**Lúmina 25(1)**
**DOI: https://doi.org/10.30554/lumina.v25n1.4926.2024**

# Modelo explicativo para estimar el área mínima cultivada en monocultivo de café, que permita alcanzar el Living Income

*Explanatory model for estimating the minimal area of monoculture coffee plantations needed to generate a living Income*

***Jackeline Londoño Rendón*[1]**
***Diego Samir Melo Solarte*[2]**

## Resumen

El objetivo de esta investigación se basa en la construcción de un modelo explicativo para determinar el área mínima a cultivar en café, para alcanzar el Living Income en fincas cafeteras de tipo monocultivo.

En el escenario de la producción sostenible de café, se evalúa el ingreso neto anual de una familia en función del logro de un nivel de vida decente para todos sus miembros, la métrica que hace esto posible es el Living Income; desafortunadamente en el segmento de pequeños caficultores existe una amplia brecha entre el ingreso neto real y el Living Income, que es un fenómeno multidimensional afectado por factores exógenos o de difícil control como el área cultivada en café y su precio de venta y variables derivadas de

Recibido: 05 de septiembre de 2023. Aceptado: 16 de abril de 2024.

**Para citar este artículo:**

Londoño, J. & Melo-Solarte, D.S. (2024). Modelo explicativo para estimar el área mínima cultivada en monocultivo de café, que permita alcanzar el Living Income. Lúmina, 25(1), E0054.
DOI: https://doi.org/10.30554/lumina.v25n1.4926.2024

**Copyright:** © Esta revista provee acceso libre, gratuito e inmediato a su contenido bajo el principio de hacer disponible la investigación al público. Esta obra está bajo una licencia Creative Commons Reconocimiento - NoComercial - Compartir Igual 4.0 Internacional (CC BY-NC-SA 4.0)

1 Universidad de Manizales. Manizales, Colombia. Magister en Gestión Estrategica de la información. Correo electrónico: jackys1206@gmail.com. ORCID: https://orcid.org/0009-0001-7400-5408

2 Universidad de Manizales. Manizales, Colombia. PhD en Desarrollo Sotenible. Correo electrónico: mdiego@umanizales.edu.co. ORCID: https://orcid.org/0000-0003-0941-6697

su modelo productivo como la productividad, los costos, y la producción de alimentos en su finca.

Por tratarse de un fenómeno multifactorial, se aplicaron algoritmos de minería de datos (árbol de decisión y random forest) para clasificar las fincas que cerrarían la brecha entre el ingreso neto y el Living Income. Las iteraciones de los modelos se hicieron sobre un dataset que recopila información histórica de costos de producción, ingresos y áreas cultivadas, de fincas localizadas en cuatro departamentos de Colombia: Caldas, Cauca, Huila y Nariño.

Se concluyó que las variables que explican la brecha en un sistema de monocultivo de café son: área cultivada en café, productividad y precio de venta.

**Palabras clave:** *café, Living Income, costeo basado en actividades, ingreso neto, minería de datos*

## Abstract

The objective of this study was to construct an explanatory model for determining the minimum area required for coffee cultivation to provide a living income on monoculture coffee farms.

In the context of sustainable coffee production, the annual net income of families is evaluated based on achieving a decent standard of living for all members. This is measured using the standard of living as a metric. Unfortunately, small coffee farmers often face a significant gap between their real net income and the amount needed to achieve a living income. This gap is influenced by various exogenous factors that are difficult to control, such as the area cultivated in coffee and its sale price, as well as variables related to the productive models employed, including productivity, costs, and food production on the farm.

Data mining algorithms, specifically decision tree and random forest, were applied to classify farms that could close the gap between net and living income. The models were iterated using a dataset that collected historical production cost, income, and cultivated area information from farms in four Colombian departments: Caldas, Cauca, Huila, and Nariño.

The variables that explain the gap in a monoculture coffee system are the area planted in coffee, productivity, and sale price.

**Keywords:** *coffee, living income, activity-based costing, net income, data mining.*

## Resumo

Esta pesquisa apresenta a construção de um modelo explicativo para determinar a área mínima a ser cultivada com café, para alcançar uma renda digna em pequenas fazendas exclusivamente cafeeiras.

No cenário de produção sustentável de café, a renda líquida anual de uma família é avaliada com base na obtenção de um padrão de vida digno para todos os seus membros. A métrica que torna isso possível é a Renda de Vida; Infelizmente, no segmento dos pequenos cafeicultores existe grande disparidade entre o rendimento líquido real e o rendimento vital, que é um fenómeno multidimensional afectado por factores exógenos ou de difícil controle, como a área cultivada com café e o preço de venda, além de outras variaveis como o nivel de produtividade, os custos de produção e os custos dos alimentos na fazenda.

Por se tratar de um fenômeno multifatorial, foram aplicados algoritmos de mineração de dados (árvore de decisão e floresta aleatória) para classificar as fazendas que fechariam a lacuna entre a renda líquida e a renda vital. As iterações dos modelos foram feitas em um conjunto de dados que compila informações históricas sobre custos de produção, renda e áreas cultivadas, nas fazendas localizadas em quatro departamentos da Colômbia: Caldas, Cauca, Huila e Nariño.

Concluiu-se que as variáveis que explicam a lacuna num sistema de monocultura cafeeira são: área cultivada com café, produtividade e o preço de venda.

**Palavras chave:** *café, renda vitalícia, custeio baseado em atividades, lucro líquido, mineração de dados*

**JEL:**
**I31 General Welfare, Well-Being**
**C23 Panel Data Models • Spatio-temporal Models**
**C53 Forecasting and Prediction Methods**

## Introducción

Si bien la industria cafetera ha disminuido su aporte al producto interno bruto (PIB) colombiano, representando menos del 1% (Federación Nacional de Cafeteros, 2020a), sigue siendo un renglón importante en la economía del país por su rol protagónico en el sector rural, se debe resaltar que los ingresos por café le representaron al país, el 13% de los ingresos del sector primario (DANE, 2021) y genera 25% del empleo agrícola, ofreciendo 730 mil empleos directos y permitiendo que dos millones de personas vivan directamente de la producción de café (Federación Nacional de Cafeteros, 2017).

Sin embargo, estos indicadores macroeconómicos positivos no necesariamente reflejan la realidad de los productores, pues debido a diversos factores, entre ellos, la fluctuación de los precios internacio-

nales del grano; la utilidad por hectárea ha sido reducida, es así como en los últimos dos años (2018-2019) estuvo alrededor de 812 USD / Ha para los pequeños y medianos caficultores de la región cafetera central (Londoño, 2020). De los ingresos recibidos por los caficultores, el 35% se asigna a los costos de producción de café, otros cultivos e insumos pecuarios (Aristizábal y Duque, 2008) (Lykke et al., 2020), se asume que el ingreso restante debería ser suficiente para cubrir los gastos de una familia cafetera, alcanzando un nivel de vida decente, conocido como *Living Income,* el cual es entendido como: "el ingreso anual neto que requiere un hogar, en un lugar particular para permitir un nivel de vida decente para todos sus miembros, incluyendo elementos como: alimentos, agua, vivienda, educación, atención médica, transporte, ropa y otras necesidades esenciales, incluida la provisión para eventos inesperados" (The Living Income community of practices, 2020).

El valor promedio de Living Income para tres regiones: Caldas, Cauca y Nariño en 2018 fue de 15´169.000 COP (4464 USD) (Task force for coffee Living Income, 2020), mostrando que existe una brecha entre el ingreso neto y el Living Income de un pequeño caficultor con un modelo productivo de monocultivo. y a pesar de que el café es un producto agrícola que facilita la inserción al mercado aún con bajos volúmenes, como los ofrecidos por agricultores pequeños y medianos, la garantía de compra no es un factor que les garantice alcanzar el Living Income.

