

# Taller 2 Big Data & Machine learning

Hugo Sabogal, Gabriela Pérez, María Paula Osuna, Juan Andrés Silva

Universidad de los Andes  
Facultad de Economía

14 de abril de 2024

[Repositorio de GitHub](#)

## 1. Introducción

La pobreza es un problema multifacético que va más allá de la falta de ingresos monetarios, también abarca aspectos como acceso a servicios básicos, educación, salud, vivienda y empleo digno. Por lo tanto, la predicción de pobreza requiere un enfoque integral que considere variables socioeconómicas y demográficas. En este contexto, la aplicación de técnicas de aprendizaje automático juega un papel fundamental al permitir identificar patrones y relaciones complejas entre las variables que pueden influir en la incidencia de la pobreza. Así, contribuir al diseño de políticas públicas más efectivas y orientadas hacia la equidad.

Este documento aborda la predicción de la pobreza en Colombia utilizando datos de la Misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad (MESEP) del año 2018. Se realizó una cuidadosa selección de variables de interés, priorizando aquellas esenciales para la predicción. Se identificaron y manejaron adecuadamente valores faltantes, eliminando observaciones con información insuficiente y seleccionando variables relevantes tanto a nivel de hogar como de individuo. Asimismo, se desarrollaron nuevas variables indispensables para el análisis que complementaron la información disponible y enriquecieron el análisis. Se realizaron estadísticas descriptivas y se exploraron diversos modelos de clasificación y regresión, destacando el modelo Extreme Gradient Boosting para la clasificación de la pobreza y Random Forest para la regresión del ingreso. Finalmente, se desarrolló un modelo de ensamblaje con pesos que combina múltiples modelos para mejorar la precisión en la predicción de la pobreza. Todo esto con el objetivo de orientar de manera eficiente las políticas públicas relacionadas con este tema en Colombia.

## **2. Datos**

### **2.1. Descripción**

Con el objetivo de realizar modelos precisos en la predicción de la pobreza en Colombia y que a su vez ayuden a reorientar de manera precisa, rápida y eficiente las diversas políticas públicas sobre esta índole, se utilizaron datos provenientes de la Misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad (MESEP) del año 2018. Dicha misión fue creada en el año 2009 como un convenio del DANE y el DNP para diseñar una nueva metodología en la medición de pobreza monetaria<sup>1</sup>; metodología la cual es el fundamento de los datos que se utilizaron en la construcción de la base de datos para el presente trabajo.

### **2.2. Procesamiento y Limpieza de los Datos**

Para la construcción de la base de datos, inicialmente se importaron 4 segmentos de datos (los cuales se encontraban a nivel hogares y personas), 2 pertenecientes al conjunto de datos de entrenamiento y 2 al conjunto de datos de prueba<sup>2</sup>. Es importante recalcar que para obtener una buena base de datos es indispensable hacer un análisis de las variables que pertenecen a cada uno de los “sets” de datos: su definición, tipo y su contenido, pues de esta forma se escogerán las variables esenciales y pertinentes para el trabajo

Inicialmente, se seleccionaron las variables de interés a estudiar a lo largo del presente trabajo: la primera, Pobre, es una variable dummy que toma el valor de 1 si la persona (u hogar) se encuentra por debajo de la línea de pobreza y por tanto es denominada, tal como su nombre lo indica, pobre y 0 de lo contrario. La segunda, Ingpcug, corresponde a una variable continua que constituye los ingresos per cápita de la unidad de gasto con imputación de arriendo a propietarios y usufructuarios. Adicionalmente, para facilitar el análisis mencionado anteriormente, se identificaron las variables que coinciden tanto en la base de datos de hogares como en la de personas y se mantuvieron éstas<sup>3</sup>

