Es una herramienta de minería de datos y de analítica predictiva, que permite resolver problemas de clasificación y regresión. La construcción de un árbol de decisión sigue un enfoque de división binaria recursiva y analiza la mejor variable para ramificarse, es decir, los algoritmos de árbol de decisión desglosan el conjunto de datos mediante la formulación de preguntas que permitan el fragmento de los datos que facilite hacer la predicción (Navada y col., 2011). Si el objetivo es una variable discreta (marca de clase), el modelo se llama árbol de clasificación y si es continua es un árbol de regresión. Su estructura es un método de representación de las decisiones de las reglas de la estructura jerárquica, compuesta por elementos de dos tipos: nodos y hojas. Sin embargo, existen los nodos raíz, los nodos de decisión y los nodos terminal que finalmente son las hojas. En los nodos se encuentran elementos decisivos de la regla y se realiza la verificación de la conformidad de los ejemplos de las reglas por algún atributo de un conjunto de aprendizaje (Hssina y col., 2014).

### **XGBoost (Extreme Gradient Boosting)**

En este método basado en un algoritmo de impulso bajo un marco de potenciación de gradientes, los árboles de decisión se crean de manera secuencial y el peso juega un papel importante en su creación. Se asignan pesos a todas las variables explicativas, que luego se introducen en un árbol de decisiones que predice los resultados (Chen & Guestrin, 2016). El peso de las variables predichas por el árbol incorrectamente aumenta y estas variables se pasan al segundo árbol de decisión. Estos clasificadores/predictores individuales luego se combinan para dar un modelo más sólido y preciso. Además, pueden trabajar en problemas de regresión, clasificación y predicción personalizada (Chen & Guestrin, 2016). XGBoost ofrece varias ventajas en comparación con otros métodos, basados en los árboles, tales

como Random Forest, AdaBoost y Gradient Boosting, especialmente, en la velocidad y la precisión, pues logra optimizar los recursos computacionales (Chen & He, 2020).

Para el análisis de la pobreza, este modelo a partir del aumento del gradiente, logra construir los árboles y obtener de una manera más precisa las puntuaciones de las características de la pobreza, por lo tanto, XGBoost define la importancia de cada característica para el modelo de entrenamiento. En otras palabras, cuanto más se utilice una función para tomar decisiones clave con árboles potenciados, mayor será su puntuación (Zheng y col., 2017).

## Resultados y Discusión

### Aproximación a la pobreza con algoritmos supervisados

A partir de la ejecución de los modelos, se determinó el más indicado a partir de su exactitud. Los modelos utilizados fueron Naive Bayes (NB), Decision Tree Classifier (DT), Random Forest Classifier (RF) y XGB Classifier (XGB). De los cuatro modelos todos tienen una sensibilidad alta, lo que significa que se hizo una correcta clasificación de los datos. En esta fase se evaluaron múltiples clasificadores a la vez con el fin de elegir cual se adapta a los conjuntos de datos y posterior a ello comparar su puntaje. La variable objetivo en el conjunto de datos Hogar fue si es pobre y no pobre y dada la alta correlación con el índice de pobreza multidimensional fue excluida de es este conjunto de datos como entrada del modelo. Para el conjunto de datos de persona la variable objetivo se determino cuando una de las personas sabia leer y escribir y para el conjunto de datos Hogar consistió en el Acueducto.

En este estudio, los conjunto de datos se dividieron en datos de prueba y datos de entrenamiento. Estos últimos a su vez fueron divididos para la validación cruzada a través de K-fold (Fushiki, 2011), generando aleatoriamente en cinco submuestras ( $k = 5$ ) de igual tamaño. Cuatro -bloques del conjunto de entrenamiento son preclasificados y el subconjunto restante que no han sido utilizados para ese entrenamiento genera la métrica

de clasificación. Lo anterior genera cinco resultados que se promediaron para producir los resultados de la tabla 1. En todos los resultados se utilizó la métrica de F1 como parámetro de evaluación de desempeño de los modelos.

