TEC | Tecnológico de Costa Rica

Escuela de Administración de Tecnologías de Información

Diseño de un modelo predictivo para la demanda y planificación de la producción de café en microbeneficios de Santa María de Dota

Trabajo final de graduación para optar al grado de Licenciatura en Administración de Tecnología de la Información

Modalidad seminario de graduación

Elaborado por Julio César Romero Chacón

Prof. tutor: M. Sc. Pedro Leiva Chinchilla

Cartago, Costa Rica

Semestre II, 2024

Noviembre, 2024

Hoja de aprobación

INSTITUTO TECNOLÓGICO DE COSTA RICA

ESCUELA DE ADMINISTRACIÓN DE TECNOLOGÍAS DE INFORMACIÓN.

GRADO ACADÉMICO: LICENCIATURA

Los miembros del Tribunal Examinador de la Escuela de Administración de Tecnologías de Información, recomendamos que el siguiente informe del Trabajo Final de Graduación del estudiante Julio Romero Chacón sea aceptado como requisito parcial para obtener el grado académico de Licenciatura de Tecnología de Información.

Firmado digitalmente por PEDRO IGNACIO LEIVA CHINCHILLA (FIRMA)
Fecha: 2024.11.29 14:54:21 -06'00'

Pedro Leiva Chinchilla
Profesor tutor

LORENA ZUÑIGA SEGURA (FIRMA)
PERSONA FISICA, CPF-01-0888-0245.
Fecha declarada: 26/11/2024 09:39:32 AM
Esta representación visual no es fuente de confianza. Valide siempre la firma.

Lorena Zúniga Segura
Lectora académica 1

Digitally signed by ISAAC ALPIZAR CHACON (FIRMA)
Date: 2024.11.26 07:26:12 +01'00'

Isaac Alpízar Chacón
Lector académico 2

TEC | Tecnológico de Costa Rica

Firmado digitalmente por YARIMA TATIANA SANDOVAL SANCHEZ (FIRMA)
Fecha: 2024.11.27 21:11:52 -06'00'

Yarima Sandoval Sánchez
Coordinadora del trabajo final de graduación

Dedicatoria

A mis padres, por su apoyo incondicional y por brindarme todas las herramientas necesarias para alcanzar mis metas. A mis hermanos, Javier y Enrique, por estar siempre presentes.

A mis amigos, José Blanco, Sebastián Córdoba, Kevin Rojas y Gustavo Calderón, gracias por su amistad y por formar, junto a mí, un excelente equipo de trabajo.

A mi profesor tutor, Pedro Leiva, por su constante motivación y orientación a lo largo de este proceso.

Con gratitud a todos ustedes, este logro también les corresponde.

Resumen

Este proyecto tiene como objetivo diseñar un modelo predictivo que apoye a los microbeneficios de café en Santa María de Dota en la planificación de la demanda y la producción, utilizando datos históricos junto con variables de mercado y del entorno.

La investigación sigue la metodología CRISP-DM, estructurando el proceso en fases de comprensión del negocio, análisis y preparación de datos, modelado y evaluación, sin implementar el modelo en un entorno operativo.

Tras consolidar el conjunto de datos a partir de fuentes relevantes, se aplicaron procesos de limpieza y tratamiento para garantizar su adecuación a los modelos predictivos. Además, se evaluaron seis técnicas de predicción en tres iteraciones, con el fin de identificar el modelo con mejor rendimiento.

Las redes neuronales, en su tercera iteración, mostraron un potencial prometedor en términos de precisión predictiva, al demostrar su capacidad para interpretar patrones complejos y no lineales en los datos, manteniendo una tasa de error relativamente baja. En el futuro, este modelo puede ayudar a reducir la incertidumbre en la producción de café y facilitar decisiones informadas en la planificación.

La principal conclusión es que el modelo seleccionado tiene el potencial de ser útil en los microbeneficios del café, lo que promueve un uso eficiente de los recursos y una gestión de la demanda más efectiva.

Palabras clave: modelo predictivo, CRISP-DM, café.

Abstract

The objective of this project is to design a predictive model to support coffee micromills in Santa Maria de Dota in demand and production planning, using historical data together with market and environmental variables.

The research follows the CRISP-DM methodology, structuring the process in phases of business understanding, data analysis and preparation, modeling, and evaluation, short of implementing the model in an operational environment.

After consolidating the data set from relevant sources, cleaning and treatment processes were applied to ensure its suitability for predictive models. Six predictive techniques were evaluated in three iterations to identify the best performing model.

Neural networks, in their third iteration, showed promising potential in terms of predictive accuracy, demonstrating their ability to interpret complex and nonlinear patterns in the data while maintaining a low error rate. In the future, this model could help reduce uncertainty in coffee production and facilitate informed decisions in planning.

The main conclusion is that the selected model has the potential to be useful in coffee micro-profits, promoting efficient use of resources and more effective demand management.

Keywords: Predictive model, CRISP-DM, Coffee.

Tabla de contenidos

Página

Índice de figuras

Página

Índice de tablas

Página

Capítulo 1. Introducción

En este capítulo se presentan los antecedentes y el contexto del sector cafetalero en Santa María de Dota, destacando el problema central de la investigación. Además, se justifica la importancia del estudio y se definen, tanto los objetivos generales como los específicos, así como los beneficios esperados para los productores y la comunidad local. Finalmente, se establece el alcance del proyecto investigativo.

1.1. Descripción general

El sector cafetalero desempeña un papel fundamental en la economía de Santa María de Dota. Muchos productores de la región optan por procesar directamente el fruto del café y vender el producto final, tanto en el mercado local como en el internacional. Según los datos del Icafé 1, la producción total de café en fruta en el cantón de Dota, medida en fanegas, aumentó de 38,406.0375 fanegas en la temporada 2017-2018 a 66,733.6650 fanegas en la temporada 2022-2023, lo que representa un crecimiento significativo. Además, la producción del cantón equivale al 3.5 % de la producción total de café del país.

Por otro lado, según el Icafé 1, en el ámbito nacional existen 211 firmas beneficiadoras que procesan volúmenes inferiores a 1,000 fanegas por temporada. Esto representa conjuntamente el 2.7 % de la producción nacional. Es decir, este sector ha experimentado un crecimiento notable en los últimos años. Sin embargo, estos pequeños beneficiadores enfrentan incertidumbres respecto a la cantidad de café que deben procesar para satisfacer la demanda futura del mercado.

Con el propósito de abordar este problema, se plantea una investigación destinada a desarrollar un modelo predictivo que asista a los microbeneficios cafetaleros de Santa María de Dota en la mejora de la planificación de la demanda y la producción. Dicho modelo utiliza datos históricos de producción, ventas y variables de mercado para generar predicciones fiables que faciliten la toma de decisiones.

Este documento constituye el trabajo final de graduación (TFG) en la modalidad de seminario de graduación. Se inicia con una introducción que contextualiza el estudio, define el problema y justifica el estudio. Enseguida, se revisa el estado del arte para fundamentar teóricamente el trabajo. El marco metodológico describe los métodos y procedimientos empleados en el proyecto investigativo. Por último, se presentan y analizan los resultados, para lo cual se destacan los hallazgos más relevantes y se discuten las limitaciones y problemas que se encontraron.

1.1.1. Antecedentes

A continuación, se presentan los antecedentes del proyecto, divididos en dos partes: los que se relacionan con la entidad donde se lleva a cabo la investigación y los antecedentes del sector cafetalero. Esta sección tiene como objetivo contextualizar el entorno en el que se desarrolla el proyecto investigativo.

1.1.1.1 Descripción de la entidad donde se lleva a cabo la investigación

La investigación se realiza en el Instituto Tecnológico de Costa Rica (TEC), específicamente en la Escuela de Administración de Tecnologías de Información. El TEC, una universidad pública fundada en 1971, es una institución nacional autónoma de educación superior universitaria. Sus actividades abarcan la docencia, el proyecto investigativo y la extensión en tecnología y ciencias afines para el desarrollo de Costa Rica. El TEC se creó mediante la Ley n.º 4.777 del 10 de junio de 1971 2.

Misión

Contribuir al desarrollo integral del país, mediante formación del recurso humano, la investigación y la extensión; manteniendo el liderazgo científico, tecnológico y técnico, la excelencia académica y el estricto apego a las normas éticas, humanísticas y ambientales, desde una perspectiva universitaria estatal de calidad y competitividad a nivel nacional e internacional 2.

Visión

El Instituto Tecnológico de Costa Rica seguirá contribuyendo mediante la sólida formación del talento humano, el desarrollo de la investigación, la extensión, la acción social y la innovación científico-tecnológica pertinente, la iniciativa emprendedora y la estrecha vinculación con los diferentes actores sociales a la edificación de una sociedad más solidaria e inclusiva; comprometida con la búsqueda de la justicia social, el respeto de los derechos humanos y del ambiente 2.

Valores

El TEC 2 destaca los siguientes valores, tanto en el ámbito institucional como en el individual.

Ámbito institucional:

- Compromiso con la democracia
- Libertad de expresión
- Igualdad de oportunidades
- Autonomía institucional
- Libertad de cátedra
- Búsqueda de la excelencia
- Planificación participativa
- Cultura de trabajo en equipo
- Comunicación efectiva
- Evaluación permanente
- Vinculación permanente con la sociedad

- Compromiso con la protección del ambiente y la seguridad de las personas.
- Compromiso con el desarrollo humano
- Rendición de cuentas

Ámbito individual:

- Respeto por la vida
- Libertad
- Ética
- Solidaridad
- Responsabilidad
- Honestidad
- Sinceridad
- Transparencia
- Respeto hacia todas las personas
- Cooperación
- Integridad
- Excelencia

Escuela de Administración de Tecnologías de Información

La carrera de Licenciatura en Administración de Tecnología de la Información forma parte de la Escuela de Administración de Tecnologías de Información, la cual surge como respuesta a la necesidad de las empresas de contar con profesionales informáticos capacitados para aplicar las mejores prácticas administrativas en la toma de decisiones. En esta disciplina, el profesional se especializa en el ámbito de la computación, con un enfoque en comprender, optimizar e innovar los procesos empresariales 3.

La carrera universitaria busca desarrollar en las personas estudiantes una variedad de habilidades, entre las que destacan: un interés por los negocios, la tecnología y las organizaciones; la capacidad de utilizar tecnologías para resolver problemas; el pensamiento abstracto y crítico; habilidades de comunicación para actuar como facilitadores entre TI y negocios; liderazgo; investigación y competencias sociales.

Equipo de trabajo

Los roles involucrados incluyen a un estudiante investigador, quien cuenta con el apoyo de un profesor de ATI. En la Tabla 1. Equipo de trabajo, se detalla cada uno de estos roles y sus respectivas funciones dentro del proyecto.

Tabla 1. Equipo de trabajo

Miembro del equipo de trabajo	**Rol en el proyecto**	**Funciones**
Estudiante	Investigador	Responsable de desarrollar y ejecutar las distintas fases del proyecto y sus respectivos entregables.
Profesor tutor	Supervisor de la investigación	Encargado de proveer orientación y supervisión al estudiante investigador. Su papel incluye ayudar en el diseño y la planificación del estudio, así como revisar y proporcionar retroalimentación sobre los informes.