Una vez realizado dicho análisis, se identificaron 20 variables categóricas que tomaban el valor de 9, la cual corresponde a que la persona (o el hogar) no sabe/no responde. Dado que esto es igual a tener un missing value, se decidió eliminar todas aquellas observaciones que tomaran dicho valor; al realizar este procedimiento se encontró que solo se perdieron el 0.043 % de las observaciones en cada una de las bases de datos. Adicionalmente, al realizar una tabla de estadísticas descriptivas a nivel personas que contuviese el % de missing values, la media y la desviación estándar de cada una de las variables de la base de datos de entrenamiento, se observó que había variables que serían esenciales en la predicción de la pobreza en Colombia pero que tenían un alto

---

<sup>1</sup> Esta información se recuperó de la documentación de la Medición de Pobreza Monetaria y Desigualdad del DANE (2019)

<sup>2</sup> En el Script se encuentran nombrados como: EP (personas) y EH (hogares) los datos de entrenamiento. Y TP (personas) y TH (hogares) los datos de prueba.

<sup>3</sup> Esto se realizó tanto para las bases de datos de entrenamiento como para las bases de datos de prueba.

porcentaje de missing values , como por ejemplo la variable Oficio la cual tenía más del % de observaciones como missing values<sup>4</sup> pero que es una variable determinística en el ingreso de una persona (u hogar). Por último, a partir de esta tabla se identificaron, además, variables demográficas que al ser incluidas en los modelos solo generarían ruido al no ser factores que determinan la pobreza en sí.

Con el objetivo de no perder variables esenciales en nuestra base de datos y tampoco generar problemas con la cantidad de missing values, se decidió realizar también una tabla de estadísticas descriptivas de las variables existentes únicamente para los jefes de hogar (esto implica que son a nivel hogar<sup>5</sup> ). A partir de ésta, y la anteriormente mencionada, se decidió dejar algunas variables a nivel personas y otras únicamente a nivel del jefe del hogar, dependiendo del % de missing values y su importancia en la predicción de la pobreza en Colombia. Es importante recalcar que se eliminaron las variables que tuviesen más del 30 % de sus observaciones como missing values a nivel de jefe del hogar. Una vez escogidas tanto las variables a nivel jefe del hogar y a nivel personas, se seleccionaron tanto en la base de entrenamiento como en la de pruebas.

### 2.2.1. Creación de nuevas variables

En este proceso se identificaron una serie de variables indispensables para el desarrollo del presente trabajo que no se encuentran en la base de datos de hogares<sup>6</sup> pero si en la base de datos a nivel personas. En primer lugar, se identificó la necesidad de una variable que contuviese la proporción de mujeres en un hogar, para esto se dividió el número total de mujeres sobre el total de personas para cada id. En segundo lugar, se creó una variable de edad, esta se divide en dos: 1. El número de niños en un hogar (entre 0 y 18 años) y 2. El número de personas de la tercera edad (de 70 años o más), indispensable para observar si acaso tener que suplir económicamente a una persona que no aporta en un hogar explica de alguna forma la presencia de pobreza en los hogares de Colombia.

En tercer lugar, es importante para la base de datos final tener una variable que describa los años de educación en un hogar, esta se construyó a partir del valor máximo que tomase la variable de “mayor nivel educativo obtenido” y se le asignó éste a la nueva variable. En cuarto lugar, se creó una variable que reflejase del número de personas en un hogar que reciben dineros de arriendos y/o pensiones, pues estos ingresos pueden contribuir a explicar la pobreza del hogar. En quinto lugar, para la construcción de la variable de ingresos no laborales (ingnolab) se asumió que, si al menos una de las personas pertenecientes a un hogar recibió dinero de otras personas, hogares, entidades no gubernamentales, dividendos, etc. Ésta tomaría el valor de 1 y 0 de lo contrario<sup>7</sup>.

Después de esto, se tomaron las bases de datos a nivel hogares y se realizó un “merge” de las nuevas variables con esta base y además, aquellas variables a nivel jefe

<sup>4</sup> Dado que ya se tienen las mismas variables tanto en las bases de datos de entrenamiento como en las de pruebas (a nivel persona) solo se realizó esto para la de entrenamiento.

<sup>5</sup> Dado que solo hay 1 jefe de hogar por hogar.