	Modelo			
	Naive Bayes	Decision Tree	Random Forest	XGBoost
Hogar	79 %	100 %	100 %	96 %
Persona	75 %	82 %	82 %	82 %
Vivienda	63 %	61 %	61 %	62 %

### Cuadro 1

*Puntaje (Score F1) de los modelos de clasificación*

*Fuente: elaboración propia.*

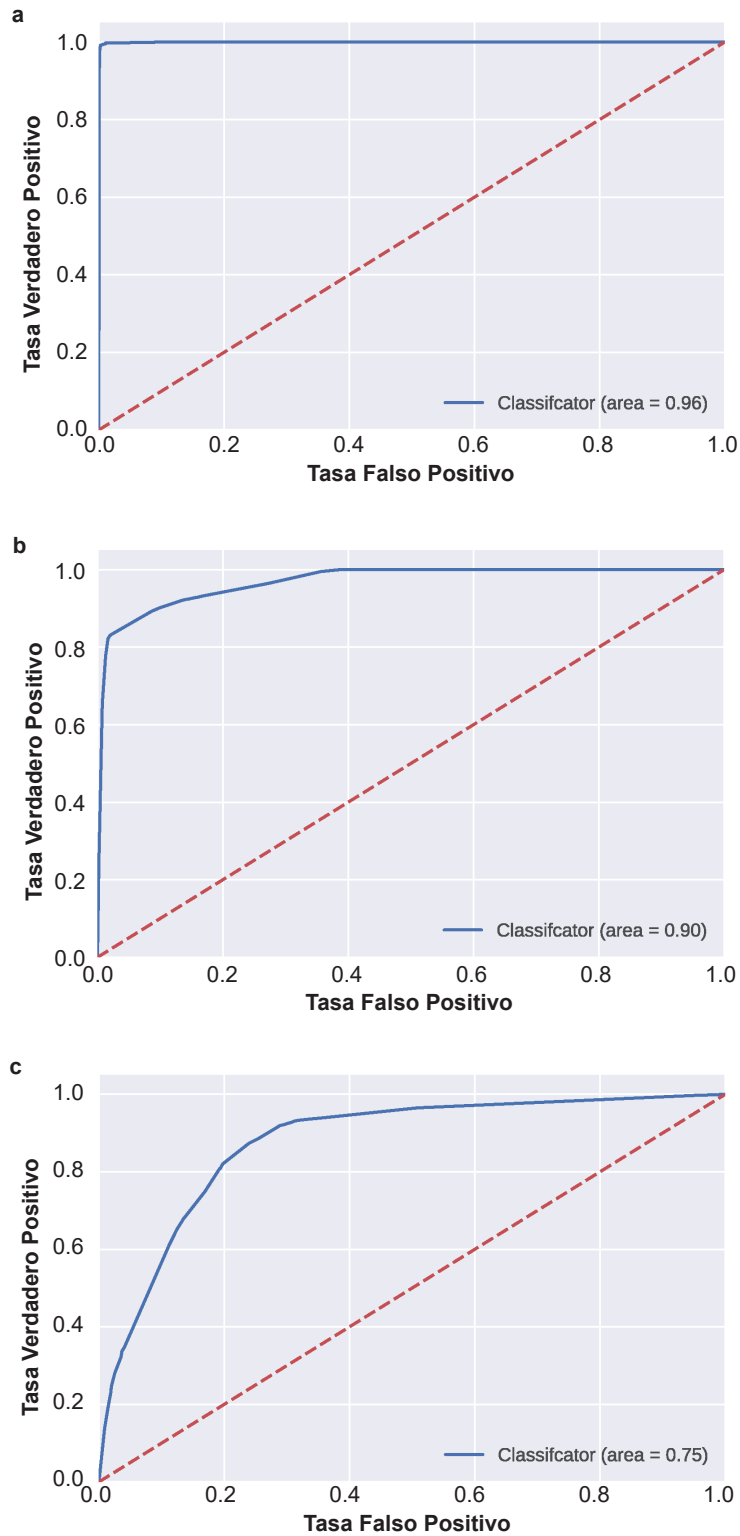
Teniendo en cuenta los resultados de cada uno de los modelos (F1-score), se seleccionó el modelo *XgBoostClassifier*. Posteriormente, se optimizaron los hiper-parámetros con el *FrameworkOptuna*, el cual realiza una optimización de hiper-parámetros usando validación. Finalmente, después de aplicar esta optimización, se logró una métrica (F1-Score) para el conjunto de datos de hogar de 99 %, de persona de 97 % y vivienda de 82 % sobre el conjunto de datos de pruebas que no ha sido utilizado en ninguno de los pasos anteriores. Adicionalmente, al realizar la validación de los modelos bajo la curva *Receiver Operator characteristic* (ROC) entre la tasa de verdaderos positivos y la tasa de falsos positivos en el modelo, se evidenció un óptimo rendimiento de modelo, pues los resultados tienen una sensibilidad alta lo cual resulta ser favorable (Figura 4). En este caso, la puntuación fue cerca a 1.0, lo que significa que tiene una forma de exactitud y precisión en el análisis en términos de sensibilidad.