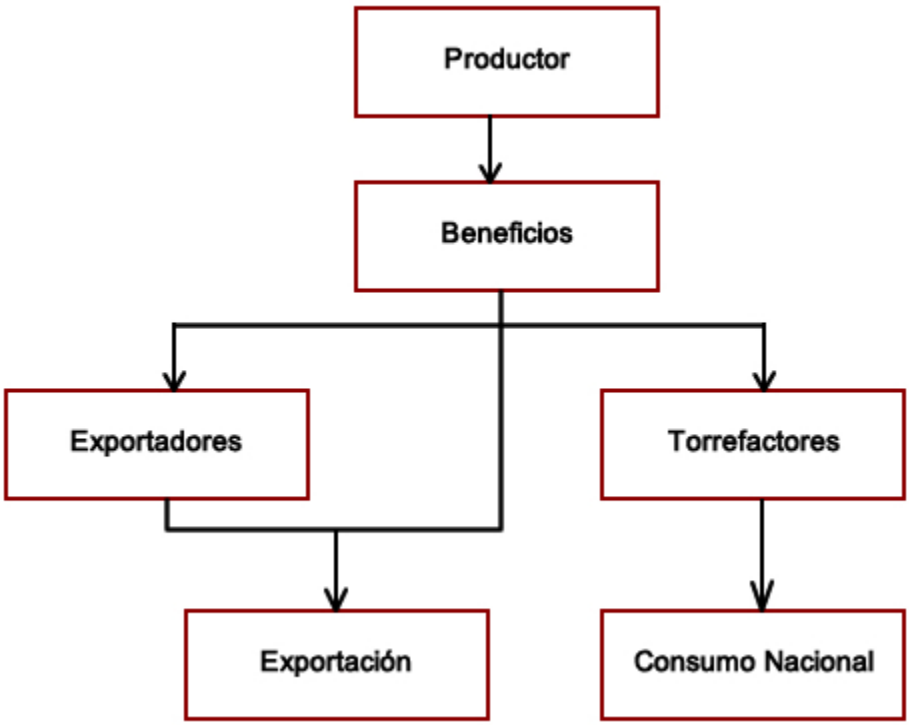
1.1.1.2 Antecedentes del sector cafetalero

A continuación, se presentan los antecedentes del sector cafetalero, explicando cómo está estructurado y quiénes son sus principales actores. Además, se describe el macroproceso de beneficiado del café.

Actores implicados

Dentro del encadenamiento productivo del sector cafetalero, se puede identificar una estructura común en cualquier escenario, ya sea para grandes productores, cooperativas o pequeños microbeneficios de café. Esta estructura revela los principales actores y sus relaciones de trabajo, funcionando como una jerarquía en el proceso de comercialización y producción del café. Los primordiales actores son: productores, beneficios, torrefactores, exportadores y consumidores nacionales. A continuación, se describe cada uno.

Ilustración 1. Estructura del sector cafetalero



Nota. Adaptado de "*Estructura del sector (Fotografía)*"*, tomado de Icafé (2015).*

- Productor: es toda persona con derecho legítimo para explotar una plantación de café y entregar el café en fruto al beneficiario. Según los datos del Icafé [4], el 92 % de los productores tiene plantaciones de café menores que cinco hectáreas (ha), lo que representa el 44 % del área total. El 6 % tiene entre 5 y 20 ha, abarcando el 21 % del área y el 2 % restante posee más de 20 ha, cubriendo el 35 % del área. Además, la densidad promedio de plantas por hectárea es de 7.000.

- Beneficiador: es el encargado de recibir, procesar, financiar y vender el café 4. Recibe la materia prima o café en fruta, de uno o varios caficultores o productores a través de los centros de acopio. En estas plantas, el café se transforma en café oro.

  Para lograr esta transformación, el grano pasa por un proceso húmedo que convierte las cerezas en café pergamino, sin mucílago, lavado y seco. Posteriormente, el café se almacena para transformarse en café oro o verde.

  Para esta investigación, los microbeneficios ejercen el rol de beneficiadores.

- Exportador: es la persona encargada de vender el café a mercados internacionales. Su función principal consiste en comprar el café al beneficiario o actuar como intermediario y, posteriormente, preparar y suministrar volúmenes de café a compañías importadoras o tostadoras que operan en los principales países consumidores. Según datos del Icafé 4 el

90 % del café producido en Costa Rica se exporta y esto representa un 15 % de las exportaciones totales del país.

- Torrefactor: son propietarios de establecimientos dedicados al tostado, molido u otros procesos industriales del grano, además de su comercialización en el ámbito nacional. En 1923 se fundó la primera industria torrefactora en Costa Rica, Café Volio, que aún existe en la actualidad. Después, en 1953, se estableció Café Rey, seguido por Café Dorado al año siguiente. Con el tiempo, se añadieron más torrefactoras, alcanzando un total de 73 empresas registradas actualmente 4.

Una vez que se identifican todos los actores en la cadena productiva del café, es importante aclarar que este estudio se enfoca específicamente en el microbeneficiado del café. Esto significa que no abarca todo el proceso productivo, sino que se concentra en el procedimiento de beneficiado.

Macroproceso del café

El proceso de beneficiado del café tiene como objetivo transformar la fruta del café en un producto final de alta calidad, conocido como café oro, que esté listo para su comercialización. Este procedimiento busca asegurar que todas las etapas se realicen de manera eficiente y cumplan con los estándares de calidad necesarios para satisfacer las demandas del mercado, tanto en el ámbito nacional como internacional.

Generalmente, el beneficiador de café es el responsable de todo el proceso, desde la recepción y el procesamiento de la fruta hasta el financiamiento y venta del producto final. Este rol implica no solo manejar el procedimiento físico del beneficiado, sino también asegurar la satisfacción de los exportadores, torrefactores y consumidores de café, quienes esperan recibir un café oro de alta calidad, procesado de manera eficiente y con transparencia.

El inicio del proceso se desencadena con la recepción de la fruta madura de café proveniente de los productores. A partir de ese momento, la fruta pasa por una serie de etapas, comenzando con su limpieza y la separación de los granos defectuosos. Después, se realiza el chancado o despulpado, seguido de un secado que dura aproximadamente 15 días. Tras el secado, el café se almacena en su estado de pergamino durante unos dos meses antes de ser pelado y transformado en café oro. Por último, se envían muestras a los compradores para determinar su calidad y se procede con la comercialización del café [5].

Para medir el rendimiento del proceso, se utilizan indicadores como el tiempo de ciclo, la cantidad de granos defectuosos (flotes) y el porcentaje de volumen perdido durante el procesamiento. Lo mencionado se detalla en el Apéndice D. Perfil del macroproceso de beneficiado de café

1.1.2. Situación problemática por analizar

En la zona de Los Santos, específicamente en Santa María de Dota, la caficultura constituye la principal actividad económica [6]. El modelo tradicional de trabajo consiste en que los productores se afilian a cooperativas o venden la fruta madura a compradores de café, quienes se encargan del proceso de beneficiado, que abarca desde el chancado hasta el secado y la posterior comercialización.

Sin embargo, hay productores que optan por tratar la fruta por sí mismos. Esta decisión les permite eliminar intermediarios y comercializar directamente el producto terminado a compradores locales, internacionales o exportadores, lo que implica mayores márgenes de ganancia para los productores.

A partir del insumo de que los microbeneficios de café comercializan su producto, se explica el proceso mediante el cual esto sucede. El comprador se encarga del beneficiado del café, convirtiéndolo en un producto terminado que se almacena en bodegas. Posteriormente, este café se negocia con compradores, con base en la calidad del producto. Sin embargo, los microbeneficios enfrentan un problema crítico, ya que no cuentan con una fuente confiable de información que les indique la cantidad de café que deben procesar. Por ende, se determina que el problema que atiende el proyecto es: "La incertidumbre en la cantidad de café por procesar que enfrentan los microbeneficios del cantón de Dota".

A continuación, se presenta el árbol del problema, el cual se fundamenta en la información obtenida mediante la reunión con el propietario de un microbeneficio de Santa María de Dota (ver Apéndice E. Minuta 1 Reunión inicial con un propietario del microbeneficio Tributos del Ota

Una de las causas que contribuyen a este problema es la volatilidad del mercado internacional. Esta volatilidad se debe a cambios en la demanda y en las preferencias de los consumidores, así como a la capacidad de producción de otros países. Los factores como sequías y otras fluctuaciones pueden afectar el mercado global, impactando a pequeños beneficios en Costa Rica.

La falta de comprensión del comportamiento del mercado en tiempo real a menudo conduce a negociaciones con los compradores con base en especulaciones en lugar de datos concretos. Esto significa que, en ocasiones, se recurre a criterios subjetivos, como la percepción individual, en lugar de utilizar información objetiva para determinar, por ejemplo, el precio del producto.

Otra de las causas es el mecanismo de negociación basado en la calidad del grano de café. Para vender el café es esencial conocer su calidad, la cual se determina mediante un proceso de catación, sin embargo, este análisis solo puede realizarse cuando el producto se ha procesado completamente. Esta necesidad de tener el producto terminado antes de conocer su calidad crea incertidumbre para los microbeneficios, ya que deben decidir cuánto café procesar sin una evaluación previa de su calidad, lo que complica la planificación y la comercialización eficiente.