Ahora bien, el precio de venta ha sido objeto de discusión para orientar la política cafetera de Colombia en busca de mejorar el nivel del ingreso de los caficultores, aunque es evidente que se trata de una variable fuera de la esfera de control del agricultor y hasta ahora, del gremio o del estado, debido a que el precio del café colombiano se calcula a partir de tres variables del mercado: precio del café en el contrato (precio internacional del café arábico lavado), diferencial o prima de calidad y tasa representativa del mercado (Centro de comercio internacional, 2011); a su vez, estos factores tienen fluctuaciones, por la influencia de variables exógenas como la oferta y la demanda mundial de café, existencias de inventarios, variaciones de la tasa de cambio, entre otras.

Estas variables exógenas, conducen a que el precio de venta del café sea altamente volátil y no presente un comportamiento paralelo al índice inflacionario, factor que sí afecta de manera directa el costo de la canasta familiar y el costo de la mano de obra para la producción

del grano, que representa el 67% de los costos de producción de café (Londoño, 2019).

Por otra parte, un factor influyente y que puede ser restrictivo para alcanzar el Living Income es el área cultivada en café, de allí se deriva la importancia de determinar el área mínima a cultivar en café bajo la modalidad de monocultivo, para tener una base que permita interpretar la caficultura a partir de su segmentación por tamaño del predio y no de un modelo productivo único y uniforme; este punto de partida ayudaría a sentar las bases para construir medidas accionables y soluciones diferenciadas, que realmente atiendan a cada segmento de familias cafeteras desde su realidad. Las iniciativas de carácter público o privado dirigidas a mejorar las condiciones de vida de los caficultores deben tener en cuenta la gran heterogeneidad que existe entre las tipologías de los diferentes agricultores (García y Ramírez, 2002).

En Colombia, a diciembre 31 de 2020 se tenía registro de 540.227 cultivadores de café, distribuidos en 654.227 predios, con un promedio de 1,29 hectáreas en café por predio y 1,56 hectáreas en café por caficultor (Federación Nacional de Cafeteros, 2020a); lo cual evidencia que la industria del café está constituida en su gran mayoría, por agricultores pequeños. Los predios se clasifican en cuanto al tamaño del cafetal: áreas menores a 5 hectáreas en café, son denominados pequeños caficultores y representan el 96,6% de las fincas del país; entre 5,1 y 10 hectáreas medianos caficultores, éstos representan el 2,5%; y predios de más de 10 hectáreas son considerados grandes caficultores y representan el 0,9% del total (Federación Nacional de Cafeteros, 2020b).

Esta investigación se encargó de analizar variables económicas y técnicas que pudiesen tener efecto directo o indirecto en el ingreso neto de los pequeños caficultores y por lo tanto en la brecha con el Living Income; de tal manera que los resultados de este trabajo puedan convertirse en insumo para la construcción de una política económica cafetera.

## Referente teórico

Con el fin de caracterizar los pilares teóricos que dan forma a esta investigación, y con el propósito de unificar el marco conceptual con el que trabajó este proyecto, se presentan a continuación los constructos de Living Income y minería de datos:

## Living Income en familias cafeteras

El 75% del café producido en el mundo, se exporta y más del 90% es exportado como "café verde" es decir, sin procesar. Los ingresos generados al mercado minorista entre 2015 y 2020 fluctuaron entre 200 y 250 mil millones USD / año, mientras el valor promedio de exportación de café verde fue menor al 10% del ingreso percibido por el mercado minorista (Panhuysen y Pierrot, 2021), es así como los tostadores y los retailers de los países importadores capturan la mayor parte del valor agregado.

Aunque se genera la percepción de tener un margen de maniobra bastante amplio para mejorar el precio de la materia prima y, por ende, el ingreso de los agricultores; es un hecho que el precio del café verde es una variable que está por fuera del área de control de los países productores. Hay cuatro factores determinantes del precio internacional ellos son: Los indicadores, el mercado de futuros, los diferenciales y el producto físico (Centro de comercio internacional, 2011), a lo cual se adhiere la tasa representativa del mercado como otro factor que influye en la volatilidad del precio del café, en virtud de la volatilidad de la tasa de cambio (Steiner et al., 2015).

Dada la amplia fuente de variabilidad que influye el mecanismo de transmisión de los precios internacionales al precio interno, orientar el diálogo cafetero hacia el precio de venta es un camino incierto. Desde la visión global, la realidad es que los países productores tienen poco o ningún control sobre el precio internacional; a partir de 1988 se liberó el comercio internacional del café con el rompimiento del pacto de cuotas que regulaba los precios a través del establecimiento de volúmenes fijos de exportación (Cano et al., 2012) y desde el contexto país, se debe tomar en cuenta que aún los picos de precios, como los que se presentaron en 2020, pueden ser insuficientes para lograr el ingreso neto por año que requiere una familia para alcanzar un nivel de vida decente; cuando existe de por medio una restricción mayor, como lo es el área cultivada en café. Destacando que en los últimos años el sector cafetero en Colombia ha presentado una tendencia hacia la pequeña propiedad por la creciente subdivisión de la tierra ocasionando una gran limitación para que un número importante de productores alcance un nivel de vida adecuado, (Aristizábal y Duque, 2008; García y Ramírez, 2002)

Es importante tener en cuenta que los ingresos de un pequeño cafetero no solamente provienen de la venta del café, sino también de

algunos productos adicionales cultivados en sus parcelas además del ingreso colectivo de los integrantes representado en otras actividades por fuera de la finca, por otra parte, los miembros del hogar que trabajan en su finca aportan una mano de obra que si bien en algunos casos no es remunerada, hace parte de los costos de producción de café (The Living Income community of practices 2020).

Debido a que la medición del Living Income se aplica al núcleo familiar, se requiere determinar el tamaño de la familia y la ocupación de la mano de obra familiar en el trabajo de la finca, en Colombia se estima que hay entre uno y dos trabajadores de tiempo completo en la familia 1,5 FTE (del inglés full time equivalent) (Anker y Anker, 2017), sin embargo, otras fuentes han calculado una mano de obra familiar del 97% de la fuerza de trabajo requerida en las fincas, que es equivalente a 2,2 personas en las actividades productivas de la finca (Aristizábal y Duque, 2008).

Estudios en Colombia han arrojado resultados para Living Income en café indicando que un caficultor pequeño de café convencional en Colombia necesitaría cultivar 12,4 hectáreas de café para alcanzar el Living Income e Incluso, con un aumento simultáneo de la productividad de 910 a 1.183 kg café pergamino seco (cps) / ha (30%) y precios al productor de 1,01 a 1,32 USD / lb GBE (green bean equivalent) no ganaría por encima del nivel de la línea de pobreza (Task force for coffee Livng Income, 2020). En México los pequeños agricultores tampoco se ganan la vida solamente cultivando su café. Mientras ganaban 2.300 USD por año, el ingreso digno estaba alrededor de 5.400 USD / año, adicionalmente, debido al tamaño de sus parcelas (2,6 ha en promedio) no trabajan tiempo completo en sus granjas, especialmente fuera de la temporada de cosecha, por lo tanto, los ingresos complementarios se logran vendiendo su mano de obra en las granjas vecinas (Toorop et al, 2017).

True Price calculó un Living Income de 6.160 USD por hogar conformado por cuatro a cinco integrantes en el departamento del Cauca (Brounen et al., 2019); en otro estudio, Task force for coffee Livng Income, (2020) determinó que, para Caldas, Cauca, y Nariño el punto de referencia de Living Income para café en 2018 era 4.467 USD /año para un hogar de cuatro personas, cabe agregar que las personas que tienen una finca cafetera pequeña, logran ingresos complementarios vendiendo su mano de obra en las granjas vecinas (Toorop et al, 2017). En cultivos de vainilla en Madagascar y Uganda se calculó el costo de

un estándar de vida decente; descontando el alimento cultivado para consumo familiar; obteniéndose un valor de EUR 5.300 al año por familia en Madagascar y EUR 6.500 en Uganda (Veldhuyzen, 2020).

Más del 50% de los productores de cacao (Ghana y Côte d'Ivoire) y té (Kenya), tendrían que duplicar los ingresos de sus hogares para obtener el Living Income y un aumento del 50% en los precios de finca, para los productores de té en Kenia, solo llevaría al 6% de los agricultores a un nivel por encima de un ingreso vital (Waarts et al., 2019).