**Figura 4**

*Rendimiento de modelos por conjunto de datos. a. hogar. b. persona. c. vivienda.*

*Fuente: elaboración propia.*

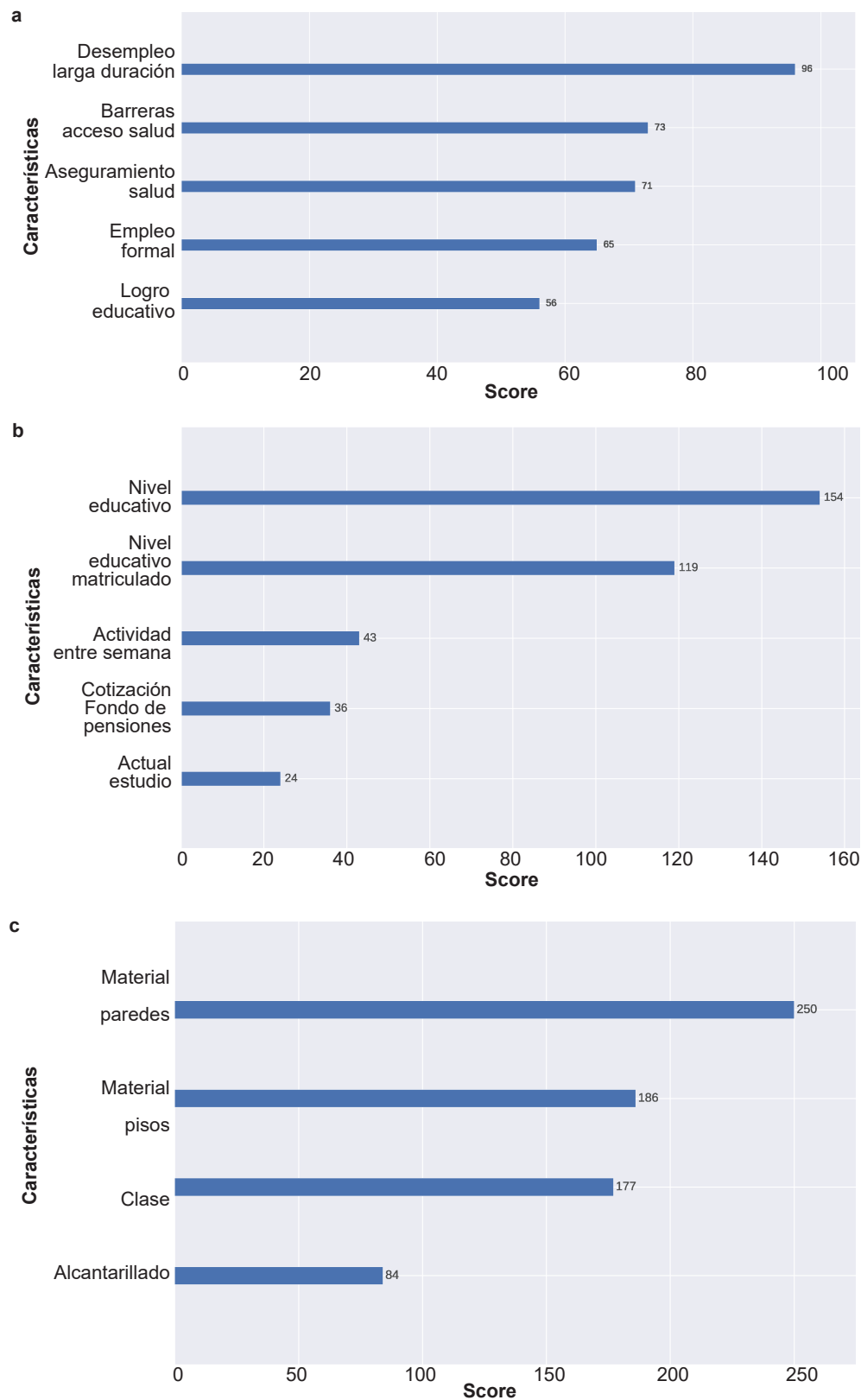


Luego de los resultados del modelo XGboostClassifier, se determinaron características relevantes que nos permitió entender las variables que influyen en la pobreza en Colombia. Con relación a las cinco dimensiones que miden el IPM podemos observar a nivel general que existen indicadores como se muestra en la figura 5 que pueden ser considerados factores consecuentes de situación de pobreza en el país. El modelo final logró identificar con una precisión del 82%, características relevantes como se visualiza en la (Figura 5) y con la métrica F1, medida que se utilizó para validar qué clasificador es realmente el mejor en términos de eficacia y calidad. El rendimiento general de la clasificación contribuyó a la evaluación de la pobreza donde cada conjunto de características se categorizaron para describir diferentes niveles de pobreza. En ellos se reflejó el déficit de necesidades básicas insatisfechas en el hogar, que en su defecto al considerar los resultados obtenidos como la privación de la actividad laboral, la educación y la salud, proporcionaron una visión significativamente de la importancia de una mayor demanda de acción e inversión para superar la pobreza.