<sup>6</sup> Para el desarrollo de estas nuevas variables es importante tener en cuenta que la variable “id” corresponde a la llave del hogar al que pertenece cada persona, por lo que utilizando el comando group\_by se realiza todo este proceso para identificar las personas de CADA UNO de los hogares.

<sup>7</sup> Las mismas variables fueron creadas para la base tanto de entrenamiento como la de prueba.

del hogar. Para eso, tomamos cada una de las variables y, junto al comando left-join, las unimos a partir del id del hogar. Adicionalmente, se eliminaron 5 variables, que tal como se mencionó anteriormente generarían no más que ruido en la estimación de la pobreza en Colombia. Como resultado de esto, se obtuvo una base de datos con el nombre EH (de entrenamiento) con 28 variables y 164,960 observaciones y una base de datos con el nombre TH (de prueba) con 27 variables<sup>8</sup> y 66,168 observaciones.

### 2.2.2. Limpieza de la base de datos final

Con el objetivo de tener una base 100 % limpia e ideal para el desarrollo futuro de los modelos, se realizó de nuevo una limpieza de los datos. Para esta nueva limpieza se realiza, se realizó la misma tabla de estadísticas descriptivas mencionada anteriormente: que contuviese el % de missing values, su media y su desviación estándar, solo que en esta nueva tabla (dado que se encuentran las variables dependientes de interés) se agrega la media y desviación estándar cuando la variable Pobre toma el valor de 1 y cuando toma el valor de 0, de manera que se facilita ver si las variables se encuentran correctamente balanceadas.

A partir de esta, se decide arbitrariamente eliminar aquellas variables con más del 30 % de missing values en sus observaciones nuevamente, obteniendo una base con 21 variables para la de prueba y 23 para la de entrenamiento.

## 2.3. Estadísticas Descriptivas

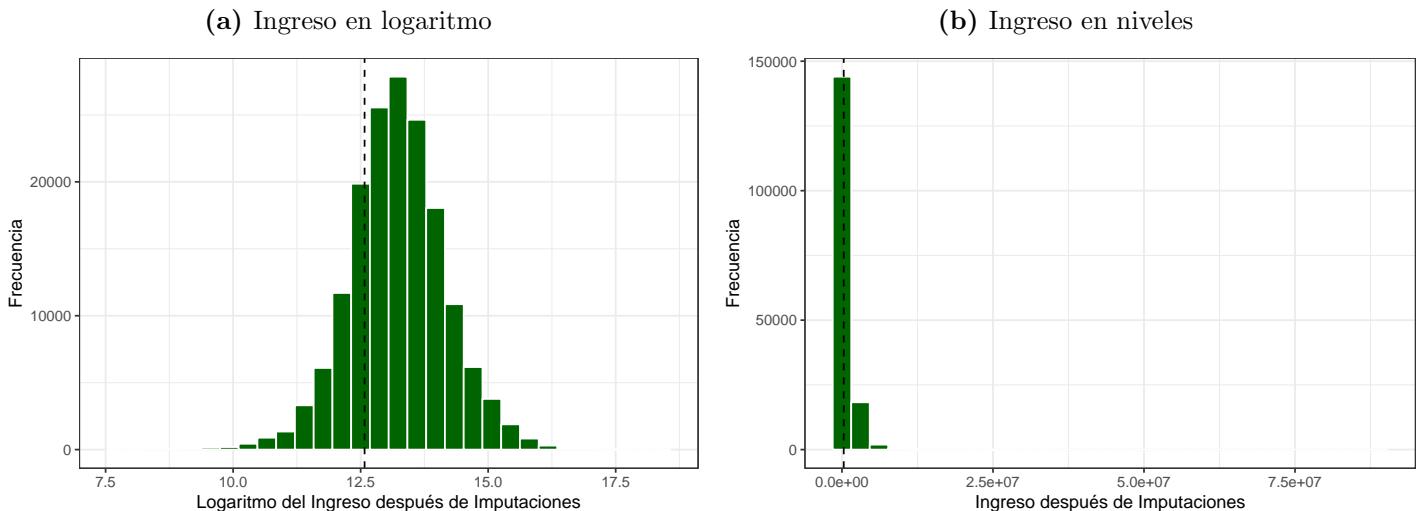
Una vez identificadas las variables relevantes para la predicción de la pobreza en Colombia, se realizaron una serie de gráficas y una tabla de estadísticas descriptivas con el fin de conocer el comportamiento y distribución de los datos de manera que tengamos un indicio de los resultados de las diferentes estimaciones<sup>9</sup>.