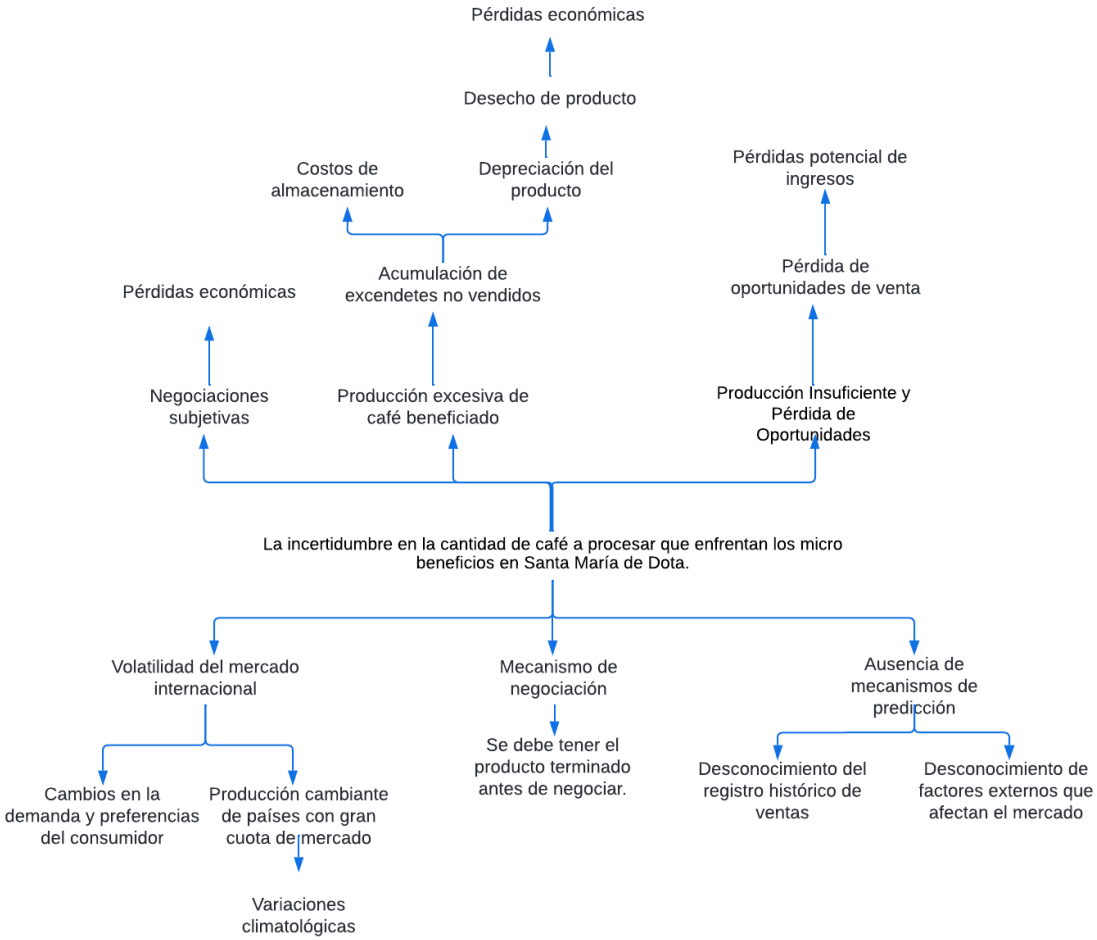
Además, la ausencia de mecanismos de predicción del comportamiento de estos mercados y el escaso aprovechamiento de los datos históricos de ventas generan incertidumbre. Sin herramientas adecuadas para prever tendencias y sin utilizar los datos existentes de manera efectiva, los microbeneficios no pueden anticipar la demanda ni planificar sus operaciones con precisión.

La falta de previsión puede llevar a una producción excesiva de café beneficiado, lo que ocasiona que se acumulen excedentes no vendidos. Esta situación puede generar pérdidas financieras debido a los costos de almacenamiento y, eventualmente, a la depreciación del producto almacenado durante largos periodos.

Por otra parte, también puede resultar en una escasez de producción, en la que los microbeneficios no procesan suficiente café para satisfacer la demanda del mercado. Esto implica oportunidades de venta desaprovechadas y una pérdida potencial de ingresos por no aprovechar completamente el mercado disponible.

En la Ilustración 2. Árbol del problema se presenta el árbol del problema, que busca representar las causas de este y también sus efectos.

Ilustración 2. Árbol del problema



1.1.3. Justificación del estudio

En este apartado se presenta la justificación del proyecto, enmarcada en las líneas de investigación de ATI y alineada con los ejes estratégicos del TEC. Además, se destaca la relevancia del proyecto en relación con los ODS establecidos por la ONU.

Primero, el trabajo se alinea con la cuarta línea de investigación aprobada por ATI: “Utilización de tecnologías innovadoras y disruptivas para mejorar la innovación y sofisticación empresarial”. Esta línea investigativa se enfoca en cómo la implementación de tecnologías avanzadas puede transformar y sofisticar los procesos empresariales, aumentar la eficiencia, reducir costos y optimizar la competitividad en el mercado.

Diseño de un modelo predictivo para la demanda y planificación de la producción de café en microbeneficios de Santa María de Dota

El concepto central de esta línea de investigación es el uso de herramientas tecnológicas de vanguardia para introducir mejoras significativas en las operaciones empresariales. Estas tecnologías pueden incluir inteligencia artificial, análisis de datos y soluciones con base en la nube, entre otras, que permiten a las empresas adaptarse rápidamente a las fluctuaciones del mercado y optimizar sus procesos.

El proyecto propuesto busca desarrollar un modelo predictivo que ayude a los microbeneficios de café en Santa María de Dota a anticipar la demanda y planificar la producción de manera eficiente. Al proporcionar una herramienta basada en datos para la toma de decisiones, se pretende reducir la incertidumbre en la cantidad de café por procesar y mejorar la capacidad de los productores para responder a las condiciones del mercado. Esta solución tecnológica innovadora no solo optimiza las operaciones de los microbeneficios, sino que también promueve prácticas más sostenibles y eficientes en la industria del café, alineándose así con la meta de sofisticación y mejora empresarial que persigue esta línea investigativa.

Este proyecto resulta idóneo para resolverse por un profesional en Administración de Tecnologías de Información, debido a su formación interdisciplinaria que combina conocimientos de tecnología, ciencia de datos, gestión y estrategia empresarial.

Para esto se requieren habilidades como la captura de requisitos, la cual es fundamental para desarrollar soluciones alineadas con las necesidades específicas de los microbeneficios cafetaleros. Además, se necesitan conocimientos de analítica empresarial, ya que en esta área se estudian diversos modelos clasificatorios y predictivos, así como marcos de trabajo aplicables a la minería de datos, como CRISP-DM, que permiten gestionar eficazmente cada etapa del desarrollo y obtener resultados óptimos.

Por otro lado, en cuanto a los ejes estratégicos del TEC, estos son áreas de conocimiento y objetos de estudio a través de los cuales la institución busca lograr su misión y fueron aprobados mediante la sesión ordinaria 105-2023 para el periodo 2023 a 2032 7.

Por ende, se considera que el proyecto se alinea con el eje estratégico de industria, el cual abarca el sector económico relacionado con la producción de bienes y la prestación de servicios. Este eje comprende los cuatro sectores de la industria moderna, siendo el primero la industria primaria, que extrae recursos naturales, lo que incluye las actividades agropecuarias, en las cuales está involucrado el sector cafetalero.

En relación con los Objetivos de Desarrollo Sostenible propuestos por la ONU, el proyecto contribuye primero al logro del objetivo n.º 2: poner fin al hambre, específicamente al cumplimiento de las siguientes metas 8:

- Meta 2.3: para 2030, se debe duplicar la productividad agrícola y los ingresos de los productores de alimentos en pequeña escala, en particular de las mujeres, los pueblos indígenas, los agricultores familiares, los pastores y los pescadores, entre otras cosas, mediante un acceso seguro y equitativo a las tierras, a otros recursos de producción e insumos, conocimientos, servicios financieros, mercados y oportunidades para la generación de valor añadido y empleos no agrícolas.

- Meta 2.c Adoptar medidas para asegurar el buen funcionamiento de los mercados de productos básicos alimentarios y sus derivados, así como facilitar el acceso oportuno a información sobre los mercados, en particular sobre las reservas. Lo anterior tiene el fin de limitar la volatilidad de los precios de los alimentos.

En segundo lugar, se contribuye al logro del objetivo n.º 12: garantizar modalidades de consumo y producción sostenible. Específicamente, se ayuda al logro de las siguientes metas 9:

- Meta 12.2: de aquí a 2030, se debe lograr la gestión sostenible y el uso eficiente de los recursos naturales.
- Meta 12.3: de aquí a 2030, se debe reducir a la mitad el desperdicio de alimentos per cápita en el ámbito mundial en la venta al por menor y en el ámbito de los consumidores, así como disminuir las pérdidas de alimentos en las cadenas de producción y suministro, incluidas las pérdidas posteriores a la cosecha.
- Meta 12.a: ayudar a los países en desarrollo a fortalecer su capacidad científica y tecnológica para avanzar hacia modalidades de consumo y producción más sostenibles.

1.1.4. Beneficios esperados

A continuación, se explican los beneficios directos e indirectos del proyecto:

Beneficios directos

- Se ayuda a los beneficiadores de café a planificar de manera eficiente la producción, lo que proporciona un sustento para la toma de decisiones informadas al conocer las tendencias del mercado y los datos históricos.
- Se fortalece la capacidad de análisis al identificar y recopilar factores clave que impactan la planificación de la producción, los cuales no solo contribuirán a mejorar las decisiones actuales, sino que también servirán como base de referencia para futuros periodos, lo que facilita la mejora continua del modelo propuesto.
- Se mejora e incentiva el uso de datos de calidad a través de la recolecta de información de fuentes fiables en la industria y mediante procesos de limpieza y transformación que garanticen su consistencia y fiabilidad.

Beneficios indirectos

- El proyecto sirve como un primer acercamiento a la innovación tecnológica en cuanto a planificar la producción de café en entornos pequeños y se ofrece una base para futuros desarrollos que pueden replicarse y perfeccionarse en otras regiones y por otros productores.
- Una mejor planificación y gestión de la producción ayuda a reducir los desperdicios y a optimizar el uso de recursos, lo que promueve prácticas agrícolas más sostenibles y amigables con el ambiente.

1.2. Objetivos del trabajo final de graduación

A continuación, se explican los objetivos del proyecto utilizando la taxonomía de Bloom.

1.2.1. Objetivo general

Diseñar, durante el segundo semestre de 2024, un modelo predictivo que apoye a los microbeneficios de café en Santa María de Dota para que se prevenga la demanda, lo que reduce la incertidumbre y optimiza sus operaciones para adaptarse a las fluctuaciones del mercado.

1.2.2. Objetivos específicos

1. Determinar los factores cualitativos y cuantitativos necesarios para que se alimente el modelo predictivo, identificando también las fuentes de información pertinentes, mediante la recolección de datos a través de informes y revisión documental.
2. Preparar los datos mediante un proceso de limpieza y transformación para que se asegure su calidad, consistencia y adecuación para su uso en el modelo.
3. Construir el modelo predictivo, por medio de un proceso iterativo de evaluación de métricas, con el fin de proporcionar a los microbeneficios de café una herramienta que les permita predecir la demanda.

1.3. Alcance de la investigación

En esta sección se describe detalladamente lo que se realiza en el proyecto. Por ende, se indica de forma explícita cuáles aspectos están incluidos en el alcance del trabajo y cuáles quedan excluidos.

1.3.1. Entregables

A continuación, se detallan los entregables del proyecto establecidos:

1. Estado del arte de la investigación que incluye estudios previos y conceptos clave desde la perspectiva de la minería de datos y desde la perspectiva del sector productivo.
2. Artículo científico que presenta de manera concisa los hallazgos más relevantes de la investigación.
3. Modelo predictivo inicial, desarrollado a partir de los factores relevantes recolectados, que tiene como objetivo prever la demanda y planificar la producción de café, acompañado de un informe sobre las pruebas de efectividad del modelo.
4. Trabajo final de graduación que abarca el desarrollo del proyecto, lo que incluye las fases ejecutadas, los resultados y las conclusiones detalladas.

1.3.2. Supuestos

En esta sección se indican de forma explícita cuáles son los factores que se asume se cumplirán o son ciertos en la realización del proyecto.