**Figura 5**

*Características relevantes por conjunto de datos.*

*Fuente: elaboración propia.*



En este experimento se logró un mayor rendimiento en el modelo XGBoost Classifier. Para este caso, se obtuvo un predictor cualitativo de clasificación de características más relevante, donde la privación para desempleo de larga duración resultó ser de mayor importancia, incluyendo las otras variables que coincidieron con los hallazgos en el análisis del árbol. Aquí se ubicaron entre las trece variables más importantes. En el análisis del árbol, el subconjunto de variables permitió identificar que tan vulnerable esta la sociedad en el acceso a la educación y su dependencia económica.

Por otra parte, el modelo no permitió identificar una estimación futura de la pobreza, sino solamente logra mostrar las características más importantes que la causan. Esto significa que al contrastar con los estudios realizados por el DANE, se logran obtener resultados preliminares que describen los indicadores de pobreza. El uso de métodos basados en árboles de decisión permitió incluir variables predictoras sobre la pobreza. Así mismo, se encontraron variables predictoras de alta precisión que determinan que indicadores pueden ser objeto de estudio para la categorización de la pobreza.

### **Algoritmos supervisados e interpretación de la pobreza**

El experimento muestra que al aplicar herramientas de aprendizaje automático con algoritmos supervisados se pueden hallar factores que son claves para identificar la pobreza según las condiciones del hogar que se calculan con el índice de Pobreza Multidimensional. Esto se logró al trabajar un marco metodológico con base en los subconjuntos de datos, en los que se entreno el modelo XGBoost que permitió determinar algunas variables que ayudan a establecer los índices que causan la pobreza. Adicionalmente, la herramienta de clasificación pronosticó variables predictoras encontrando características relevantes en la educación, el trabajo y la salud.

De acuerdo con los resultados del modelo, en el conjunto de datos de hogar, el nivel más alto de importancia fue el desempleo de larga duración seguido de barreras en el acceso a la salud, aseguramiento salud, empleo formal y logro educativo. Estos elementos

fueron importantes para la variable objetivo cuando se clasifica si es pobre o no, evidenciando un diagnóstico del problema en la educación, la salud, la seguridad social y el trabajo que carece un hogar. Para el conjunto de persona, se puede establecer que la educación y la seguridad social influyen cuando en el hogar uno de los integrantes sabe leer y escribir. El tipo de lugar de residencia, categorizado como rural o urbano también influye en la clasificación en el acceso al alcantarillado. Se podría argumentar que esta se relaciona con la región de donde proviene. Sin embargo, no se sitúa qué características de la región influyeron en la clasificación anterior y esto va más allá del alcance de la presente investigación.

Es evidente que el modelo XGBoost se escala automáticamente, mientras que Random Forest necesita escalar la característica manualmente. Tanto Random Forest como XGBoost son robustos a valores atípicos en diferentes escenarios. Las estadísticas descriptivas del valor de pobreza pronosticado en las variables predictoras están basadas en la técnica de aprendizaje automático, el hallazgo encontrado respecto a la educación, el trabajo y la salud son importantes para el desarrollo del país y bienestar de los hogares, aplicando métodos para descubrir la desigualdad de ingresos.

En el caso de la evaluación general de la superación en la reducción de la pobreza el índice de la población pobre es un indicador elemental para el análisis. No obstante, en algunos casos, por ejemplo, para analizar el impacto en los pobres o condición socio económicos de las medidas, el uso de este indicador a menudo se utiliza para evaluar los efectos derivados que conllevan la pobreza en la población vulnerable. Esto se debe a que los beneficiarios de asistencia social pueden y no salir del grupo de los pobres, sino que es esencial mejorar su nivel de ingresos. En estos casos, para el análisis, la clasificación de precisión a menudo se presentó como una relación porcentual del número de predicciones correctas de todas las predicciones realizadas, las causas de la pobreza causadas no sólo por factores estructurales, sino también por las circunstancias familiares y personales de una persona.