---

<sup>8</sup> Solo se dejó en la base de entrenamiento la variable de la línea de pobreza porque solo es necesaria ésta en los valores de entrenamiento para la predicción.

<sup>9</sup> Dado que el objetivo es realizar un análisis descriptivo del comportamiento de las variables, y teniendo en cuenta que éste es el mismo tanto en la base de entrenamiento como en la de prueba, solo se utilizaron los datos de la primera.

**Figura 1:** Distribución del ingreso



**Fuente:** Elaboración propia con GEIH2018, 2024.

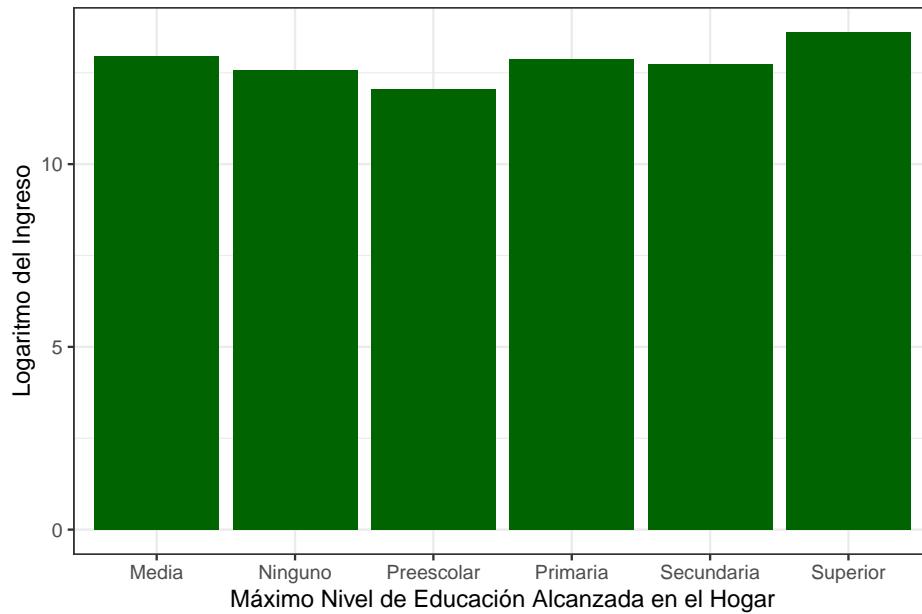
En primer lugar, observando la figura 1, se identificó que los ingresos después de imputaciones se encuentran distribuidos hacia la izquierda de la gráfica; es decir que la mayoría de los hogares cuentan con ingresos bajos. Esta asimetría dificulta la interpretabilidad de los resultados y refleja una alta varianza del error de esta variable. Por esto, es necesario realizar una estandarización de la misma, de manera que se transforme hacia una distribución normal estándar; para el análisis gráfico, se decidió transformar la variable a su forma logarítmica (véase la figura 1b), pero, para obtener una estandarización más precisa se determinó el uso del comando `scale(.)10` el cual resta la media y divide por su desviación estándar.

Adicionalmente, se decidió presentar la distribución del ingreso después de imputaciones en logaritmo con a finalidad de observar detalladamente la masa de probabilidad a la izquierda de la linea punteada de color negro la cual representa la linea pobreza. Como se observa en la figura 1a existe una masa de probabilidad a la izquierda de la linea de pobreza, sin embargo, no es lo suficientemente grande para considerar que la mayoría de los individuos del conjunto de datos son pobres, por el contrario, el la figura 1a indica que es mas probable no sufrir las condiciones de pobreza monetaria.