Es importante aclarar que, la base para estimar la brecha entre el ingreso neto y el Living Income es el cálculo de los costos de producción de café, dado que el ingreso neto es el remanente de dinero después de descontar los costos de producción de la actividad productiva. El método de costeo basado en actividades es aplicable a empresas agrícolas y pecuarias donde los costos de producción se calculan al final de los ciclos productivos; de esta manera, el modelo concibe que el costo de un producto o servicio debe abarcar todas las actividades necesarias para fabricarlo dentro de una cadena de valor; Los costos se asignan a las actividades y las actividades a los productos; los recursos son consumidos por las actividades y estas, por los productos (Heredia Gutiérrez, 2008). Una actividad se define como un evento o transacción que opera como inductor o impulsor de costo, es decir, actúa como factor causal en la incurrencia de costos en una empresa (García S., 2009).

Para la aplicación de este modelo de costeo en café en Colombia, se han segmentado los costos de producción en nueve actividades: Recolección, beneficio, fertilización, manejo de arvenses, control fitosanitario, otras labores, manejo de lotes en renovación, gastos administrativos y gastos financieros (FNC, citada por Araque, 2015).

## Minería de datos en la agricultura

Las técnicas de minería de datos se utilizan para predecir y extraer patrones de datos para comprender el comportamiento de las entidades y clasificar las que tienen atributos multivariados. Se utiliza para identificar patrones implícitos y explícitos en los datos. El patrón implícito define una relación preexistente en los datos que no está fácilmente disponible, mientras que los patrones explícitos son más evidentes en los datos (Hira y Deshpande, 2015).

En el campo agrícola, hay que tener claro que cada día la cantidad de datos aumenta significativamente y se hacen necesarias técnicas analíticas capaces de procesar y analizar grandes cantidades de datos para obtener información más confiable y predicciones mucho más precisas. Por lo tanto, la minería de datos debe jugar un papel importante en la agricultura inteligente para gestionar el análisis de datos masivos en tiempo real (Ait Issad et al.,2019).

En agricultura se han llevado a cabo estudios usando diferentes algoritmos de minería de datos: En el año 2003 se usó un algoritmo de árboles de decisión contrastado imágenes hiperespectrales de parcelas cultivadas con ensilaje o maíz en grano y cultivadas con labranza convencional, labranza reducida o sin labranza (Yang et al., 2003). Otro estudio aplicó la regresión logística para desarrollar un sistema avanzado de mapeo de rendimiento de cítricos mediante un sistema de visión artificial en una máquina limpiadora de residuos de cítricos (Shin et al., 2012). Con una técnica de aprendizaje no supervisado a partir de capturas de video de los brotes de la vid, se creó un sistema automático para la estimación del rendimiento de la vid (Liu et al, 2017).

Aunque en agricultura, la aplicación de la minería de datos ha sido predominantemente con fines clasificatorios aplicados a imágenes, también se han usado modelos estimadores como la regresión que es una técnica de modelado predictivo que investiga la asociación entre una variable dependiente (objetivo) y una variable independiente (predictor). En India obtuvieron un modelo predictivo del rendimiento de varios cultivos utilizando diversas técnicas de regresión (lineal, logística, polinomial). A partir de datos abiertos del gobierno se obtuvo una ecuación que predice la producción de un cultivo cuando se asigna al área y el cultivo (Surya y Aroquiaraj, 2018).

Si bien la aplicación de la minería de datos ha ido ganando terreno en diferentes sectores de la economía, por sus resultados en la reducción de costos en muchas aplicaciones industriales y comerciales, no se desconoce que la tasa de fallos llega al 60%, porque ni todos los resultados se aplican, ni todos los proyectos terminan con éxito (Nadali et al., 2011). En este sentido, es de gran utilidad la aplicación de una metodología que ordene las tareas y ayude a la optimización de recursos. El proceso estándar de Cross Industry Standard Process for Data Mining (CRISP-DM) define una secuencia no rígida de seis fases que permite la construcción e implementación de un modelo de

minería de datos para ser utilizado en un entorno real (Moro et al., 2011). las cuales según Nadali et al. (2011) describen el proceso de la siguiente manera:

- *Entendimiento del negocio*: se orienta a la comprensión de los objetivos y requisitos del proyecto desde una perspectiva comercial, lo que permite delimitar el problema en términos de minería de datos.
- *Entendimiento de los datos*: comprende microprocesos que ayuden a identificar problemas de calidad de los datos, allí se pueden descubrir las características más importantes de los datos como la identificación de columnas, nivel de agregación, naturaleza de las variables, entre otras
- *Preparación y preprocesamientos de datos*: incluye las actividades necesarias para construir el conjunto de datos que se llevará al modelado a partir de los datos sin procesar. Las tareas incluyen actividades como la detección y tratamiento de outliers (datos extremos), la identificación de valores nulos, la detección y tratamiento de valores erróneos.
- *Modelado de datos*: La fase de modelado construye el modelo que representa el conocimiento aprendido. Se pueden seleccionar y aplicar varias técnicas de modelado y sus parámetros se calibran a valores óptimos. Para el presente estudio se aplicarán dos algoritmos de DM: Random forest y Decision tree.
- *Evaluación*: antes de la implementación final del modelo, se debe evaluar más a fondo y revisar los pasos que se ejecutaron en su construcción para asegurarse de que cumpla adecuadamente con los objetivos. Un objetivo clave es determinar si hay alguna variable importante que no ha sido considerada.
- *Implementación del modelo*: la fase de implementación puede ser tan simple como generar un informe o tan compleja como implementar un proceso de minería de datos escalable a toda la compañía.

En el sentido de esta investigación, se utilizaron técnicas de minería de datos a través de regresiones no lineales para analizar los datos históricos de la producción de pequeños caficultores entre los años 2017 y 2019, con el fin de construir un modelo predictivo que permita identificar el área mínima a cultivar en café para alcanzar el Living Income en una finca cafetera pequeña con un modelo productivo de monocultivo.

## Metodología

El desarrollo de esta investigación corresponde a un enfoque cuantitativo, con un alcance explicativo, orientado a identificar las causas y relaciones de los eventos y/o fenómenos como variables que explican el cierre de la brecha entre el ingreso neto y el Living Income de un pequeño cafetero y su familia.

La información del estudio corresponde a costos de producción reales, tomados de los registros que manejan los agricultores de forma directa, recopilados por la organización Solidaridad Network a partir de históricos de los costos de producción de fincas cafeteras de cuatro departamentos (Caldas, Cauca, Huila, y Nariño), el dataset recopila información de producción, ingresos, área cultivada y costos de producción del café entre 2017 y 2019; en un número variable de fincas (año 2017: 253 fincas, año 2018: 969 fincas y año 2019: 1077 fincas). Las fincas del conjunto de datos pertenecen a los segmentos de fincas medianas y pequeñas, que producen café sostenible y están certificados con uno o varios estándares voluntarios de sostenibilidad (EVS).

El proceso investigativo se desarrolló aplicando las fases propuestas por la metodología CRISP-DM: Entendimiento del negocio, entendimiento de los datos, preparación de datos, modelamiento, evaluación y despliegue (Schröer et al., 2021), que se expresan a continuación.

## Construcción del modelo

La construcción del modelo se hizo con información de costos de producción reales, tomados de los registros que manejan los agricultores. A partir de ellos se obtuvieron las variables derivadas que se incorporaron en los algoritmos de Random Forest y Decision Tree (fase de modelado)

- **Entendimiento de los datos**

Está fase del proceso se orienta a la comprensión de la información determinando el conocimiento que aporta cada variable y su naturaleza; para este proyecto investigativo se contó con 15 variables continuas y una variable categórica (variables crudas), estas variables representan las nueve actividades de la estructura de costos basado en actividades, información de áreas cultivadas en café y variables relacionadas con la producción y los ingresos por concepto de las ventas de café; adicionalmente, se cuenta con variables que permiten caracterizar e

identificar cada finca: Id finca, departamento, municipio y organización que apoya al caficultor.