1. Se supone que los microbeneficios del café en el cantón de Dota y las instituciones implicadas proporcionan la información y los datos necesarios para el desarrollo del modelo.
2. Se asume que se dispone de suficientes registros y de calidad para entrenar el modelo predictivo mediante la técnica que se seleccionó.

1.3.3. Exclusiones

En esta sección se indican los entregables o productos que pueden esperarse del proyecto, pero que no forman parte de su alcance.

1. La integración del modelo predictivo en los procesos operativos diarios de los microbeneficios, es decir, la fase de despliegue de la metodología CRISP-DM, no se incluye.
2. No se extenderá el modelo predictivo a otras regiones ni a otros tipos de café que no sean los beneficiados en Santa María de Dota.
3. No se incluye el mantenimiento continuo ni las actualizaciones del modelo predictivo tras su desarrollo.
4. No se realiza un análisis de la rentabilidad ni de los costos financieros asociados con la implementación del modelo.
5. Debido al alcance limitado del proyecto, la fase de entendimiento del negocio se realiza una sola vez, omitiendo la naturaleza iterativa de CRISP-DM en esta etapa.

1.3.4. Limitaciones

En esta sección se indican los factores que, en alguna medida, restringen realizar el proyecto y sobre los cuales no se tiene control.

1. El proyecto se enfoca exclusivamente en los microbeneficios del café en Santa María de Dota, lo que puede limitar la aplicabilidad y generalización de las soluciones y los datos que se obtienen para otras regiones.
2. La disponibilidad limitada de las personas participantes puede dificultar su asistencia a algunas reuniones, lo que puede afectar la recopilación de datos y la colaboración en el proyecto.
3. Existe una alta dependencia de información externa, lo cual puede influir en el diseño del modelo, debido a que la disponibilidad y la calidad de los datos necesarios no están garantizadas.

Capítulo 2. Estado del arte

A continuación, se presenta el estado del arte, en el cual se investigan estudios previos y se exploran conceptos clave desde la perspectiva de la minería de datos y el sector productivo relevante para el estudio. Esta revisión proporciona una base sólida para el desarrollo del proyecto, garantizando la coherencia y el rigor metodológico en la implementación y el análisis.

2.1 Trabajos similares

Diversas investigaciones han abordado la predicción de la oferta y la demanda de diferentes cultivos, utilizando modelos predictivos que permiten optimizar la producción agrícola y maximizar los rendimientos económicos. Un ejemplo de este es el estudio de Garzón 10, centrado en la predicción de la oferta de aguacate Hass en el municipio de Herveo, Tolima. Este estudio utiliza modelos con base en series de tiempo para analizar datos históricos y proyectar el comportamiento futuro de la producción y la oferta del cultivo. En particular, se aplicaron métodos como los promedios móviles y el suavizamiento exponencial, los cuales permiten identificar patrones en los datos históricos y realizar predicciones más precisas. El estudio concluyó que la aplicación de estas técnicas mejora la toma de decisiones agrícolas y reduce los riesgos asociados con la variabilidad en la oferta de productos agrícolas.

En el caso del café, varios estudios recientes han investigado diferentes modelos predictivos para mejorar la planificación de la producción. Por ejemplo, Suhardi *et al.* 11 aplicaron métodos de análisis de series temporales, como medias móviles, medias móviles ponderadas y suavizamiento exponencial, para predecir la demanda de café tostado. Asimismo, Vijayan *et al.* 12 utilizaron el modelo Arima para pronosticar las tendencias futuras de la producción de café arábica y robusta en India, implementando los algoritmos en el lenguaje de programación R.

Además, se llevó a cabo una comparación entre diferentes modelos predictivos, como el suavizamiento exponencial, la media móvil y la regresión, en la producción de café en Filipinas. Los resultados mostraron que la media móvil era el método más preciso, ya que presentaba la tasa de error más baja en las predicciones 13.

Técnicas más avanzadas, como las redes neuronales artificiales (RNA) y la regresión lineal múltiple (MLR), también se han utilizado para predecir el rendimiento del café arábica. En un estudio de Kittichotsatsawat 14, las RNA mostraron una alta precisión en las predicciones (R^2 = 0,9524). Para llevar a cabo estas predicciones, se recolectaron datos de seis variables: áreas, zonas de productividad, precipitaciones, humedad relativa, temperatura mínima y temperatura máxima, a lo largo de 180 meses (15 años), desde 2004 hasta 2018.

Por último, Khumaidi 15 utilizó la metodología CRISP-DM junto con métodos de regresión lineal múltiple para desarrollar un modelo predictivo enfocado en la producción de café. Este modelo examina la relación entre variables independientes, como la superficie de plantación, las precipitaciones, la presión atmosférica y la radiación solar, con la producción de café. La calidad del modelo se evaluó mediante el valor de la raíz del error cuadrático medio (RMSE), lo que permitió medir su precisión.

Estos estudios destacan el potencial de diversos modelos predictivos para optimizar la planificación de la producción de café y satisfacer la demanda del mercado.

El valor agregado de este estudio frente a las investigaciones previas radica en la inclusión de una mayor cantidad de variables en el entrenamiento de los modelos predictivos. Por otro lado, la mayoría de los estudios existentes se enfoca en modelos básicos de series de tiempo, específicamente en modelos univariables que solo capturan tendencias en una única serie temporal, el presente trabajo explora modelos multivariables capaces de captar las variaciones específicas de cada microbeneficio de café, así como de incorporar otras variables de entorno y de mercado.

Además, esta investigación utiliza la metodología CRISP-DM, que proporciona un marco estructurado para proyectos de minería de datos. Esto asegura que el proceso sea robusto y que los datos que llegan a la etapa de modelado sean de alta calidad y útiles para maximizar el rendimiento de los modelos. Asimismo, se adopta un enfoque iterativo para ajustar los parámetros de los modelos. Lo anterior tiene el fin de identificar aquellos que ofrecen los mejores resultados en términos de precisión y capacidad explicativa.

2.2 Microbeneficiado del café en Costa Rica

A continuación, se presentan algunos conceptos relevantes del microbeneficiado del café que son importantes para comprender la dinámica y los procesos involucrados en esta industria. Estos términos son esenciales para quienes participan en la cadena de valor del café, desde la producción hasta la comercialización y permiten apreciar de mejor forma los retos y oportunidades que enfrenta el sector.

Tabla 2. Conceptos del sector productivo

Concepto	**Definición**
Variedades de café	Diferentes subespecies de plantas de café, como Bourbon, Geisha o Caturra. Cada variedad tiene características únicas en cuanto a sabor, resistencia a enfermedades y adaptabilidad al clima.
Despulpado o chancado	Proceso en el que se elimina la cáscara o pulpa del fruto del café (cereza) para liberar las semillas (granos de café).
Lavado	Proceso de lavar el café para eliminar cualquier resto de mucílago. Este procedimiento puede realizarse con abundante agua y es típico de los cafés *lavados*.
Secado	Los granos de café, ahora sin mucílago, deben secarse para reducir su contenido de humedad al nivel adecuado. Esto puede hacerse al sol (en

Concepto	**Definición**
	patios o camas elevadas) o mediante secadores mecánicos.
Café en pergamino	Es el estado en el que se encuentra el grano después del secado, aún envuelto en una fina capa conocida como *pergamino*.
Pelado	Proceso en el que se elimina la capa de pergamino seco que envuelve al grano de café. El resultado es el grano de café verde, que es la forma en la que se comercializa antes de la tostión.
Café oro	Es el café que ha sido trillado y está listo para ser tostado. Es conocido como café verde en otras partes del mundo.
Catación	Proceso de evaluación sensorial de los cafés mediante la cata. Esto incluye evaluar el aroma, sabor, acidez, cuerpo y otros atributos.
Cajuela	Unidad de medida que se utiliza en la recolección de café. Esta es una caja de tamaño estandarizado que se usa para medir la cantidad de café en cereza (fruta) que un recolector ha cosechado. Una cajuela generalmente equivale a 12.9 kg de café en cereza y es una unidad común en la etapa de recolección antes de que el café sea procesado.
Fanega	Unidad de medida que se utiliza en la industria cafetalera para cuantificar la cantidad de café en fruta. Una fanega equivale a 258 kg de café en fruta o 20 cajuelas y, aproximadamente, 46 g de café oro.
Quintal	Unidad de medida que se utiliza principalmente para cuantificar grandes cantidades de café procesado. Un quintal de café puede referirse a diferentes equivalencias según el contexto: por lo general, se refiere a 46 kg de café oro (grano seco listo para la exportación) o 100 libras (aproximadamente 45.36 kg).

2.3 Métodos de minería de datos

La minería de datos es el proceso de extraer información valiosa a partir de grandes volúmenes de datos. Esta información se obtiene mediante el descubrimiento de patrones y relaciones ocultas en los datos, utilizando una combinación de análisis matemático, estadístico y algoritmos 16. En esencia, la minería de datos transforma datos inicialmente poco útiles en conocimiento aplicable para la toma de decisiones. Esta capacidad ha convertido este proceso en un componente esencial en sectores como la agricultura, la salud y la gestión empresarial, entre otros.

Una de las principales razones por las que CRISP-DM se ha convertido en una de las metodologías ampliamente utilizadas en minería de datos es su flexibilidad. CRISP-DM no se limita a aplicaciones, herramientas o problemas específicos, lo que significa que puede aplicarse en cualquier sector y con cualquier tipo de *software* 17. Esta versatilidad la hace accesible y útil en una amplia variedad de escenarios. Además, su enfoque iterativo permite el refinamiento continuo del modelo a medida que se obtienen nuevos datos, lo cual resulta crucial en entornos cambiantes.

Aunque existen otras metodologías como SEMMA (Sample, Explore, Modify, Model, Assess), desarrollada por SAS Institute y KDD (Knowledge Discovery in Databases), CRISP-DM se distingue por incluir una fase de entendimiento del negocio. Esta fase se centra en comprender los objetivos del proyecto desde una perspectiva empresarial, lo cual facilita lograr una mejor alineación entre el análisis de datos y las metas organizacionales 18.

En el siguiente apartado se exponen los conceptos relevantes para la investigación desde la perspectiva de minería de datos, lo que incluye la metodología de minería de datos CRISP-DM y los modelos de predicción con base en series de tiempo. Esta revisión proporciona el marco teórico necesario para comprender y aplicar estas metodologías y técnicas en el contexto del beneficiado del café.