---

<sup>10</sup> Véase la línea 306 del Script de manipulación

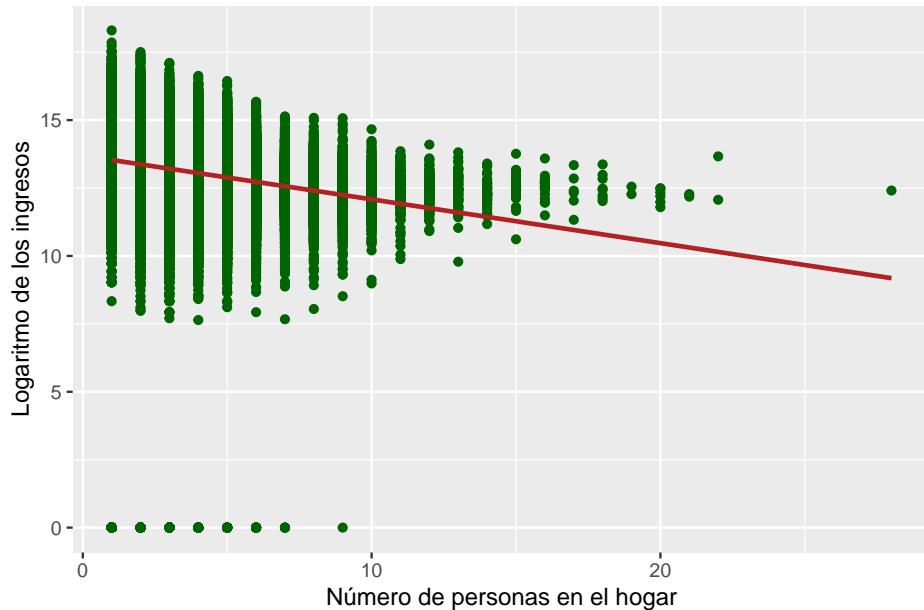
**Figura 2:** Log del ingreso promedio vs nivel educativo



Fuente: Elaboración propia con GEIH2018, 2024.

En segundo lugar, se realizó una gráfica de las medias del logaritmo del ingreso para cada categoría de la variable maxedu. Esto con el objetivo de observar el nivel de ingreso de hogares según su nivel máximo de educación alcanzado ; se identifica que a medida que un hogar tiene mayor educación alcanzada, mayor son los ingresos del mismo. Es importante recalcar que, dado que se imputó el valor de 1 a aquellos hogares que no sabían o no informaban la respuesta a esta pregunta, la media del ingreso para este grupo es relativamente alta con respecto a los hogares con un nivel de educación preescolar y esto puede explicarse porque hay un mayor número de observaciones (con ingresos bajos) en ésta.

**Figura 3:** Log del ingreso promedio vs numero de personas por hogar



Fuente: Elaboración propia con GEIH2018, 2024.

En tercer lugar, con el objetivo de observar la relación entre el número de personas que viven en un hogar y los ingresos de este se realizó la figura 3. De esta se puede concluir que a medida que un hogar cuenta con más personas en él, el logaritmo de los ingresos disminuye drásticamente; esto puede explicarse por varias razones, entre estas es que en un hogar con alto número de niños y/o viejos, no existe un aporte económico significativo por parte de estos dada su inactividad laboral y en realidad se consideren un gasto para éste.

**Tabla 1:** Estadísticas descriptivas de variables principales

Variable	Observaciones	Media	Desviacion Estandar	Min	Max
1 Numero de cuartos	164960	3.39		1.24	1.00 98.00
2 Numero de dormitorios	164960	1.99		0.90	1.00 15.00
3 Numero de niños	164960	0.98		1.16	0.00 15.00
4 Numero de tercera edad	164960	0.21		0.49	0.00 5.00
5 Jefe de hogar cotiza salud[1=sí]	164960	0.94		0.23	0.00 1.00
6 Horas trabajadas	164960	46.90		15.32	1.00 130.00
8 Tamaño empresa jefe de hogar	164960	3.81		3.31	1.00 9.00
9 % de mujeres en el hogar	164960	0.53		0.28	0.00 1.00
10 numero de personas en el hogar	164960	3.29		1.77	1.00 28.00

Fuente: Elaboración propia con GEIH2018, 2024.