- **Preprocesamiento de los datos**

Corresponde a la preparación de datos de la metodología CRISP-DM, donde fue posible integrar tres datasets correspondientes al mismo número de años de recolección de información, a partir de este conjunto de datos unificado se hizo una primera segmentación para eliminar los registros correspondientes a fincas con cultivos de café mayores a 10 hectáreas, debido a que el modelo explicativo está focalizado en fincas medianas y pequeñas, es decir, menores o iguales a 10 hectáreas en café, logrando un dataset depurado con 1939 registros.

Como parte de este mismo proceso, se identificaron outliers (valores fuera de rango) y valores nulos, usando dos variables, productividad y costo por arroba, como variables de contraste para identificar outliers, debido a que un costo por arroba extremo puede originarse en datos erráticos del área en café o de los costos de producción. Finalmente, esta última variable no se incluyó en el modelo, por ser derivada de la relación entre la productividad y el costo total.

Adicionalmente, dentro del preprocesamiento de datos se aplicaron técnicas de *Feature Engineering*, la cual es una tarea central en la preparación de datos para el aprendizaje automático, que consiste en crear variables adecuadas a partir de unas características dadas, esto conduce a un mejor rendimiento predictivo del modelo. Implica la aplicación de funciones de transformación en determinadas variables para generar otras nuevas (Nargesian et al. 2017). A partir de esto, se identificaron tres variables explicativas (precio de venta por arroba, costos por hectárea y productividad) y la variable dependiente (brecha) que serían usadas en el proceso de análisis. Adicionalmente se identificaron 11 variables derivadas, como se describe en la Tabla 1 algunas de las cuales se usaron en os análisis de relaciones bidireccionales (estadística inferencial)

La brecha se calculó a partir del valor promedio de Living Income del año 2018 (5645 USD[3]) de los dos estudios más representativos para café en Colombia, el estudio de True Price (Brounen et al., 2019) y el de CIAT y Sustainable Food Lab (Task force for coffee Livng Income, 2020). Este valor se ajustó para cada año con el referente del índi-

3 Tasa representativa del mercado promedio para 2018 = 2956.55 COP / USD

ce de inflación, teniendo en cuenta que los factores que componen el Living Income son rubros afectados directamente por la inflación (alimentación, servicios públicos, vivienda, transporte, entre otros), es pertinente realizar estos ajustes. Para el año 2017 se aplicó una deflactación de 3,18% y para el año 2019, se hizo un ajuste por inflación del 3,8% (DANE, 2021).

**Tabla 1: Descripción de las variables derivadas**

| ID | Variable | Unidad de medida | Descripción / Cálculo |
|---|---|---|---|
| 16D | Porcentaje de área renovada | Porcentaje | Es la relación entre el área en café renovada (menor de 1 año) y el área total cultivada en café, multiplicada por 100. |
| 17D | Área en café | Hectárea | Variable explicativa; es la suma del área en producción y el área en levante |
| 18D | Área en café por rango | | Variable categorizada a partir de la variable área en café, segmentando la información en fincas menores o iguales a 5 hectáreas y entre 5 y 10 hectáreas |
| 19D | Costos totales | Pesos | Variable explicativa, Incluye los costos operacionales y los costos indirectos, que en nuestra estructura de costos se denominan gastos administrativos. Igualmente, involucra los gastos financieros relacionados directamente con el cultivo de café. |
| 20D | Costos por hectárea | Pesos / Hectárea | Variable explicativa, corresponde a los costos totales diferidos en las hectáreas cultivadas en café. |
| 21D | Precio de venta | Pesos / arroba | Variable explicativa, relación entre ingresos y arrobas de cps vendidas |
| 22D | Productividad | Arrobas / hectárea | Variable explicativa; para el presente estudio, la productividad se define como la cantidad de café pergamino seco producido por unidad de área sembrada en café (hectárea). |
| 23D | Ingreso neto | Pesos | Variable explicativa, es el ingreso final del agricultor después de descontar los costos de producción y los gastos administrativos y financieros, en este caso, del cultivo de café. |
| 24D | Brecha | Pesos | Variable dependiente del modelo, es la diferencia entre el Living Income y el ingreso neto anual. |
| 25D | Brecha_D | Valor dicotómico | Variable dicotómica transformada a partir de la variable cuantitativa Brecha. |
| 26D | Utilidad / hectárea | Pesos / hectárea | Variable usada en análisis de relaciones bidireccionales, es la diferencia entre los costos totales por hectárea y los ingresos por hectárea. |

**Nota:** esta tabla muestra una descripción detallada de todas las variables derivadas que fueron utilizadas en el diseño del modelo. Elaboración propia.

Para el cálculo de esta variable se incorporó un supuesto que afecta al ingreso neto: se asumió que toda la mano de obra utilizada en la producción del cultivo fue contratada, es decir, no se estimó la porción de la mano de obra familiar incorporada al proceso de la producción de café, que es uno de los drivers del ingreso neto, esta decisión fue basada en el hecho de que el dataset no contenía estos datos y no tenía las variables crudas suficientemente discriminadas para inferir el valor de la mano de obra de cada finca, así que incluir esta variable en el análisis hubiese implicado la incorporación de varios supuestos al modelo.

Cabe resaltar que dentro del modelo Crisp-dm, la fase de preprocesamiento es una etapa crucial que objetiva lograr un conocimiento preliminar de los datos y garantizar su calidad, a fin de tener un alistamiento de los mismos que facilitará su uso y comprensión, para lo cual se desarrolló las siguientes actividades.

- **Discretización de variables**

para la creación del modelo explicativo se discretizó la variable brecha, inicialmente era una variable continua y se transformó a binaria con el fin de aplicar los algoritmos de minería de datos.

Análisis exploratorio de datos (EDA): se realizó con el software SPSS, en esta fase se aplicó la prueba kolmogorov-smirnov, para identificar si los datos presentaban una distribución normal y a partir de allí definir la aplicación de las técnicas de estadística inferencial y las pruebas no paramétricas pertinentes (Cazacu y Titan, 2021).

Análisis de estadística inferencial: se realizaron análisis de correlaciones entre todas las variables derivadas y regresiones entre las variables explicativas y la variable respuesta, una vez identificado que los datos no seguían una distribución normal se usó una correlación no paramétrica a través de la prueba de Spearman. La regresión lineal es una técnica usada para pronosticar y modelar series de tiempo y descubrir la relación del efecto causal entre las variables, indica las relaciones significativas entre la variable dependiente y la variable independiente e indica la fuerza del impacto de múltiples variables independientes sobre una variable dependiente (Surya y Aroquiaraj, 2018).

- **Modelado**

Los datos fueron modelados usando algoritmos de Random Forest (RF) y Árbol de Decisión utilizando la herramienta Jupyter notebook (lenguaje Phyton), los algoritmos de minería de datos permitieron identificar las variables determinantes de la brecha. RF es un método de clasificación y regresión no lineal que se basa en la construcción de un ensamble de árboles de decisión; Los árboles de decisión son modelos en forma jerárquica, donde los datos se dividen en cada nodo de decisión en subconjuntos utilizando alguna regla; en el modelo, los nodos representan el resultado de la observación (Saarela y Jauhiainen, 2021).

Para el entrenamiento de los datos, inicialmente se hizo un Split de datos asignando 70% para entrenamiento y 30% para testeo, sin embargo, este fue uno de los parámetros que se modificó en las iteraciones posteriores buscando el mejor desempeño del modelo, para lo cual se ejecutó el modelo en RF en varias iteraciones, cambiando los siguientes parámetros:

- Porcentaje de datos de entrenamiento y datos de testeo
- Profundidad
- Número de estimadores

El desempeño del modelo se evaluó a través de la matriz de confusión (González et al, 2011) y el F1 Score (combinación entre precisión y sensibilidad) que permiten hacer un mejor análisis de desempeño de un modelo en bases de datos desbalanceadas (DeVries et al. 2021). A partir de la matriz de confusión se calculan los siguientes parámetros: la sensibilidad, especificidad, precisión y exactitud del modelo.