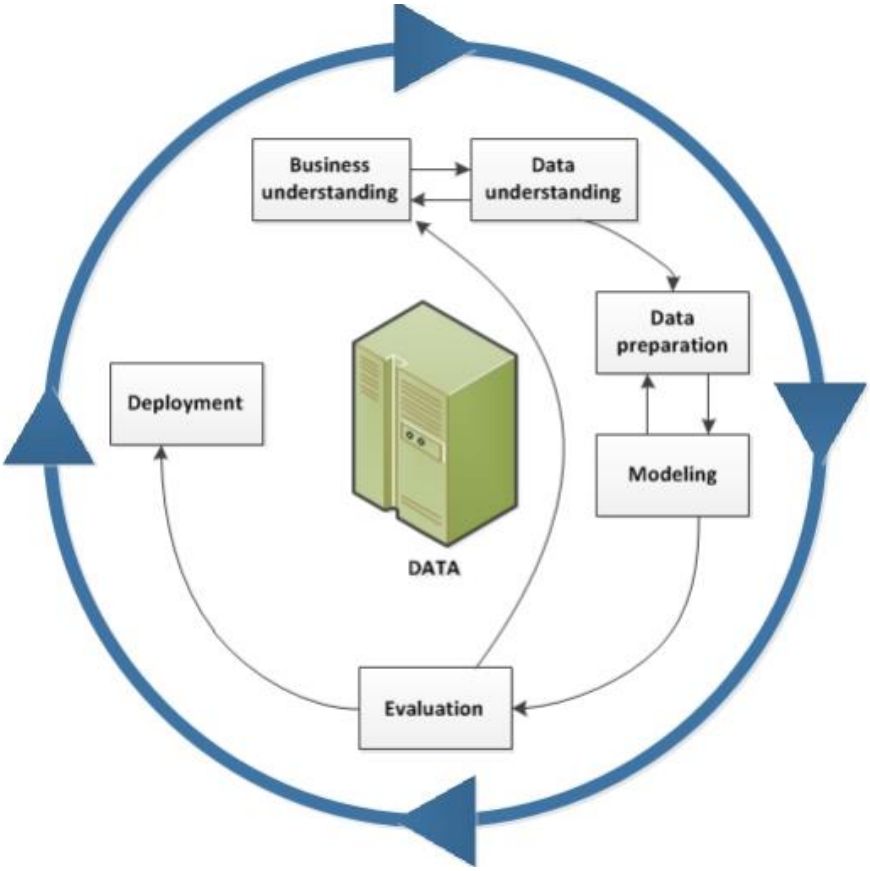
2.3.1 Metodología CRISP-DM

CRISP-DM (CRoss Industry Standard Process for Data Mining) se desarrolló a fines de 1996 por Daimler-Benz, ISL, NCR y OHRA con el objetivo de establecer un modelo estándar para la minería de datos. Sin embargo, no fue sino hasta el año 2000 que se presentó la versión 1.0, consolidándose como un modelo ampliamente aceptado en la industria 17.

La metodología describe el ciclo de vida de la minería de datos 19 y establece tareas que son necesarias en la mayoría de los proyectos de minería de datos, desde comprender el contexto en el que se está inmerso hasta el despliegue final del modelo en ambientes operativos.

Debido a que las necesidades empresariales están en constante cambio, CRISP-DM permite un proceso repetitivo y no necesariamente lineal, que soporta la mejora de los modelos a través del tiempo y da la posibilidad de retroceder a las diversas etapas si es necesario (ver la Ilustración 3. Ciclo de vida de la minería de datos).

Ilustración 3. Ciclo de vida de la minería de datos



Nota. Adaptado de ciclo de vida de minería de datos (Fotografía), tomado de IBM (2018).

Las etapas de la metodología son: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación e implementación. Enseguida, se explican estas fases. Es importante mencionar que este estándar es adaptable y puede requerir ciertas actividades específicas en contextos de minería particulares que no son aplicables a todos los proyectos. Por lo tanto, la ejecución de cada tarea depende de las necesidades de cada proyecto.

Entendimiento del negocio

Esta fase se centra en comprender las necesidades del cliente. Como en cualquier proyecto, la primera actividad consiste en establecer los objetivos empresariales, es decir, lo que el cliente desea conseguir o conocer y los criterios de éxito para lograrlos. Posteriormente, se evalúa la situación, al identificar los datos disponibles, los recursos humanos y los riesgos asociados. Luego, se especifican los objetivos de minería de datos, traduciendo los objetivos empresariales en objetivos alcanzables mediante la minería de datos desde un punto de vista técnico 20. Es recomendable que estos objetivos sean medibles y concretos. La última tarea de esta fase consiste en producir un plan de proyecto, que incluye un cronograma con las tareas necesarias y los recursos asociados para realizarlas 21.

Comprensión de los datos

Durante esta etapa, se recopilan los datos necesarios para el estudio a través de las fuentes establecidas, ya sean internas o externas a la organización. Asimismo, se describen los datos mediante sus propiedades o metadatos, como el número de registros o el formato, para determinar si la cantidad es suficiente o si se requieren adicionales. Además, se realiza una exploración de los datos por medio de visualizaciones o por medio de la búsqueda de relaciones entre ellos y se comprueba su calidad, evaluando la existencia de valores nulos, la dispersión de los datos, la relevancia de los datos y otras inconsistencias 20.

Preparación de los datos

Esta etapa incluye la selección de los datos que se conservarán para el estudio y la limpieza de estos para asegurar su calidad. Durante la limpieza de los datos, se pueden emplear diferentes tipos de soluciones, según los problemas de calidad, como reemplazar valores nulos con la moda, la mediana o la media de los datos o eliminar registros inconsistentes. Además, puede ser necesario obtener datos adicionales derivados de los existentes, lo que requiere su construcción y, si hay diversas fuentes de datos, integrarlos mediante fusión (agregando columnas) o adición (agregando filas nuevas) 21. Finalmente, se revisa el formato, ya que el modelo potencial puede requerir datos en una forma específica.

Modelado

En esta fase, los datos preparados se utilizan en el modelo elegido, buscando que los resultados logren los objetivos propuestos en la etapa de comprensión del negocio 21. Este proceso se repite varias veces, probando distintos modelos o ajustando los parámetros para lograr una mayor precisión; incluso puede retroceder a la etapa de preparación de los datos.

Se comienza seleccionando los modelos de minería de datos o algoritmos que se probarán, para lo cual se consideran sus requerimientos, como la cantidad recomendada de registros, el tipo de datos que soportan y si requieren datos de entrenamiento y prueba. Luego, se establecen los criterios para evaluar los modelos, tales como la tasa de error en las predicciones o clasificaciones 22. Posteriormente, se construyen los modelos a partir de los datos de entrenamiento o del conjunto de datos definido y se evalúan con los datos de prueba, ajustando los parámetros para mejorar los resultados.

Evaluación

En esta etapa se evalúa si los modelos cumplen con los criterios de éxito establecidos por el negocio, es decir, si satisfacen las necesidades y ayudan en la toma de decisiones empresariales 22. Tras este proceso, se seleccionan los modelos que cumplen los objetivos para incluirlos en el informe final. Además, es necesario realizar una retrospectiva para aprender de posibles errores y considerar formas de simplificar el procedimiento. Según los resultados, se puede optar por continuar con la fase de despliegue, refinar los modelos existentes o descartarlos 20.

Despliegue

En esta fase, el nuevo conocimiento se traslada a un ambiente operativo empresarial, implementando las mejoras 21. Se requiere un plan de despliegue y un plan de monitoreo y mantenimiento para mantener el modelo vigente de acuerdo con las necesidades empresariales. La complejidad de esta etapa varía según los requerimientos, que pueden ir desde la generación de un reporte hasta la integración del modelo con sistemas empresariales. Finalmente, se elabora un informe final que detalla el problema original, el procedimiento realizado, los costos asociados, los resultados, el plan de despliegue y las recomendaciones para futuros proyectos de minería de datos.

2.3.2 Modelos predictivos

En esta sección se estudian diferentes modelos predictivos que pueden ser útiles en la etapa de modelado. Asimismo, se presentan modelos con base en series de tiempo, como la media móvil, la media móvil ponderada y el suavizamiento exponencial, así como modelos más avanzados, como Arima y sus variantes Sarima y Arimax. Además, se incluyen modelos multifactoriales, como redes neuronales, *random forest*, *gradient boosting* y *k-nearest neighbors* (KNN), que ofrecen mayor flexibilidad para capturar relaciones no lineales y multifactoriales. Estos modelos constituyen un conjunto de herramientas potenciales para abordar problemas predictivos de diversas naturalezas y se evaluaron según su aplicabilidad en el contexto del proyecto.

2.3.2.1 Modelos con base en series de tiempo

Los modelos predictivos con base en series de tiempo resultan especialmente relevantes para el presente estudio, ya que utilizan datos históricos para predecir tendencias futuras, incorporando factores como la no linealidad y la ausencia de un patrón de comportamiento definido. Estos aspectos son, en la mayoría de los casos, inherentes a la agricultura, tales como la estacionalidad, las condiciones climáticas y las fluctuaciones en los mercados.

Entre los modelos básicos se destaca la media móvil, la media móvil ponderada y el suavizamiento exponencial, los cuales se explican en detalle a continuación. Este estudio es fundamental, ya que proporciona las bases para seleccionar los modelos predictivos más adecuados según diversos factores. Además, estos modelos primordiales constituyen el fundamento para el desarrollo de modelos predictivos más complejos, como Arima, Sarima o Arimax.

Media móvil

Este método utiliza el promedio de valores pasados para predecir valores futuros. En este enfoque, todos los valores pasados tienen el mismo peso en la determinación del valor futuro. La idea central de este modelo se basa en considerar que el valor actual de la serie puede preverse a partir de los errores pasados, es decir, la diferencia entre los valores reales y los pronosticados.

La media móvil se define como 11:

$$F_t = \frac{\sum Demanda\ anterior\ en\ el\ periodo\ n}{n} = \frac{A_{t-1} + A_{t-2} + \cdots + A_{tn}}{n}$$

Donde:

Tabla 3. Leyenda de variables de la fórmula de la media móvil

Valor	Definición
F_t	El valor previsto para el siguiente periodo
A_{t-i}	Valor actual del periodo anterior
n	Número de periodos usados

Media móvil ponderado:

A diferencia del método anterior, este otorga mayor importancia a los valores recientes. La media móvil ponderada está definida por 11:

$$F_t = \frac{\sum(\ Peso\ para\ el\ periodo\ n)(demanda\ en\ el\ periodo\ n)}{\sum peso} =$$

$$\frac{W_1 A_{t-1} + W_2 A_{t-2} + \cdots + W_n A_{tn}}{W_1 + W_2 + \cdots + W_3}$$

Donde:

Tabla 4. Leyenda de variables de la fórmula de la media móvil ponderada

Valor	Definición
F_t	El valor previsto para el siguiente periodo
A_{t-i}	Valor actual del periodo anterior
W_1	Valor del peso
n	Número de periodos usados

Suavizamiento exponencial

Este método, al igual que el de media móvil ponderada, otorga mayor relevancia a los datos más recientes. Sin embargo, a diferencia de la media móvil ponderada, los pesos no se asignan de manera explícita, sino que se determinan de forma exponencial. Es decir, el valor de cada peso aumenta o disminuye exponencialmente en lugar de definirse de forma directa. La fórmula correspondiente a este método es la siguiente 11:

$$F_t = F_{t-1} + a(A_{t-1} - F_{t-1})$$

Donde:

Tabla 5. Leyenda de variables de la fórmula del suavizamiento exponencial

Valor	definición
F_t	El valor previsto para el siguiente periodo
F_{t-1}	El valor previsto para el periodo anterior
A_{t-i}	Valor actual del periodo anterior
a	Constante entre 0 y 1

Arima

El modelo Arima (*autoregressive integrated moving average*) es una herramienta versátil para analizar y pronosticar series temporales, utilizando datos históricos como insumo 23. A diferencia de modelos más simples, Arima presenta una mayor complejidad, ya que integra tres componentes principales 24:

1. Autorregresión (AR): determina cuántos valores pasados se deben utilizar para realizar predicciones.
2. Integración (I): establece cuántas diferencias son necesarias para que los datos sean estacionarios.
3. Media móvil (MA): define cuántos errores pasados se deben considerar.