Los indicadores utilizados en las estadísticas para evaluar la situación con el nivel de vida de la población en general y la pobreza en particular son muy poco informativos para mostrar diferencias cualitativas en la pobreza. En consecuencia, para distinguir a los pobres, generalmente se compara la privación de condiciones del hogar y aquellos cuyos ingresos están por debajo del mínimo de subsistencia ("línea de pobreza"), que debe garantizar la supervivencia humana básica. Entonces, el papel del Estado en la lucha contra la pobreza se limita, por lo general, a una asistencia mínima a los más pobres, la cual se concentra simplemente en la supervivencia física.

La investigación también determina que las variables importantes de predicción de la pobreza coinciden con la investigación de (Sharma y col., 2019), en la cual se determinó que los indicadores experimentan privaciones en la educación y el trabajo. Aunque, al aplicar cualquier política para contrarrestar la pobreza, se debe recordar que sus características económicas siempre se combinan con instituciones del contexto económico del país y, en este caso, se considera que el desempleo de larga duración es un indicador significativo para los hogares encuestados, dado que los antecedentes de este factor recaen considerablemente en el acceso y la continuidad de la educación.

A pesar de que el gobierno implementa ayudas económicas, aún existen hogares que presentan dificultades en la asistencia social. Al mismo tiempo, la cantidad de asistencia a los hogares pobres depende, en primer lugar, de las capacidades de la sociedad y del nivel de su desarrollo socio-económico. Se recalca el empleo formal en los diferentes escenarios al tener presente cuando un hogar es considerado pobre o no pobre, lo cual radica en la existencia de condiciones respecto al factor resultante, principalmente asociado con la escasez de empleos. Los empleos proporcionan estabilidad laboral, perspectivas de carrera y un salario aceptable que repercute al tener un empleo formal.

Como planteamiento para combatir la pobreza se podría estimular la creación de puestos de trabajo utilizando recursos de inversión social que promuevan actividades con valor agregado así como el desarrollo de sistemas educativos accesibles e integrados con las

necesidades de los procesos productivos maduros y no en etapas de implementación. También se observa en los resultados del modelo que se debe fortalecer en los integrantes del hogar, el nivel educativo y el logro educativo. Frente a la vivienda se puede evidenciar la probabilidad de que una persona se encuentre en nivel de pobreza, ya que esta privación influye considerablemente en la calidad de vida y prolonga la condición de pobreza.

### Conclusiones

De acuerdo con los resultados obtenidos y con las características identificadas en este estudio, se puede evidenciar varios elementos que contribuyen a la predicción de la pobreza, así como definir algunas herramientas que pueden sugerir algunos planes de acción para reducir la pobreza en Colombia. En la selección del mejor modelo predictivo, se evaluaron diferentes algoritmos de clasificación donde fueron evaluados con la métrica desempeño F1-score y ROC. El clasificador XGBClassifier resultó ser el de mayor eficiencia porque superó las métricas de rendimiento utilizadas y permitió predecir hogares, persona y vivienda que son pobres y no pobres respecto a su privación. Así mismo, su tasa de error fue significativamente menor o igual a un 0.03.

Las variables predictoras de pobreza fueron útiles para enfocar lo que parece más importante priorizar como es la educación y el trabajo, variables vitales para el desarrollo del país. El aumento del nivel educativo podría fortalecer la reducción de la pobreza y este estudio proporcionó algunos detalles para determinar la estructura causal de las dimensiones de medición de la pobreza por esta variable. Desde el punto de vista del nivel de vida de los hogares, este indicador parece ser subestimado, incluso cuando se utiliza la metodología para medir la pobreza desarrollada en la Universidad de Oxford.

Esta metodología está específicamente enfocada sobre el análisis de la pobreza en los países en desarrollo más atrasados y se basa en un enfoque de privación a la pobreza, según con la que se singulariza a los pobres sobre la base de las dificultades que experimentan. Según la metodología de Oxford, las personas que están al borde de la

pobreza. experimentan privaciones en un 33% sobre los indicadores evaluados. Se podría entender que el restante sería las personas que sufren "pobreza severa" de los demás indicadores. De esta forma y de acuerdo con los hallazgos de la investigación, se proporcionaron algunos elementos que deberían direccionar de manera más precisa los procesos de urbanización, especialmente en aspectos esenciales para mejorar la calidad de vida (agua potable, alcantarillado y calidad de los materiales de la vivienda), pues direccionar las mejoras exclusivamente al acceso de la tecnología y a procesos tecnológicos (en los que varias políticas públicas se ha basado en los últimos años) dejaría de lado aspectos que son esenciales para reducir la pobreza estructural. Adicionalmente, otra forma que podría ayudar a combatir la pobreza es fortalecer el acceso al sistema educativo a los jóvenes en las zonas rurales y urbanas de acuerdo con las especialidades requeridas, es decir, el fortalecimiento de las capacidades técnicas y oficios, pero donde el mercado laboral sea activo en la contratación por relevancia y no por obligatoriedad. Situación que sucede con los planes de contratación de aprendices del Sena.