Para determinar el correcto funcionamiento del modelo *se* probaron dos dataset del año 2020, uno de ello con 933 registros de costos de producción e ingresos de fincas menores a cinco hectáreas cultivadas en café, ubicadas en los departamentos de Caldas, Cauca, Huila y Nariño (dataset 1); y otro conjunto de datos con 669 registros de fincas cafeteras (menores a cinco hectáreas cultivadas en café) de otros departamentos: Antioquia, Caquetá, Cundinamarca, Meta, Risaralda, Santander y Tolima (dataset 2).

Ahora bien, para comprobar el poder clasificatorio del modelo se asignaron valores 1 ó 0 a cada registro (concordante con los

valores de la variable binaria), de acuerdo con los rangos de referencia de los regresores, arrojados por el modelo, como se muestra a continuación:

> Valor 1= área café > 3,56 **Y** productividad > 141,5 **Y** precio venta > 23,07 USD[4]
>
> Valor 0 = área café <= 3,56 **O** productividad <= 141,5 **O** precio venta <= 23,07 USD

Se creó una nueva variable: brecha real, calculada a partir del ingreso neto, el precio de venta real de 2020 y el Living Income ajustado a 2020. La validación del modelo se ejecutó en la herramienta Power Bi.

El contraste de los valores binarios con los valores de la variable derivada (brecha real), se determinó a través de un DAX (Data Analysis Expressions) que para el valor 1 implicase cumplir las tres condiciones y para el valor 0, el no cumplimiento de al menos una de las tres condiciones.

## Resultados del modelo con mejor desempeño

Los parámetros y variables que arrojaron el modelo con mejor desempeño fueron los siguientes:

- Split de los datos para entrenamiento y testeo: 80/ 20 (80% para entrenamiento y 20% para testeo)
- Máxima profundidad = 4
- Número de estimadores =140
- Variables explicativas: Productividad, precio de venta y área en café.

## Árbol de decisión

Este algoritmo arrojó un F1 score de 73% y la matriz de confusión mostró la precisión del modelo clasificando las fincas como positivos si alcanzaron el Living Income (brecha <=0) y negativos en el caso contrario, como se puede identificar en la Tabla 2

4 Tasa representativa del mercado promedio para 2019 = 3289.32 COP / USD

**Tabla 2:Matriz de confusión árbol de decisión**

| Observado | Pronosticado | |
|---|---|---|
| | Brecha_dic | |
| | 0 | 1 |
| 0 | 327 | 9 |
| 1 | 17 | 35 |

Nota: esta tabla muestra la matriz de confusión reveló que hubo 327 verdaderos negativos, y 9 falsos positivos, con estos referentes se calculó la especificidad obteniendo un valor de 97%.
Construcción propia.

Dado el propósito de la investigación, esta métrica es la que mejor se ajusta a las características de los datos y los objetivos del modelo explicativo debido a que demuestra el desempeño del modelo para clasificar correctamente las fincas que no alcanzarían el Living Income.

## Random Forest

Para una mejor comprensión de los resultados encontrados a partir de RF es necesario tener presente que el resultado previsto de un modelo de RF es la moda o la media de las predicciones de los árboles individuales o árbol de decisión (Saarela y Jauhiainen, 2021). Por esta razón la matriz de confusión del árbol de decisión es válida para este algoritmo y en la Tabla 3 se puede visualizar las métricas obtenidas.

**Tabla 3: Métricas de performance del modelo Random Forest**

| Valores binarios | Precisión | Recall | F1-Score |
|---|---|---|---|
| 0 | 0,97 | 0,99 | 0,97 |
| 1 | 0,89 | 0,77 | 0,82 |
| Accuracy | | | 0,96 |
| Promedio | 0,93 | 0,88 | 0,9 |

**Nota:** esta tabla muestra el ensamble de árboles de RF, indicando que se mejoraron todas las métricas de medición de performance del modelo, notándose una mejora sustancial en el F1 Score, con un 90%.
Construcción propia

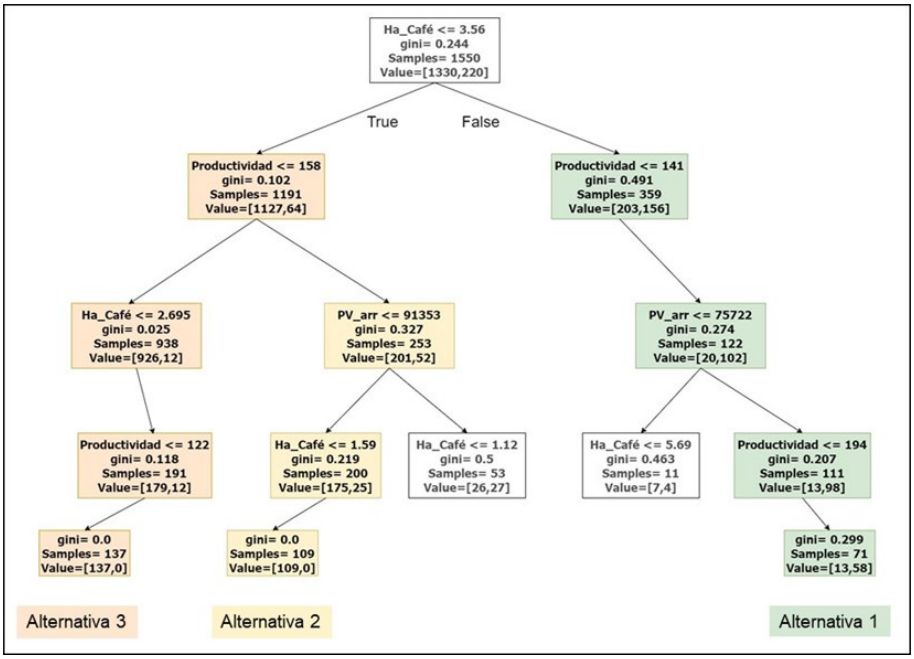


Figura 1: Ramas de las clasificaciones finales del árbol de decisión.

Para identificar las mejores alternativas del árbol de decisión se tuvieron en cuenta dos criterios; el primero, el índice de Gini, que mide el grado de impureza de los nodos; se buscaron las alternativas con los valores Gini más cercanos a 0. El segundo criterio fue llegar al nodo terminal con un tamaño de muestra significativo.

Como se observa en la Figura 1, la raíz del árbol fue el área en café, siendo 3,56 hectáreas el primer punto de división de la población, (fincas menores o iguales a 3,56 ha, ramificación izquierda = “True” y mayores a 3,56 ha ramificación derecha = “False”).

De acuerdo con el análisis de los nodos de la alternativa 1, hay un 81% de probabilidad de alcanzar el Living Income (LI), por parte de las fincas que reúnan las siguientes condiciones: área cultivada en café superior a 3,56 hectáreas, productividad superior a 141 arrobas de café pergamino seco (cps) por hectárea y precio de venta por encima de 23,07 USD por arroba de cps, siguiendo la ramificación de la derecha.

**Tabla 4: Resultados finales árbol de decisión**

| Alternativas (ramas del árbol) | Resultado | Predictores | | | GINI | Probabilidad |
|---|---|---|---|---|---|---|
| | | Área café | Precio Venta | Productividad | | |
| 1 | Lograr LI | > 3,56 Ha | > 23,07 USD / @ | > 141,5 @ / ha | 0,299 | 81% |
| 2 | No Lograr LI | < 3,56 Ha | | < 122 @ / ha | 0,000 | 100% |
| 3 | No Lograr LI | < 1,6 Ha | < 27,77 USD / @ | > 158 @ / ha | 0,000 | 100% |

Nota: esta tabla muestra los resultados finales obtenidos con la construcción del árbol de decisión. Construcción propia

Los valores de referencia de los regresores difieren del modelo propuesto por el Índice de Desarrollo Humano (IDH), el cual determinar que el área cultivada en café para alcanzar el Living Income debería ser de 4,7 hectáreas, en las siguientes condiciones: finca con café certificado (similar a las fincas del dataset que se analizó en este estudio), precio de venta 1,14 US/ lb GBE (equivalente a 77,600 COP /@) y una productividad de 146 @ / ha (Task force for coffee Living Income, 2020), similar a la del modelo RF. Aunque hay una diferencia amplia en cuanto al área en café, un punto de convergencia importante es que la propuesta de IDH calcula la brecha a partir de las mismas variables del RF: área en café, precio de venta y productividad.