Por último, de la tabla de estadísticas descriptivas, se pueden concluir varias cosas. Primero, se identifica que observaciones con un número de cuartos igual a 98 son datos atípicos que pueden afectar las estimaciones de los modelos y por tanto se sobreestime o subestime el efecto y por tanto la predicción de la pobreza. Segundo, en promedio, los hogares cuentan con 2 dormitorios y con 3 personas en él, lo que refleja que la mayoría

de los hogares colombianos cuentan con viviendas de tamaño proporcional al número de personas en él. En promedio, los hogares cuentan con 1 niño menor de 18 años y casi ningún adulto mayor a 70 años. Tercero, la mayoría de los hogares cuentan con un jefe de hogar cotizante a una seguridad de salud, por lo que se puede decir que en promedio los hogares de la base de datos hacen parte del sector formal. Por otro lado, en promedio el jefe del hogar trabaja 46 horas a la semana, lo que sigue dando evidencia de pertenecer al sector formal. Asimismo, el oficio que ocupa el jefe del hogar consta de 99 diferentes categorías. Además, la empresa a la que pertenece el jefe del hogar tiene en promedio entre 4 a 5 personas, lo que da indicios de que pertenece a pequeñas o medianas empresas. Por último, se observa que en promedio la proporción de mujeres se encuentra balanceada en los hogares, es decir, que existe una misma cantidad de mujeres que de hombres en el hogar.

### 3. Análisis predictivo

Para el análisis predictivo se encontró que la mejor base de datos para estimar los modelos era la base grande con mayor numero de observaciones, donde se le dio tratamiento a los valores faltantes. Como estrategia de submuestreo, utilizamos validación cruzada repetida de tamaño 5. Para los modelos de clasificación se escogió el modelo con mayor F1 fuera de muestra, mientras que para los modelos de regresión se escogió el modelo con menor error cuadrático medio fuera de muestra.

#### 3.1. Clasificación

Para abordar el problema de clasificación, se exploraron diferentes algoritmos de aprendizaje automático con el objetivo de predecir la probabilidad de pertenecer a una de las dos clases: cero (no pobre) y uno (pobre). En este análisis utilizamos Elastic Net, Árbol de Decisión, Random Forest y Extreme Gradient Boosting. El modelo más efectivo para la clasificación se seleccionó bajo la métrica F1. En este caso, el mejor modelo de predicción fue el Extreme Gradient Boosting. Este algoritmo funciona construyendo una serie de árboles de decisión de forma secuencial, donde cada árbol intenta corregir los errores del árbol anterior.

El modelo utiliza todas las variables de la base final de entrenamiento y los valores finales de los hiperparámetros se explican a continuación: Se realizaron en total 1000 iteraciones o rondas, donde cada una representa la construcción de un nuevo árbol. Asimismo, la profundidad máxima óptima para cada árbol es de 6 niveles. Un árbol más profundo puede capturar relaciones más complejas en los datos de entrenamiento, pero también puede conducir al sobreajuste. Por otro lado, la learning rate controla la contribución de cada árbol al modelo final. Un valor bajo de la tasa de aprendizaje generalmente requiere más árboles para lograr un rendimiento óptimo, pero puede mejorar la generalización del modelo. En este caso, la tasa de aprendizaje toma el valor de 0.01, un valor pequeño, pero que enfoca el ajuste del modelo hacia la generalización en lugar del sobreajuste. Gamma establece el umbral mínimo para dividir un nodo en el árbol. Esto ayuda a controlar la complejidad del modelo al prevenir divisiones que no proporcionan una mejora significativa en la función de pérdida. Su valor igual a