Existen variaciones del modelo Arima que pueden ser útiles según la naturaleza de los datos disponibles:

1. Sarima (*seasonal autoregressive integrated moving average*): extiende Arima al incorporar la estacionalidad, lo que permite ajustar los modelos para datos con patrones estacionales regulares o fluctuaciones periódicas.
2. Arimax (*Arima with exogenous variables*): permite la inclusión de variables externas que pueden influir en la serie temporal, lo que proporciona una mayor flexibilidad para incorporar información adicional que afecta las predicciones.

2.3.2.2 Otros modelos multifactoriales

En el contexto de los modelos predictivos, además de los modelos con base en series de tiempo, es importante destacar otras técnicas que ofrecen una perspectiva integral al considerar múltiples factores que pueden influir en la producción. Enseguida, se explican cuatro modelos adicionales que pueden considerarse posteriormente para implementarse en este proyecto: redes neuronales, *random forest*, *gradient boosting* y *k-nearest neighbors* (KNN). Estos modelos pueden ofrecer una mayor flexibilidad para abordar problemas no lineales y multifactoriales, lo que resulta útil para capturar interacciones complejas entre múltiples variables.

Redes neuronales

Las redes neuronales son modelos no lineales que imitan el funcionamiento del cerebro humano para aprender patrones complejos 25. En el contexto de la predicción de la demanda de café, las redes neuronales pueden aprender de múltiples variables, tales como la temperatura, la precipitación, los precios de mercado y los datos históricos de producción, para generar predicciones más precisas. Este modelo presenta la ventaja de ajustarse adecuadamente a datos complejos y no lineales. No obstante, su principal desventaja radica en que requiere un gran volumen de datos para entrenarse de manera adecuada, lo que puede representar un desafío si no se dispone de una base de datos robusta.

Random forest

El modelo de *random forest* es un método de ensamble que combina múltiples árboles de decisión para mejorar la precisión de las predicciones 26. Este modelo resulta útil cuando se dispone de un conjunto de datos con muchas variables, ya que puede manejar interacciones complejas entre ellas sin requerir la normalización previa de los datos. Además, *random forest* es robusto frente al sobreajuste, lo que lo convierte en una alternativa sólida para predecir la demanda de café. Sin embargo, dicho modelo puede volverse computacionalmente costoso al trabajar con grandes volúmenes de datos y su interpretación no es tan directa como la de otros modelos más simples.

Gradient *boosting*

Gradient *boosting* es otro modelo de ensamble que optimiza el rendimiento al combinar múltiples árboles de decisión, sin embargo, a diferencia de *random forest*, este modelo construye los árboles de manera secuencial. Cada árbol corrige los errores del anterior, lo que lo convierte en un modelo potente para datos con patrones complejos 27. En el contexto del café, *gradient boosting* puede ayudar a capturar relaciones no lineales y a corregir predicciones erróneas que otros modelos pueden no identificar. No obstante, su desventaja radica en que puede ser susceptible al sobreajuste si no se ajustan correctamente sus parámetros, lo que requiere un control riguroso.

K-nearest neighbors (KNN)

El modelo *k-nearest neighbors* (KNN) es un método basado en la proximidad entre los datos. Para predecir la demanda de café, KNN busca los datos más cercanos en el conjunto histórico y basa la predicción en la media de esos valores. Este modelo es fácil de implementar y entender y funciona bien en contextos donde las relaciones entre las variables no son complejas. No obstante, su principal limitación es que puede resultar ineficiente para grandes volúmenes de datos, ya que necesita revisar todos los registros para realizar una predicción, lo que lo vuelve lento. Además, puede ser más eficiente para tareas de clasificación que para asignaciones de predicción 28.

Capítulo 3. Marco metodológico

A continuación, se describe el marco metodológico que incluye el tipo de investigación, enfoque, diseño, población, fuentes, sujetos, variables e instrumentos de investigación. Asimismo, se establece el procedimiento metodológico de la investigación y, por último, se presenta una tabla resumen que sintetiza dicho procedimiento.

3.1 Tipo de investigación

Según los tipos de investigación establecidos por Baena 29, este estudio se clasifica como investigación aplicada o utilitaria. Este tipo se centra en el análisis de problemas con el objetivo de encontrar soluciones prácticas y directas. En este caso, se eligió la investigación aplicada porque el proyecto busca resolver problemas específicos y mejorar las operaciones de los pequeños beneficiadores de café y ayudarles a adaptarse a las fluctuaciones del mercado.

3.2 Enfoque de investigación

Según Ulate y Vargas 30, existen tres enfoques de investigación: cualitativa, cuantitativa y mixta. La investigación cualitativa se realiza de manera abierta y se basa en las experiencias y puntos de vista de los individuos; la cuantitativa se fundamenta en la recolección de datos mediante la medición y el análisis estadístico y, por último, la mixta implica recopilar datos tanto cuantitativos como cualitativos, así como su integración y discusión conjunta.

Por lo tanto, se considera que el enfoque de investigación más adecuado para el proyecto es el mixto. Este enfoque permite recolectar datos sobre las experiencias y necesidades específicas de los microbeneficios de café, así como identificar patrones y tendencias en la demanda de café mediante mecanismos de análisis cuantitativo.

3.3 Diseño de la investigación

Las modalidades de diseño de investigación consideradas en este estudio son las propuestas por Hernández Sampieri 31 para la investigación mixta. Entre ellas se encuentran:

- Diseño exploratorio secuencial (Dexplos): este diseño implica una fase inicial de recolección y análisis de datos cualitativos, seguida de una fase de recolección y análisis de datos cuantitativos.
- Diseño explicativo secuencial (Dexplis): comienza con la recolección y el análisis de datos cuantitativos y continúa con la recopilación y la evaluación de datos cualitativos.
- Diseño transformativo secuencial (Ditras): este diseño sigue un enfoque similar al de los anteriores, pero está guiado por una perspectiva teórica o ideológica transformativa.
- Diseño de triangulación concurrente (Ditriac): en este diseño, se recolectan datos cualitativos y cuantitativos de manera simultánea para validar, corroborar o complementar los hallazgos de ambos enfoques.
- Diseño anidado o incrustado concurrente de modelo dominante (DIAC): en este enfoque, uno de los métodos (cualitativo o cuantitativo) predomina y guía el estudio, mientras que el otro se incorpora para enriquecer y complementar la investigación recolectándose ambos tipos de datos al mismo tiempo.

- Diseño anidado concurrente de varios niveles (Diacniv): en este caso, se recopilan datos cualitativos y cuantitativos en diferentes niveles. Por ejemplo, en un nivel se obtienen y analizan datos cualitativos, mientras que en el siguiente nivel se recogen y evalúan datos cuantitativos y así sucesivamente.
- Diseño transformativo concurrente (Distrac): similar al diseño transformativo secuencial, pero los datos cualitativos y cuantitativos se recolectan de manera simultánea en lugar de secuencial, siempre con una perspectiva teórica o ideológica.
- Diseño de integración múltiple (DIM): combina el enfoque cualitativo y cuantitativo en múltiples niveles y etapas del proceso investigativo, buscando una integración continúa, desde la recolección de datos hasta el análisis y la interpretación.

Por lo tanto, se considera que el diseño de investigación más adecuado para este contexto es el diseño de integración múltiple. Este enfoque permite combinar de manera iterativa métodos de investigación cualitativos y cuantitativos, lo que ofrece una visión más completa del tema y aborda el problema desde diferentes perspectivas.

3.4 Población y muestra del estudio

A continuación, se define una muestra de estudio que sea representativa de la población total. La población objetivo de este estudio son los microbeneficios de café de Santa María de Dota, es decir, empresas que procesan volúmenes inferiores a 1.000 fanegas por año, estas empresas constituirán el insumo necesario para la contrucción del modelo. Según la información brindada por Icafé [1], existen 12 firmas beneficiadoras que cumplen con este requisito. Para determinar la muestra se utiliza la siguiente fórmula, la cual corresponde al muestreo aleatorio simple con población finita conocida:

$$Tamaño\ de\ la\ muestra = \frac{\frac{z^2 * p(1-p)}{e^2}}{1 + (\frac{z^2 * p(1-p)}{e^2 N})}$$

Donde:

- N es el tamaño de la población.
- E es el margen de error.
- Z es el nivel de confianza.
- p representa la proporción de la población que pertenece al estrato h.

$$11{,}63 = \frac{\frac{1{,}96^2 * 0{,}5(1-0{,}5)}{0.05^2}}{1 + (\frac{1{,}96^2 * 0{,}5(1-0{,}5)}{0{,}05^2 * 12})}$$

Se determinó un tamaño de muestra de 12 individuos, la cual es adecuada para lograr un nivel de confianza del 95 % y un margen de error del 5 %. Este cálculo garantiza que los resultados del estudio sean estadísticamente significativos y representativos de la población en su totalidad.

3.5 Fuentes de información

A continuación, en la Tabla 6. Fuentes primarias se presentan las fuentes de información primaria relevantes para elaborar el presente estudio. Las fuentes primarias se definen como aquellas que proporcionan datos de primera mano, es decir, información extraída directamente de quienes la produjeron 30.

Las fuentes presentadas en este apartado, a excepción de los modelos de predicción, han sido recopiladas mediante una proceso de revisión documental, cuyo objetivo principal es conformar la base de datos para alimentar el modelo predictivo. Debido a la gran cantidad de estas fuentes, no se listan individualmente en este apartado, pero pueden consultarse en detalle en el Apéndice O.

Tabla 6. Fuentes primarias

Fuente	**Importancia**
Datos de producción y ventas históricos recolectados de compendios estadísticos.	Proporcionan datos objetivos necesarios para modelar y predecir la demanda futura.
Revisión documental de informes presentados por el Icafé	Debido a que la institución es el ente rector de la caficultura en Costa Rica, los informes presentados proporcionan una perspectiva objetiva de la situación actual.
Datos de mercado facilitados por Icafé y la Organización Internacional del Café (OIC).	Proporcionan información crucial sobre las tendencias del mercado cafetalero, tanto en el ámbito local como internacional, por ejemplo, precios o tipo de cambio.

Por otro lado, en la Tabla 7. Fuentes secundarias se presentan las principales fuentes secundarias, que son resúmenes o compilaciones de fuentes primarias o desarrollan un tema a partir de una recopilación propia de datos.