Los resultados del algoritmo de Feature Importances (Saarela y Jauhiainen, 2021) demostraron que la variable área en café tiene una influencia de 41.5% sobe la variable respuesta; como segunda variable en importancia se encuentra la productividad con un valor de 33,8% y finalmente el precio de venta con 24.6%.

- **Validación del modelo**

La validación del modelo se ejecutó con dos conjuntos de datos, uno correspondiente a los departamentos que hicieron parte del estudio (dataset 1) y el segundo con siete departamentos nuevos (dataset 2).

De esta manera, Para el primer dataset con 933 registros, se presentó un alto porcentaje de falsos negativos (19%), es decir, fincas que el modelo pronosticó que no alcanzaría el LI pero si lo hicieron; mientras que para el segundo dataset con 669 registros se observó una tendencia más baja de falsos negativos (13%). Aunque el porcentaje de falsos negativos en ambos datasets fue alto, la especificidad fue del 99 % y 100% respectivamente, es decir el modelo logró clasificar correctamente las fincas que no alcanzaron el LI.

## Análisis de resultados

Los resultados más relevantes de este proceso investigativo se abordan en tres segmentos: estadística inferencial, modelado de minería de datos y validación del modelo.

### Estadística inferencial

Un resultado destacable en el análisis de las relaciones bilaterales entre predictores es el coeficiente de la correlación costo por hectárea y utilidad por hectárea, el cual tiene signo positivo y un valor de 0,501 con un nivel de confianza del 99%, es un resultado aparentemente contradictorio porque a un mayor nivel de costos de producción debería haber menor utilidad, es decir que la relación debería ser negativa. La explicación de este hallazgo se puede encontrar analizando las relaciones entre otras variables independientes, en este sentido, la relación positiva entre los costos por hectárea y la productividad es la más alta de todas (0,908) y a su vez, la relación entre la productividad y la utilidad por hectárea es la segunda (0,753), mientras que la relación entre el costo por hectárea y la utilidad por hectárea, siendo alta es inferior a la relación anterior, lo que significa que la influencia de la productividad sobre la utilidad por hectárea es mayor que la influencia del costo por hectárea.

Finalmente, pareciera ser que la relación de causalidad entre productividad y costos por hectárea es más fuerte en la dirección productividad – costo por hectárea, que, en dirección contraria, por lo tanto, cuando se presentan costos de producción por hectárea altos, no solamente no afectan negativamente la utilidad por hectárea, sino que por el contrario son el reflejo o la consecuencia de una alta productividad.

Similar resultado se obtuvo en un estudio que buscaba medir los factores que influyen en la rentabilidad del cultivo de camu camu en la selva peruana, el estudio encontró una débil correlación entre los costos de producción y la rentabilidad del cultivo; concluyendo que la variación en los costos de producción explicaría solo en un 15,7% la variación en la rentabilidad, y aunque los costos de producción y la rentabilidad no están correlacionados significativamente, mostraron una relación positiva (Flores y Miranda, 2017).

Un resultado contrario reveló el estudio de Cancino et al. (2018), los autores desarrollaron un modelo para explicar la rentabilidad del

cultivo de durazno, hallando una relación inversa entre los costos de producción y la rentabilidad económica, validando las hipótesis iniciales que apuntaban hacia un valor beta con signo negativo de la variable costos de producción.

El resultado obtenido en el presente estudio tiene sentido en sistemas productivos donde el mayor porcentaje de los costos de producción corresponde a costos variables (relacionados de manera directa con el volumen de producción), por lo tanto, los costos que estén asociados directamente al nivel de producción estarán afectados por sus variaciones. En el caso del café, este tiene un componente de costo variable superior al 60%, representado en los costos de recolección y beneficio (Londoño, 2020; Ospina et al.,2003).

## Minería de datos

El área cultivada en café se reveló como la variable de mayor influencia sobre la variable dependiente, posicionándose como el nodo raíz en el árbol de decisión, y es allí donde se localiza el regresor que más influencia tiene sobre la variable dependiente. El valor de 3,56 hectáreas la ratifican como una variable restrictiva para lograr el Living Income teniendo en cuenta que el área promedio cultivada en café, en Colombia, es de 1,56 hectáreas (Federación Nacional de Cafeteros, 2020a), sin embargo, este resultado debe tomarse como un posible punto de segmentación que facilite la implementación de estrategias diferenciadas para contribuir al cierre de la brecha.

Un modelo desarrollado por Fairtrade, para fijar precios de referencia de LI, consideró la participación de la mano de obra familiar como base para estimar que el área de cultivo mínima viable para producción de café convencional es de 2,8 hectáreas (Fairtrade International, 2021). Si bien este valor es inferior al que arroja el modelo RF, no es conveniente introducir el concepto de área mínima viable, porque esto significaría dejar por fuera de las estrategias de intervención a la mayoría de caficultores de Colombia, los cuales están en un rango de área inferior a 2,8 hectáosreas.

Con respecto al performance del modelo, evaluado en términos de la especificidad, se observó su alto desempeño para identificar los verdaderos negativos, lo cual es fundamental en el tipo de datos que se estaban analizando y el objetivo perseguido con su modelado; se requería un modelo que pudiera clasificar correctamente los verdade-

ros negativos (las fincas que no alcanzan el Living Income), porque las estrategias conducentes al cierre de la brecha se deben dirigir a este segmento de la población.

### Validación del modelo

Si bien el porcentaje de falsos negativos fue alto en ambos datasets, la especificidad, que es la métrica relevante fue del 100%, este resultado demuestra la bondad del modelo, puesto que estas fincas tuvieron un precio de venta promedio de 109.100 COP por arroba de cps (29,54 USD), que es superior 44% al precio de venta mínimo que se identificó en el RF, sin embargo, al no cumplir con las condiciones mínimas de área cultivada en café o de productividad, no alcanzaron el LI. La explicación para el alto porcentaje de falsos negativos que arrojó el modelo radica en el incremento sin precedentes que tuvo el precio del café en el año 2020, si bien el precio ha demostrado ser volátil; en el año 2020 tuvo un incremento del 33%.

Desde marzo de 2020, cuando se declaró el covid-19 como una pandemia, se presentó una fuerte volatilidad en los precios del café como resultado principalmente de perturbaciones en la cadena de suministro. Se identificaron dos efectos en el sector cafetero: en primer lugar, afectó al suministro de mano de obra, o bien directamente debido a enfermedad o indirectamente debido al limitado flujo de trabajadores agrícolas y trabajadores migrantes por medidas de distancia social, confinamientos y restricciones de desplazamiento. En segundo lugar, las perturbaciones en las redes logísticas internas ocasionaron retrasos de la exportación y aumento de los costos de comercio y transacción (ICO - International Coffee Organization, 2020).

Estos efectos causaron preocupación en el mercado a nivel mundial por un posible desabastecimiento a corto plazo, lo que a su vez generó la tendencia al alza de los precios durante 2020 y 2021.

Es importante tener en cuenta que el precio de venta tiene un poder explicativo del 25% en el modelo, por lo tanto, las variaciones extremas que tenga esta variable tendrán un efecto importante en el poder predictivo del mismo.

## Conclusiones

Las conclusiones de esta investigación son abordadas desde dos aristas: el entendimiento de los regresores y los resultados del modelo.