Con base en los resultados, estos análisis utilizan una forma subjetiva de evaluar el bienestar personal para que de alguna manera estos factores relevantes sean un paso inicial de reducir la pobreza en Colombia, en las medidas de lucha contra la pobreza, será vital: seguir estudiando su estructura con mayor profundidad, ya que con el fin de reducir el nivel de pobreza, es necesario no solo proporcionar asistencia social a los que están por debajo de la línea de pobreza, sino también crear las condiciones para reducirla. Por ejemplo, esto puede incluir el aumento de la disponibilidad de servicios de salud, educación, proporcionar empleo formal en el mercado laboral y aplicar un plan estratégico que resulte de la priorización que se de con las herramientas de aprendizaje automático, lo cual resulta ser más práctico para implementar un plan que determine con mayor precisión si un individuo/hogar se clasifica como pobre o no pobre. Trabajos futuros plantean la oportunidad de utilizar los modelos para simular escenarios que permitan generar datos sobre las oportunidades de disminución de la pobreza ante diferentes condiciones de



políticas publicas.

Las personas pobres de múltiples maneras. Ahora mas que nunca sera necesario trabajar para combatir la pobreza y la vulnerabilidad ante esta en todas sus formas. Por eso es tan importante el índice de pobreza multidimensional que en este estudio fue de gran importancia. Si bien tenemos datos históricos sobre la pobreza, existe una gran brecha para las privaciones que aumentan la vulnerabilidad de los ciudadanos en Colombia. A futuro se espera que esta encuesta del índice de pobreza multidimensional (IPM) cierre la brecha; y que esta desviación significativa de la forma que se generaliza y mide la pobreza pueda aplicar técnicas de aprendizaje automático que conlleven a un cambio en el país a futuro.

### Referencias

- Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework.  
<http://arxiv.org/pdf/1907.10902v1:PDF>
- Alkire, S. & Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of Public Economics*, 95(7), 476-487.  
<https://doi.org/https://doi.org/10.1016/j.jpubeco.2010.11.006>
- Alkire, S. & Santos, M. E. (2013). A Multidimensional Approach: Poverty Measurement & Beyond. *Social Indicators Research*, 112(2), 239-257.  
<https://doi.org/10.1007/s11205-013-0257-3>
- Anand, S. & Sen, A. (1997). Concepts of human development and poverty! A multidimensional perspective. *United Nations Development Programme, Poverty and human development: Human development papers*, 1-20.
- Arestis, P. & Caner, A. (2010). Capital account liberalisation and poverty: how close is the link? *Cambridge Journal of Economics*, 34(2), 295-323.
- Atkinson, A. B. & Bourguignon, F. (1982). The Comparison of Multi-Dimensioned Distributions of Economic Status. *The Review of Economic Studies*, 49(2), 183-201.  
<https://doi.org/10.2307/2297269>
- Bahamón, M., Domínguez, J. & Núñez, J. (2013). La pobreza en Colombia, 2001-2005. Curvas globales, dominancia y aspectos inferenciales. *Revista de Economía Institucional*, 15(29).
- Bastiaansen, J., Herdt, T. D. & D'Exelle, B. (2005). Poverty reduction as a local institutional process. *World Development*, 33(6), 979-993.  
<https://doi.org/https://doi.org/10.1016/j.worlddev.2004.09.019>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

- Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>
- Chen, T. & He, T. (2020). xgboost: eXtreme Gradient Boosting. <https://mran.microsoft.com/snapshot/2020-07-15/web/packages/xgboost/vignettes/xgboost.pdf>
- Collins, P. D. (2012). Governance and the Eradication of Poverty: an Introduction to the Special Issue. *Public Administration and Development*, 32(4-5), 337-344. <https://doi.org/https://doi.org/10.1002/pad.1640>
- Dutt, P. & Tsetlin, I. (2020). Income distribution and economic development: Insights from machine learning. <https://doi.org/doi:10.1111/ecpo.12157>
- Epstein, G. S. & Gang, I. N. (2009). Poverty and Governance: The Contest for Aid. *Review of Development Economics*, 13(3), 382-392. <https://doi.org/https://doi.org/10.1111/j.1467-9361.2009.00496.x>
- Espinosa-Espinosa, A., Madero-Jirado, M., Rodríguez-Puello, G. & DíazCanedo, L. C. (2020). Etnicidad, espacio y desarrollo humano en comunidades pobres urbanas: la comuna 6 en Cartagena de Indias, Colombia. *Cuadernos de Economía*, 39(81), 635-665. <https://doi.org/10.15446/cuad.econ.v39n81.77333>
- Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, 21(2), 137-146.
- Gates, S., Hegre, H., Nygård, H. M. & Strand, H. (2012). Development Consequences of Armed Conflict. *World Development*, 40(9), 1713-1722. <https://doi.org/https://doi.org/10.1016/j.worlddev.2012.04.031>
- Grindle, M. S. (2004). Good Enough Governance: Poverty Reduction and Reform in Developing Countries. *Governance*, 17(4), 525-548. <https://doi.org/https://doi.org/10.1111/j.0952-1895.2004.00256.x>

- Hegre, H., Østby, G. & Raleigh, C. (2009). Poverty and Civil War Events: A Disaggregated Study of Liberia. *Journal of Conflict Resolution*, 53(4), 598-623.  
<https://doi.org/10.1177/0022002709336459>
- Hssina, B., Merbouha, A., Ezzikouri, H. & Erritali, M. (2014). A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications*, 4(2).
- Kelleher, J. D., Namee, B. M. & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. The MIT Press.
- Monroe, D. B., Walter S Stuit. (1933). The Interpretation of the Coefficient of Correlation. *The Journal of Experimental Education*, 1, 186-203.  
<http://www.jstor.org/stable/20150268>
- Navada, A., Ansari, A. N., Patil, S. & Sonkamble, B. A. (2011). Overview of Use of Decision Tree Algorithms in Machine Learning. *2011 IEEE Control and System Graduate Research Colloquium*, 37-42.  
<https://doi.org/10.1109/ICSGRC.2011.5991826>
- Ramírez, J. M., Díaz, Y. & Bedoya, J. G. (2017). Property Tax Revenues and Multidimensional Poverty Reduction in Colombia: A Spatial Approach. *World Development*, 94, 406-421.  
<https://doi.org/https://doi.org/10.1016/j.worlddev.2017.02.005>
- S. P. Subash, R. R. K. & Aditya, K. S. (2018). "Satellite data and machine learning tools for predicting poverty in rural India. *Agricultural Economics Research Association*, 31, 231-240. <http://10.0.23.70/0974-0279.2018.00040.X>.
- Salazar, R. C. A., Cuervo, Y. D. & Pardo, R. (2011). *Indice de Pobreza Multidimensional para Colombia* (Archivos de Economía N.º 009228). Departamento Nacional de Planeación.

- Sánchez Torres, R. (2015). Descomposiciones de los cambios en la pobreza en Colombia 2002-2012. *Revista Desarrollo y Sociedad*, (75), 349-398 los procesos cooperativos. <https://doi.org/10.13043/dys.75.9>
- Sánchez Torres, R., Maturana Cifuentes, L. & Manzano Murillo, L. (2020). Estimación alternativa de la pobreza multidimensional en Colombia. *Revista de Economía Institucional*, 22(43), 137-168. <https://doi.org/10.18601/01245996.v22n43.07>
- Sen, A. (1983). Development: Which Way Now? *The Economic Journal*, 93(372), 745-762. <https://doi.org/10.2307/2232744>
- Sen, A. (1999). *Development as freedom*. Anchor Books.
- Sen, A. (2008). Violence, Identity and Poverty. *Journal of Peace Research*, 45(1), 5-15. <https://doi.org/10.1177/0022343307084920>
- Sharma, A., Rathod, J., Pol, R. & Gajbhiye, S. (2019). Poverty Prediction Using Machine Learning. *Int. J. Comput. Sci. Eng*, 7(0), 946-949. <https://doi.org/10.26438/ijcse/v7i3.946949>
- "Wijaya, D.R., Paramita, N.L.P.S.P., Uluwiyah. (2020). "Estimating city level poverty rate based on e commerce data with machine learning. *Electronic Commerce Research*, 0, 27. <https://doi.org/10.1007/s10660-020-09424-1>
- Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. (2017). *Data Mining: Practical Machine Learning Tools and Techniques* (4.<sup>a</sup> ed.). Morgan Kaufmann.
- Zheng, H., Yuan, J. & Chen, L. (2017). Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation. *Energies*, 10(8). <https://doi.org/10.3390/en10081168>