Tabla 7. Fuentes secundarias

Fuente	**Importancia**
Modelos de predicción multifactoriales y de series de tiempo	Los modelos de predicción multifactoriales y de series de tiempo son herramientas estadísticas y matemáticas diseñadas para analizar y prever patrones y tendencias en datos. Aunque los modelos de series de tiempo son útiles para identificar ciclos regulares en la producción y comercialización del café, se complementan con otros modelos predictivos que integren múltiples factores que puedan potenciar la predicción.
Café informado	Permite comprender la evolución de los costos de procesamiento, relevante para el análisis económico de la industria cafetalera.
Costos de beneficiado de café aceptados por ley	Aunque el estudio se centra en métodos de pronóstico de cosecha, también ofrece información valiosa sobre la predicción de la demanda del café.
Cambio climático Cerros de Dota	Proporciona información climática local que afecta la producción de café, fundamental para estudiar el impacto del clima en la calidad y cantidad del café.

Fuente	Importancia
Grupo Banco Mundial	Ofrece datos sobre el crecimiento económico global, lo cual impacta la demanda mundial de café y sus precios.
Informe sobre la actividad cafetalera de Costa Rica	Presenta análisis de producción nacional y oferta/demanda mundial, cruciales para evaluar la posición competitiva de Costa Rica en el mercado global.
Precios históricos del café en Nueva York	Registra precios internacionales, clave para entender las fluctuaciones en el mercado y su impacto en los ingresos de los productores.
Precio de liquidación final	Refleja los precios pagados a los productores al final de cada cosecha, esencial para el análisis de rentabilidad y sostenibilidad en el sector.

3.6 Sujetos de información

Los sujetos de estudio son personas o grupos que forman parte de una investigación y aportan una diversidad de características, puntos de vista y experiencias que son importantes para el proyecto investigativo. En este caso, los sujetos de investigación son:

Tabla 8. Sujetos de investigación

Rol del sujeto	Caracterización	Justificación de la importancia de este sujeto para la investigación
Microbeneficios de café	Pequeñas empresas procesadoras de café ubicadas en el cantón de Dota, que transforman las cerezas de café en grano verde para su comercialización.	Su participación es crucial porque son los actores directamente afectados por la incertidumbre en la cantidad de café a procesar. Además, aportan datos operativos y comerciales esenciales para construir y validar el modelo predictivo, permitiendo al proyecto abordar una problemática real y práctica.
Experto de minería de datos	Se refiere a un profesional con experiencia en el análisis de grandes volúmenes de datos, especializados en técnicas de minería de datos y modelado predictivo.	Su conocimiento es fundamental para el desarrollo y validación del modelo predictivo, ya que aportan criterios acerca de metodologías y herramientas que se pueden aplicar al proyecto.

3.7 Variables de la investigación

Enseguida, se presentan las variables de investigación que se relacionan con los objetivos específicos propuestos:

Tabla 9. Cuadro de variables

Nombre de la variable	**Tipo de variable**	**Definición conceptual**	**Indicador**	**Definición instrumento**
Objetivo específico n.º 1: determinar los factores cualitativos y cuantitativos necesarios para que se alimente el modelo predictivo, identificando también las fuentes de información pertinentes, mediante la recolección de datos a través de informes y revisión documental.				
Factores cualitativos y cuantitativos relevantes	Factores que describen características cualitativas (como condiciones específicas de mercado) y cuantitativas (como volúmenes de ventas y precios) relevantes para el análisis.		Lista de factores relevantes que se identifican.	Revisión documental. Entrevista con experto.
Factores relevantes disponibles	Factores con los que se tiene disponibilidad de información y que actuarán como insumo para el modelo.		Porcentaje de disponibilidad de factores relevantes.	Revisión documental
Objetivo específico n.º 2: preparar los datos mediante un proceso de limpieza y transformación para que se asegure su calidad, consistencia y adecuación para su uso en el modelo.				
Proceso de limpieza y transformación de datos.	Procedimientos aplicados a los datos para eliminar errores, inconsistencias y preparar los datos para el análisis.		Cantidad de reglas de limpieza y transformación realizadas.	Scripts de limpieza de datos
Calidad y consistencia de datos	Estado final de los datos después de ser limpiados y transformados, asegurando que estén libres de errores y sean uniformes para su uso efectivo en el modelo predictivo.		Indicadores de calidad de datos, como valores nulos o duplicados eliminados.	Hoja de recogida de datos.
4. **Objetivo específico n.º 3:** Construir el modelo predictivo, por medio de un proceso iterativo de evaluación de métricas, con el fin de proporcionar a los microbeneficios de café una herramienta que les permita predecir la demanda.				
Modelos predictivos	Herramientas con base en modelos matemáticos o estadísticos que permiten predecir la demanda de café y		Cantidad de modelos predictivos entrenados.	Hoja de recogida de datos.

Nombre de la variable	Tipo de variable	Definición conceptual	Indicador	Definición instrumento
	planificar la producción de manera eficiente.			
Capacidad de predicción de la demanda y planificación de la producción	La capacidad de los modelos para predecir la demanda futura del café y facilitar la planificación de la producción.		Cumplimiento de los objetivos de minería de datos.	Hoja de comprobación.

3.8 Instrumentos de investigación

En esta sección se detallan los instrumentos propuestos y se incluyen enlaces a los apéndices donde se encuentran las plantillas correspondientes.

Entrevista

La entrevista es una interacción estructurada entre dos personas que tiene un propósito definido. En ella, el entrevistado ofrece su perspectiva u opinión sobre un tema específico, mientras que el entrevistador se encarga de recopilar e interpretar esa visión particular. El objetivo principal consiste en obtener información a través del relato del entrevistado 32.

Existen tres tipos principales de entrevistas:

1. Entrevista abierta: el entrevistador permite que el entrevistado hable libremente, sin restricciones en sus respuestas. Las preguntas son generales, lo que fomenta una conversación más espontánea y detallada.
2. Entrevista cerrada: se estructura con preguntas específicas y respuestas limitadas. El entrevistado debe ceñirse a opciones predeterminadas, lo que permite un mejor control sobre la información obtenida.
3. Entrevista semiabierta: combina características de las entrevistas abiertas y cerradas. El entrevistador formula preguntas estructuradas, pero deja espacio para que el entrevistado amplíe su respuesta, logrando un balance entre flexibilidad y enfoque.

En el Apéndice F. se incluye la plantilla que se utiliza para la entrevista, la cual se llevará a cabo en un formato semiabierto y estará dirigida a una persona experta en minería de datos.

Revisión documental

La revisión documental consiste en recopilar y analizar diversas fuentes escritas, como informes, diarios y cartas, con el propósito de sustentar y contextualizar la investigación en curso. Es esencial verificar la autenticidad de estas fuentes para garantizar la validez del proyecto investigativo. Esta revisión permite comprender de forma más profunda el fenómeno en estudio 31.

En esta investigación se emplea este instrumento para explorar estudios previos que evidencian el impacto de diversos factores en la predicción de la demanda y la producción de café. La plantilla correspondiente se incluye en el Apéndice G Plantilla de revisión documental – Factores relevantes.

Posteriormente, mediante esta técnica se identifican las fuentes de información que contienen los factores necesarios para alimentar el modelo predictivo. La plantilla que se utiliza para la técnica se encuentra en el Apéndice H. Plantilla de revisión documental-fuentes de información

Hojas de recogida de datos

Las hojas de recogida de datos son formularios o documentos impresos diseñados para recopilar datos de manera sistemática. Estos documentos suelen presentar un formato tabular o en columnas, lo que facilita la anotación y organización de los datos 33. Este instrumento se utiliza para recopilar métricas sobre la calidad del conjunto de datos empleado, cuya estructura se presenta en el Apéndice J. Plantilla de hoja de recogida de datos para métricas de calidad de datos

Además, se utiliza para documentar las diferentes configuraciones de los parámetros de los modelos entrenados, según lo detallado en el Apéndice K. Plantilla de hoja de recogida de datos para parámetros de los modelos

Scripts de limpieza de datos

Son bloques de código que se utilizan para realizar transformaciones sistemáticas en un conjunto de datos, con el objetivo de corregir errores, eliminar inconsistencias, gestionar valores nulos y preparar los datos para su análisis posterior. En el Apéndice L se encuentra la plantilla del instrumento.

Hojas de comprobación

Una hoja de comprobación es un instrumento diseñado para recopilar y registrar datos de manera sistemática y ordenada. Su propósito principal es facilitar la interpretación y el análisis de los datos mediante una estructura sencilla y directa 33.

Las hojas de comprobación son útiles para identificar tendencias, patrones y comportamientos en los datos, lo que permite un análisis visual claro y efectivo. Estos están diseñados para responder a preguntas específicas, lo que las hace menos flexibles para aplicaciones generales o en otros contextos de recolección de datos 33.

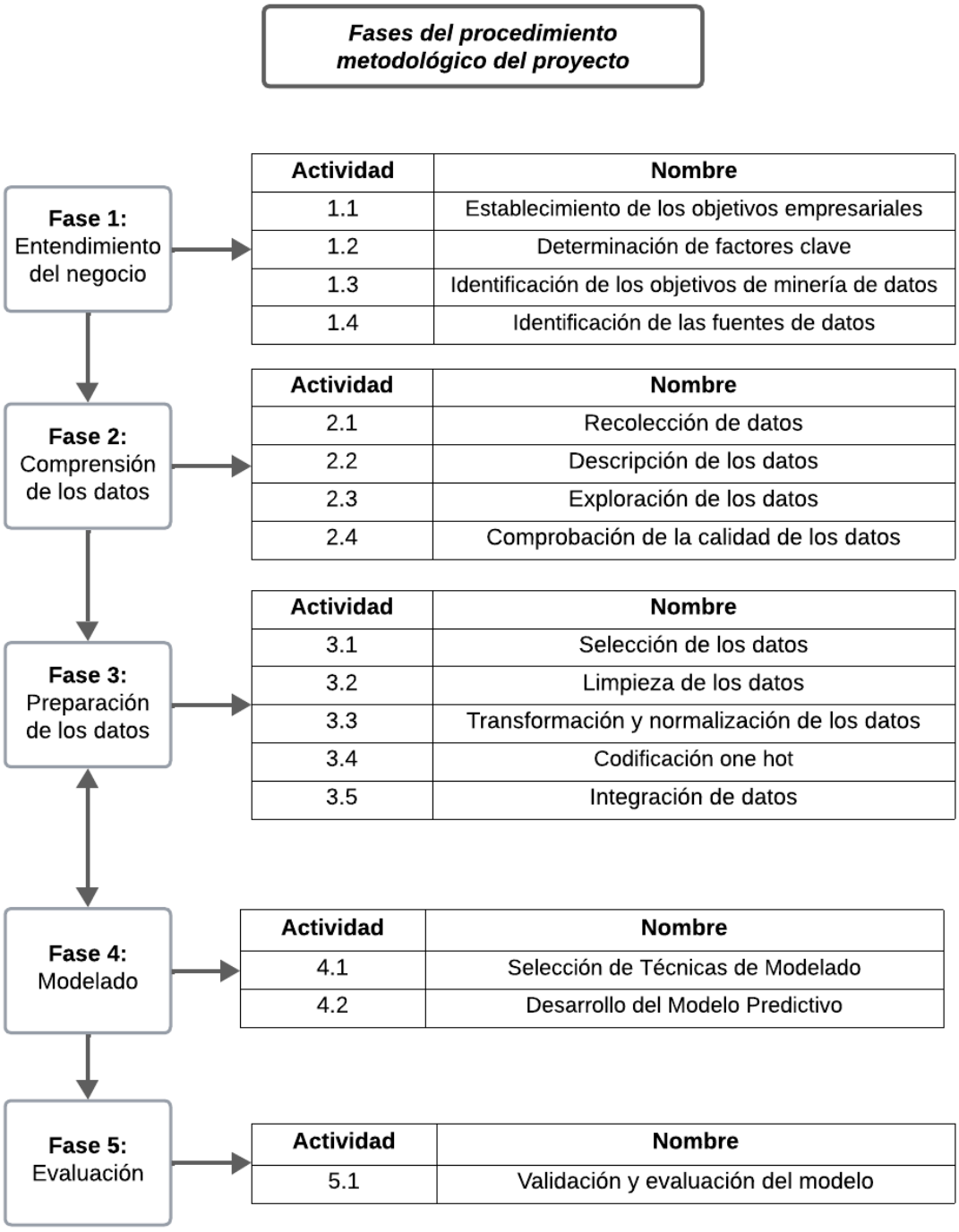
Para la presente investigación se utiliza esta herramienta, con la finalidad de contrastar los resultados de los diferentes modelos construidos. Este proceso ayuda a determinar cuáles modelos cumplen con los objetivos de minería de datos propuestos. La plantilla del instrumento se encuentra en el Apéndice M. Plantilla de hoja de comprobación