## Entendimiento de los regresores:

La alta correlación entre las variables productividad y costos por hectárea (0,908) que tiende a la colinealidad, desmitifica la creencia histórica de los costos de producción como la causa raíz de los problemas de la caficultura. Un estudio que evaluaba la eficiencia técnica de fincas cafeteras en Colombia concluyó que la eficiencia técnica de los productores colombianos era menor que la de Vietnam, lo cual podía tomarse como reflejo de la baja productividad y competitividad del sector cafetero en los mercados internacionales, y así mismo inducía a pensar que los caficultores en Colombia no estaban minimizando sus costos de producción y por ende tampoco maximizaban las ganancias que deberían obtener en la actividad cafetera (Perdomo et al., 2007) empleados en procesos productivos para conocer el maximo nivel producido y cantidad optima utilizada de insumos acorde con sus precios. El pre- sente estudio maneja datos microeconomicos de caficultores peque- flos, medianos y grandes en los departamentos de Caldas, Quindio y Risaralda, para determinar la eficiencia tecnica y asignativa mediante el metodo no parametrico Analisis Envolvente de Datos - DEA (Data Envolpment Analysis, siglas en ingles.

Un estudio precedente, esbozaba que cuando un agricultor puede vender su producto a menor precio que sus pares, se debe principalmente a que sus costos de producción son inferiores, según el autor, cuando los productores compiten en precio en el mercado deben acceder a tecnologías que apunten a reducir costos de producción, en este sentido el autor afirma: ”es necesaria la investigación económica sobre las prácticas de producción que ayuden a los caficultores en la comprensión de la mejor manera de reducir los costos unitarios de la producción obtenida”, de acuerdo con el estudio, bajo cualquier escenario, la selección de tecnologías es un factor determinante en la definición del costo de producción y por tanto, en la competitividad por precio de venta (Duque, 2004).

Si bien, para la cadena de valor es importante desplegar estrategias orientadas a la optimización de costos de producción, estas acciones pueden no ser tan efectivas para aumentar de manera significativa la utilidad de las fincas cafeteras, pues el impacto de intervenir los costos de producción no es contundente, debido al fuerte efecto de la productividad sobre el margen por hectárea (coeficiente de correlación de 0,753**) y sobre el costo de producción. Este tipo de hallazgos debe conducir a la industria a focalizar los objetivos de las acciones

de intervención orientadas a mejorar el ingreso neto de las familias cafeteras.

## Resultados del modelo

Aunque el regresor más influyente sobre el cierre de brecha es una variable que da poco margen de maniobra. El área en café tiene que ser precisamente el foco de atención de las estrategias orientadas al cierre de brecha, porque debe ser uno de los criterios de segmentación de las tipologías de caficultores. El hecho de que el negocio cafetero, en Colombia y en el mundo, sea un negocio de pequeños agricultores debe poner más presión para contribuir al logro del Living Income para las familias productoras, esta realidad implica evaluar los drivers del ingreso neto y del Living Income, estimar su influencia y desplegar estrategias de intervención cuya restricción no sea el área cultivada en café.

"Es importante establecer qué tanto y de qué forma dependen los pequeños propietarios del cultivo del café. Resolver este interrogante, requiere conocer la estructura de ingresos (enfatizando en su ingreso extrapredial y por otros cultivos), costos de producción y gastos familiares" (García y Ramírez; 2002).

De acuerdo con los resultados del modelo RF, el precio de venta es la variable con menor influencia sobre la brecha, no obstante, hay iniciativas bien intencionadas que promueven un precio de referencia de Living Income, según su definición es "el precio necesario para que un hogar de agricultores típico con un tamaño de finca viable y un nivel de productividad sostenible pueda ganarse la vida con las ventas de su cultivo" (Fairtrade International, 2021).

Existe un riesgo al centrarse en el incremento de precios como única estrategia para cerrar la brecha, puesto que los aumentos de precios generan beneficios a corto plazo para los agricultores, pero a largo plazo podría generar nuevamente presión sobre los precios. Las intervenciones de precios en los mercados globales tienen efectos limitados cuando solo se aplican a nivel nacional o regional. Si las políticas de cada país aumentan los precios, los compradores commodities pueden comprarlos en otros países donde los precios son más bajos. Además, el aumento de los precios hará que más agricultores los cultiven y estimulará el aumento de la producción (Waarts et al., 2019).

Una de las ramificaciones determinantes del árbol de decisión revela que una finca con un área inferior a 1,6 hectáreas cultivadas en café

y un precio de venta inferior a 27,77 USD por arroba de cps, no tiene probabilidad de alcanzar el Living Income. Este escenario debería remitir a la industria y a la institucionalidad cafetera a contemplar medidas diferentes a las estrategias enfocadas en el precio de referencia para cerrar la brecha, porque de lo contrario, estarían quedando fuera de la solución la inmensa mayoría de caficultores colombianos.

"Al observar las diferencias en los ingresos por tipo de explotación, resulta claro que la distribución por tamaño de las unidades de producción es una característica estructural que influye profundamente en el bienestar y en el nivel de rentabilidad de las familias caficultoras" (García y Ramírez; 2002).

## Agradecimientos

Los autores agradecen a solidaridad Network por la generosidad al autorizar el uso de la información recopilada durante 10 años, para la ejecución de este estudio.

A las familias cafeteras que con su trabajo incansable construyen esperanza para el país

A los aliados de la Plataforma de comercio sostenible quienes se han comprometido con el acompañamiento y soporte técnico a las familias cafeteras de Colombia

# Referencias

Ait Issad, H., Aoudjit, R., y Rodrigues, J. J. P. C. (2019). A comprehensive review of Data Mining techniques in smart agriculture. In Engineering in Agriculture, Environment and Food (Vol. 12, Issue 4, pp. 511–525). Elsevier B.V. https://doi.org/10.1016/j.eaef.2019.11.003

Anker, R., y Anker, M. (2017). Living Wages Around the World. In Living Wages Around the World. https://doi.org/10.4337/9781786431462

Araque, H. (2015). Variables Tecnológicas que Determinan la Productividad de las Fincas Cafeteras del Departamento de Caldas. http://bdigital.unal.edu.co/49567/

Aristizábal, C., y Duque, H. (2008). Identificación de los patrones de ingreso en fincas de economía campesina de la zona cafetera central de Colombia. Cenicafé, 59(4), 321–342. http://biblioteca.cenicafe.org/bitstream/10778/219/1/arc059%2804%29321-342.pdf

Brounen, J., De Groot, A., Isaza, C., y Van Keeken, R. (2019). The true price of climate-smart coffee- Quantifying the potential impact of Climate-Smart Agriculture for Colombian coffee. https://www.solidaridadnetwork.org/wp-content/uploads/migrated-files/publications/TP%20CSA%20Coffee%20COL.pdf

Cancino, S., Cancino, G., y Quevedo, E. (2018). Explanatory model of the economic profitability of peach crop in the province of pamplona, colombia. Económicas Cuc, 39(2), 63–76. https://doi.org/10.17981/econcuc.39.2.2018.04

Cano, C. G., Vallejo, C., Caicedo, E., Amador, J. S., y Tique, E. Y. (2012). El mercado mundial del café y su impacto en Colombia. Borradores de Economía; No. 710. http://repositorio.banrep.gov.co/handle/20.500.12134/5733

Cazacu, M., y Titan, E. (2021). Adapting CRISP-DM for social sciences. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *11*(2Sup1), 99-106. https://www.researchgate.net/publication/344732208_Adapting_CRISP-DM_for_Social_Sciences

Centro de comercio internacional. (2011). Guía del Exportador de Café. In Guía del Exportador de Café. https://doi.org/10.18356/8f83c5c4-es

DANE. (2021). Índice de precios al consumidor. In Boletín estadístico (Vol. 258). https://www.dane.gov.co/index.php/estadisticas-por-tema/precios-y-costos/indice-de-precios-al-consumidor-ipc/ipc-historico

DeVries, Z., Locke, E., Hoda, M., Moravek, D., Phan, K., Stratton, A., Kingwell, S., Wai, E. K., y Phan, P. (2021). Using a national surgical database to predict complications following posterior lumbar surgery and comparing the area under the curve and F1-score for the assessment of prognostic capability. Spine Journal, 0(0). https://doi.org/10.1016/j.spinee.2021.02.007