0.5 representa un equilibrio en términos de su influencia en el proceso de división de nodos en el árbol de decisión, no es ni muy “liberal” ni muy “conservador” a tomar esta decisión. Subsample y colsample\_bytree controlan el muestreo de filas y columnas durante la construcción de cada árbol. Subsample determina la proporción de muestras utilizadas para entrenar cada árbol, mientras que colsample\_bytree determina la proporción de características (columnas) utilizadas para entrenar cada árbol. Sus valores iguales a 0.5 y 0.6, respectivamente, demuestran, nuevamente, un equilibrio en la toma de estas decisiones. Por último, otro parámetro que controla la complejidad del árbol durante su proceso de construcción es min\_child\_weight. En este caso, 10 es la cantidad mínima de observaciones (instancias de datos) necesarias en un nodo hijo para considerar continuar dividiendo ese nodo.

Las características más importantes para la predicción de pobreza, según los resultados de XGBoost, son la cantidad de niños en el hogar, el estado efectivo de cotización de pensión, la educación máxima alcanzada por el jefe de hogar, la composición del hogar (personas, mujeres, personas de la tercera edad, ect), algunos dominios (Rural) y algunos oficios específicos.

### **3.2. Regresión**

En el caso de los modelos que buscaban predecir el ingreso directamente, se estimaron distintos tipos de especificaciones y algoritmos de estimación. Entre ellos se estimaron modelos lineales con boosting, Elastic Net, Random Forest y Extreme Gradient Boosting. El segundo mejor modelo, perteneció a este conjunto y se abordara a detalle en la siguiente a continuación.

El modelo usa todas las variables predictoras descritas en la sección de descripción de los datos. Los hiper-parametros de este modelo fueron seleccionados óptimamente a través de validación cruzada (5 fold cross-validation). En particular, se escogió un numero de 500 árboles para el algoritmo con una profundidad de árbol de 2. Respecto al numero de observaciones con nodo, el algoritmo de validación cruzada escogió óptimamente el numero de 20 observaciones por nodo. Por otro lado, se definió escogieron los valores de 0.1 y 0.01 para la tasa de aprendizaje y el algoritmo de validación cruzada escogió una tasa de aprendizaje de 0.1. Acompañado de la tasa de aprendizaje se definió el parámetro gamma para que variara entre los valores de 0 y 1. Por último, se fijó el parámetro de sub-muestra y se mantuvo constante durante el algoritmo en un valor de 70 % de la muestra.

### **3.3. Modelo final**

El modelo con mayor F1 de todos los estimados por el equipo fue un modelo de clasificación de ensamblaje con pesos. Este modelo se estimó utilizando las predicciones de los once mejores modelos previamente estimados como nuevos predictores. Se estimaron diferentes combinaciones de los valores predichos para cada observación y se calcularon nuevos valores. La forma funcional de este modelo comprende la combinación lineal ponderada de las predicciones de los modelos anteriores. Los modelos utilizados tienen un puntaje F1 que va entre 0.49 hasta 0.54, y dependiendo del F1 obtenido en Kaggle

se definieron los pesos de cada modelo en la predicción final. Donde el mejor modelo de 0.54 no se descontó (es decir se ponderó con 1), 0.53 con 4/5, 0.51 con 2/5, 0.5 con 1/5 y los modelos de 0.49 se redujeron con un factor de 1/10. Finalmente, se hizo un análisis de la distribución de esta combinación lineal y se definió un punto de corte de .5 para clasificar a las observaciones como pobre y no pobre, ya que representaba el valor del tercer cuartil (puesto del 75 % de la muestra).

El modelo de ensamble resultó superior a los otros dos expuestos, según la métrica escogida, porque se aprovechó la varianza en las predicciones de los modelos para incrementar la correcta clasificación de las observaciones, cuando este mejor modelo se equivocaba. Así, si un número suficiente de los demás modelos predice una ocurrencia de pobreza, aunque el mejor modelo no lo captura, la participación de los demás modelos puede corregir estos errores clasificación. Por la falta de una base de entrenamiento, no se pudieron optimizar los pesos de cada variable usando estimaciones fuera de muestra. En este sentido, una combinación lineal distinta pueda incrementar la capacidad predictiva del ensamble.