## Apéndice

## Estructura de los conjuntos de datos

ID	Campo	Tipo	Descripción
1	periodo	discrete	Año
2	directorio	contin	Secuencia numérico directorio
3	secuencia encuesta	discrete	Secuencia encuesta
4	secuencia p	discrete	Secuencia hogar encuestada
5	p5010	discrete	No personas que duermen hogar
6	p8526	discrete	Tipo servicio sanitario
7	p8530	discrete	Preparación de agua
8	fex c	contin	Factor de expansión Hogar
9	region	discrete	Región
10	departamento	discrete	Departamento
11	personas	contin	No personas Hogar
12	paredes	discrete	Privación inadecuado material de paredes
13	pisos	discrete	Privación inadecuado material de pisos
14	alcantarillado	discrete	Privación inadecuado eliminación excretas
15	acueducto	discrete	Privación por no acceso fuente agua
16	empleo formal	discrete	Privación por tasa desempleo
17	desempleo larga duracion	discrete	Privación por desempleo
18	barreras acceso salud	discrete	Privación por barreras acceso a la salud
19	aseguramiento salud	discrete	Privación por no aseguramiento a la salud
20	trabajo infantil	discrete	Privación por trabajo infantil
21	atención integral	discrete	Privación por atención primera infancia
22	inasistencia escolar	discrete	Privación inasistencia escolar
23	rezago escolar	discrete	Privación rezago escolar
24	alfabetismo	discrete	Privación por Alfabetismo
25	logro educativo	discrete	Privación por bajo logro educativo
26	hacinamiento	discrete	Privación por hacinamiento crítico
27	ipm	contin	Índice de Pobreza Multidimensional
28	pobre	discrete	Pobre
29	fexp	contin	Factor de expansión Personas

Cuadro A1

*Descripción de las características del conjunto de datos Hogares*

ID	Campo	Tipo	Descripción
1	directorio	contin	Secuencia numérico directorio
2	secuencia encuesta	discrete	Secuencia encuesta
3	secuencia p	discrete	Consecutivo persona encuestada
4	orden	discrete	Consecutivo encuesta orden
5	p6040	contin	Años cumplidos
6	fex c	contin	Factor de expansión
7	p6090	discrete	Cotizante o Beneficiario
8	p5665	discrete	Algún problema en salud
9	p8563	discrete	Algún tratamiento en salud
10	p51	discrete	Donde o quien permanece mayor parte de la semana
11	p55	discrete	Recibe o toma desayuno o almuerzo en lugar donde permanece mayor parte de la semana
12	p774	discrete	Paga por esta alimentación
13	p6160	discrete	sabe leer y escribir
14	p8586	discrete	actualmente estudia
15	p8587	discrete	Cuál es el nivel educativo alcanzado
16	p8587s1	discrete	Grado o año aprobado
17	p1088	discrete	En qué nivel está matriculado y qué grado cursa
18	p1088s1	discrete	Grado o año que cursa
19	p6180	discrete	recibe en el plantel educativo alimentos
20	p6240	discrete	¿En que actividad ocupó la mayor parte del tiempo SP?
21	p6250	discrete	Además de lo anterior, ¿realizó SP alguna actividad paga por una hora o más?
22	p6260	discrete	¿tenía durante esa semana algún trabajo o negocio por el que recibe ingresos?
23	p6270	discrete	¿trabajó SP en un negocio por H más sin que le pagaran?
24	p6351	discrete	Si le hubiera resultado algún trabajo, ¿estaba disponible SP para empezar a trabajar?
25	p6390s1	contin	¿A qué actividad se dedica principalmente la empresa o negocio en la que realiza su trabajo?
26	p7250	contin	¿Durante cuántas semanas ha estado o estuvo buscando trabajo?
27	p6920	contin	¿Está cotizando actualmente a un fondo de pensiones?
28	periodo	discrete	Año

**Cuadro A2**

*Descripción detallada de las características del conjunto de datos Persona*

ID	Campo	Tipo	Descripción
1	directorio	contin	Secuencia numérico directorio
2	secuencia encuesta	discrete	Secuencia numérico encuesta
3	secuencia p	discrete	Secuencia numérico vivienda
4	P3	discrete	Clase
5	P4005	discrete	Material predominante de las paredes exteriores
6	P4015	discrete	Material predominante de los pisos
7	P8520S3	discrete	Alcantarillado
8	P8520S5	discrete	Acueducto
9	periodo	discrete	Año

**Cuadro A3**

*Descripción detallada de las características del conjunto de datos Vivienda*