Duque, H. (2004). Como reducir los costos de producción en la finca cafetera (p. 103), Federación Nacional de Cafeteros - Cenicafé. https://www.cenicafe.org/es/publications/Como_reducir_los_costos_de_producción_en_la_finca_cafetera.pdf

Fairtrade International. (2021). Fairtrade Living Income Reference Prices for Coffee from Colombia. https://www.fairtrade.net/news/decent-coffee-prices-to-stay

Federación Nacional de Cafeteros. (2017). FNC en Cifras. https://federaciondecafeteros.org/static/files/FNCCIFRAS2017.pdf

Federación Nacional de Cafeteros. (2020a). Colombia Cafetera - Federación Nacional de Cafeteros. https://federaciondecafeteros.org/servicios-al-caficultor/colombia-cafetera/

Federación Nacional de Cafeteros. (2020b). Economía Cafetera. Ensayos Sobre Economía Cafetera, 2(6), 553–559. https://federaciondecafeteros.org/app/uploads/2019/12/Economía-cafetera-No.-32-Final-mayo-2020.pdf

Flores, J., y Miranda Edwin. (2017). Factores que influyen en la rentabilidad económica de la producción del cultivo de camu camu en la selva peruana. Angewandte Chemie International Edition, 6(11), 951–952. https://doi.org/10.26495/rtzh179.121610

García, J., y Ramírez, J. (2002). Sostenibilidad económica de las pequeñas explotaciones. In Ensayos sobre economía cafetera 15(18):73-89. 2002. http://biblioteca.cenicafe.org/handle/10778/762

García S., O. L. (2009). Nociones de Costeo ABC. In Administración Financiera: Fundamentos y Aplicaciones. https://www.academia.edu/10712342/Capítulo_Complementario_3_NOCIONES_DE_COSTEO_ABC

González Ferrer, V. (2011). Curvas receiver–operating characteristic y matrices de confusión en la elaboración de escalas diagnósticas. RevistaeSalud.com, 7(26). https://dialnet.unirioja.es/servlet/articulo?codigo=4201589

Heredia Gutiérrez, C. D. (2008). Metodología para implantar un sistema de costeo ABC a la industria de la confección. Dictamen Libre, 7, julio-diciembre, 10–30. https://dialnet.unirioja.es/servlet/articulo?codigo=5786249

Hira, S., y Deshpande, P. S. (2015). Data Analysis using Multidimensional Modeling, Statistical Analysis and Data Mining on Agriculture Parameters. Procedia Computer Science, 54, 431–439. https://doi.org/10.1016/j.procs.2015.06.050

ICO - International Coffee Organization. (2020). Impact of covid-19 on the global coffee sector: the demand side. American Health & Drug Benefits, 13(3), 1–9. https://www.ico.org/documents/cy2019-20/coffee-break-series-1e.pdf

Liu, S., Cossell, S., Tang, J., Dunn, G., y Whitty, M. (2017, May 1). A computer vision system for early stage grape yield estimation based on shoot detection. Computers and Electronics in Agriculture, 137, 88–101. https://doi.org/10.1016/j.compag.2017.03.013

Londoño, J. (2019). Costos de producción de café 2019 Colombia. https://solidaridadlatam.org/wp-content/uploads/attachments/200607_informecostos_2019.pdf

Londoño, J. (2020). Costos de producción de café 2020 - Colombia. https://solidaridadlatam.org/wp-content/uploads/2022/02/200607-informeCostos_2020.pdf

Lykke, P. , Andersen, E., Anker, R., y Anker, M. (2020). Informe sobre el salario vital Costa caribeña de Colombia Contexto: Sector bananero Mayo 2018 (incluye actualización hasta enero 2020) Preparado para: The Global Living Wage Coalition. https://www.globallivingwage.org/wp-content/uploads/2020/06/LW-Report_Colombia_2019_es.pdf

Moro, S., Laureano, R. M. S., y Cortez, P. (2011). Using data mining for bank direct marketing: An application of the CRISP-DM methodology. ESM 2011 - 2011 European Simulation and Modelling Conference: Modelling and Simulation 2011, 117–121. https://core.ac.uk/download/pdf/55616194.pdf

Nadali, A., Kakhky, E. N., y Nosratabadi, H. E. (2011). Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system. ICECT 2011 - 2011 3rd International Conference on Electronics Computer Technology, 6(January 2016), 161–165. https://doi.org/10.1109/ICECTECH.2011.5942073

Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B., y Turaga, D. (2017). Learning feature engineering for classification. IJCAI International Joint Conference on Artificial Intelligence, 2529–2535. https://doi.org/10.24963/ijcai.2017/352

Ospina, O., Duque, O., Farfán V. (2003) Análisis económico de la producción de fincas cafeteras convencionales y orgánicas en transición, en el departamento de caldas. Cenicafé, 54(3), 197-207, https://www.cenicafe.org/es/publications/arc054%2803%29197-207.pdf

Panhuysen, S. and Pierrot, J. (2021). Coffee Barometer 2020. https://coffeebarometer.org/

Perdomo, J. A., Huet, D., y Mendieta, J. C. (2007). Factores que afectan la eficiencia técnica y asignativa en el sector cafetero colombiano: una aplicación con análisis envolvente de datos. Revista Desarrollo y Sociedad, 60, 1–45. https://doi.org/10.13043/dys.60.1

Toorop, R., Ruiz, A., Maanen, E., Brounen, J., Casanova, L., Garcia, R. (2017). The true price of climate smart coffee - quantifying the potential impact of climate-smart agriculture for Mexican coffee. https://sustainablefoodlab.org/wp-content/uploads/2018/04/The-True-Price-Of-Climate-Smart-Coffee-Solidaridad.pdf

Saarela, M., y Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. SN Applied Sciences, 3(2), 1–12. https://doi.org/10.1007/s42452-021-04148-9

Schröer, C., Kruse, F., y Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526-534. https://www.sciencedirect.com/science/article/pii/S1877050921002416

Shin, J. S., Lee, W. S., y Ehsani, R. (2012). Postharvest citrus mass and size estimation using a logistic classification model and a watershed algorithm. Biosystems Engineering, 113(1), 42–53. https://doi.org/10.1016/j.biosystemseng.2012.06.005

Steiner, R., Salazar, N., y Becerra, A. (2015). La politica de precios del cafe en Colombia. http://hdl.handle.net/11445/3166

Surya, P., y Aroquiaraj, I. L. (2018). Crop yield prediction in agriculture using data mining predictive analytic techniques. International Journal of Research and Analytical Reviews, 5(4), 783-787. https://ijrar.org/papers/IJRAR1905206.pdf

Task force for coffee Livng Income. (2020). Strategy handbook: A Fact-Based Exploration of the Living and Pricing Strategies that Close the Gap. In IDH - the Sustainable Trade Initiative. https://www.idhsustainabletrade.com/uploaded/2020/02/TCLI_FR_3.2_Lres_singlepages5-full-report.pdf

The Living Income community of practices. (2020). Living Income | livingincome. https://www.living-income.com/

Veldhuyzen, C. (2020). Fairtrade Living Income Progress Report. https://files.fairtrade.net/2019_RevisedExplanatoryNote_FairtradeLivingIncomeReferencePriceCocoa.pdf

Waarts, Y. R., Janssen, V., Ingram, V. J., Slingerland, M. A., van Rijn, F. C., Beekman, G., ... y van Vugt, S. M. (2019). A living income for smallholder commodity farmers and protected forests and biodiversity: how can the private and public sectors contribute?: White Paper on sustainable commodity production (No. 2019-122). Wageningen Economic Research. https://library.wur.nl/WebQuery/wurpubs/556298

Yang, C. C., Prasher, S. O., Enright, P., Madramootoo, C., Burgess, M., Goel, P. K., y Callum, I. (2003). Application of decision tree technology for image classification using remote sensing data. Agricultural Systems, 76(3), 1101–1117. https://doi.org/10.1016/S0308-521X(02)00051-3