3.9 Procedimiento metodológico de la investigación

A continuación, se presenta el diagrama propuesto que ilustra las fases del procedimiento metodológico, acompañado de una explicación detallada de cada etapa. Es importante destacar que estas etapas están alineadas con las actividades estipuladas en el modelo estándar de minería de

datos CRISP-DM (Cross-Industry Standard Process for Data Mining), a excepción de las exclusiones mencionadas.

Ilustración 4. Procedimiento metodológico de la investigación



A continuación, se explica cada una de las fases:

3.9.1 Fase 1: entendimiento del negocio

En esta fase se establece una comprensión profunda de los objetivos empresariales en el contexto del microbeneficiado de café en Santa María de Dota. Se definen claramente los objetivos empresariales mediante reuniones y entrevistas con los beneficiadores de café y otros actores clave de la industria. Después, se determinan los factores clave que influyen en el modelo predictivo, tales como los factores de mercado y los datos históricos de producción. Además, se identifican los objetivos específicos de minería de datos que alinearán la solución técnica con los objetivos del negocio. Por último, se realiza una identificación exhaustiva de las fuentes de datos relevantes, asegurando que se cubran todos los aspectos necesarios para la modelización, incluidos los registros históricos y los datos sectoriales.

3.9.2 Fase 2: comprensión de los datos

En esta etapa se recopilan los datos necesarios desde las fuentes identificadas en la fase anterior, verificando su representatividad y adecuación al objetivo del proyecto. Se exploran los datos mediante herramientas de visualización y análisis para identificar patrones y relaciones, así como para detectar problemas de calidad, como duplicados y valores faltantes.

Además, se realizan análisis de correlación para identificar las variables más influyentes en la demanda y se generan estadísticas descriptivas que facilitan la comprensión de los datos. Este proceso asegura que los datos estén organizados y listos para la siguiente etapa

Fase 3: preparación de los datos

En esta etapa se limpian y transforman los datos para que sean aptos para el análisis predictivo. La limpieza incluye la eliminación de datos inconsistentes o incompletos. La transformación incluye la normalización de las variables, estandarizando las unidades de medida para facilitar su análisis.

También se codifican las variables categóricas en formatos compatibles mediante técnicas como la codificación "one hot". Finalmente, se integran los datos recopilados en un único repositorio que permite su uso eficiente en la etapa de modelado

3.9.3 Fase 4: modelado predictivo

En esta fase se seleccionan las técnicas de modelado que mejor se adaptan a los objetivos del negocio y las características de los datos, incluyendo algoritmos de regresión y modelos basados en series de tiempo. Se desarrollan y entrenan los modelos predictivos utilizando los datos preparados, y se ajustan los parámetros para mejorar su desempeño.

Además, se realizan pruebas iterativas con diferentes configuraciones de algoritmos, evaluando su desempeño y seleccionando el modelo que ofrece las predicciones más precisas y adecuadas a los objetivos planteados.

3.9.4 Fase 5: evaluación

Por último, se lleva a cabo una validación exhaustiva del modelo mediante técnicas como la validación cruzada, asegurando que el modelo sea robusto y fiable. Por último, se evalúan los resultados utilizando métricas estadísticas para garantizar que el modelo cumpla con los objetivos de minería de datos establecidos y finalmente se procede a seleccionar el mejor modelo.

3.10 Tabla resumen del procedimiento metodológico de la investigación

Esta sección ofrece un resumen de los métodos que se utilizan en el proyecto. La Tabla 10 detalla los objetivos del trabajo, la fase en la que se abordaron, las herramientas empleadas y la parte del proyecto en la que se desarrollaron, así como los principales hallazgos.

Tabla 10. Resumen del procedimiento metodológico de la investigación

Objetivo	Fase	Desarrollo	Apéndice/Anexo	Conclusiones
Determinar los factores cualitativos y cuantitativos necesarios para que se alimente el modelo predictivo, identificando también las fuentes de información pertinentes, mediante la recolección de datos a través de informes y revisión documental.	Fase 1: entendimiento del negocio	Entendimiento del negocio	Apéndice N Apéndice Ñ Apéndice O	Capítulo 6. Sección 6.2 Conclusiones
Preparar los datos mediante un proceso de limpieza y transformación para que se asegure su calidad, consistencia y adecuación para su uso en el modelo.	Fase 2: comprensión de los datos	Comprensión de los datos Preparación de los datos	Apéndice P Apéndice Q Apéndice R Apéndice S Apéndice T Apéndice U Apéndice W Apéndice V Apéndice X Apéndice Y Apéndice Z Apéndice AA	
	Fase 3: preparación de los datos			
Construir el modelo predictivo, por medio de un proceso iterativo de evaluación de métricas, con el fin de proporcionar a los microbeneficios de café una herramienta que les permita predecir la demanda.	Fase 4: modelado	Modelado predictivo Evaluación	Apéndice BB Apéndice CC Apéndice DD	
	Fase 5: evaluación			

Capítulo 4. Resultados

A continuación, se presenta el análisis de los resultados de la investigación, en el que se desarrollan las diferentes fases del proyecto, utilizando los instrumentos, las fuentes y los sujetos de estudio descritos anteriormente.

4.1 Entendimiento del negocio

En esta sección se lleva a cabo la fase inicial de la metodología, que se enfoca en comprender el negocio. Aquí se aclaran los objetivos empresariales del microbeneficiado de café en Santa María de Dota, alineando las metas del modelo predictivo con las necesidades del sector. Además, se definen los factores clave y se identifican las fuentes de datos pertinentes.

4.1.1 Establecimiento de los objetivos empresariales

Como ya se ha mencionado, el objetivo empresarial principal del modelo es mejorar la eficiencia operativa de los microbeneficios de café de Santa María de Dota al implementar un modelo predictivo que optimice planificar la producción y minimice las incertidumbres en la demanda del mercado.

4.1.2 Determinación de factores clave

Es importante aclarar que, para evitar confusiones entre las variables que se utilizan en el modelo predictivo y las empleadas en el resto de la investigación, a las variables del modelo no se les denominará *variables* en este documento. En su lugar, se les referirá como *factores relevantes*. De este modo, se busca mantener una distinción clara entre las variables de estudio que abarcan todo el trabajo final y los factores que se emplean específicamente en el modelo predictivo.

Identificación de factores relevantes:

En primer lugar, se llevó a cabo una revisión documental de investigaciones que analizan los factores que influyen en el mercado del café. Posteriormente, se buscó la validación de un experto en minería de datos para validar decisiones tomadas.

En el Apéndice N se pueden observar, mediante el instrumento de revisión documental, los factores que han sido estudiados previamente en diversos estudios. Se ha concluido, de una u otra forma, que estos factores tienen un impacto en la producción de café y en el mercado en general de este producto. Por lo tanto, se pretende extrapolar estos factores y adaptarlos para que sirvan como insumo del modelo predictivo.

- Factores locales

Primero, se tienen en cuenta los factores locales, es decir, aquellos factores en los que los microbeneficios tienen una inherencia directa y que, de alguna forma, dependen de la gestión interna y de las decisiones operativas de cada microbeneficio.

Entre los factores locales se encuentra la cantidad de café procesado en periodos anteriores. Este factor constituye el pilar de la investigación, ya que es la variable que se desea predecir. Seguidamente, se presentan los costos de beneficiado, los cuales abarcan todos los gastos involucrados en el procesamiento del café. Estos costos desempeñan un papel importante en la

toma de decisiones, ya que influyen directamente en la rentabilidad del negocio y en la capacidad de los beneficiadores para mantenerse competitivos en el mercado.

Además, se determinó que las variedades de café procesadas son un factor que incide en el precio y en la demanda del mercado, especialmente en el sector de cafés de especialidad, donde la calidad del café se determina, entre otros factores, por la variedad.

El último factor local identificado es la capacidad de beneficiado, que se refiere a la habilidad de las instalaciones para manipular y transformar las cerezas de café en granos de café verde. Este factor es determinante, ya que limita la cantidad de café que puede procesarse en un periodo determinado.

- Factores climáticos

La siguiente categoría de factores relevantes son los climáticos, los cuales son factores ambientales del entorno que, si bien es cierto, no pueden controlarse y responden al ámbito de producción y no al beneficiado de forma directa, estos afectan a la materia prima del beneficiado, que es el café cereza.

En esta categoría se encuentra la temperatura media, la cual se refiere al promedio de las temperaturas ambientales durante el ciclo de cultivo del café. Un aumento en la temperatura media puede reducir significativamente la producción, lo que subraya la necesidad de incorporar este factor ambiental en el modelo predictivo.

Las precipitaciones se refieren a la cantidad de lluvia que recibe la región donde se cultiva el café. Un nivel adecuado de lluvias resulta crucial para el desarrollo de las plantas de café, ya que influye en el total y calidad de la cosecha. Las variaciones extremas, ya sea por exceso o por falta de lluvias, pueden impactar negativamente la producción.

- Factores económicos

Por otro lado, se consideran los factores económicos, los cuales son externos y afectan los costos y la rentabilidad del café. Entre estos factores se contempla el tipo de cambio, que representa el valor de la moneda local en relación con las divisas extranjeras, mayoritariamente con respecto al dólar. Para los microbeneficios, las fluctuaciones en el tipo de cambio pueden afectar de manera significativa los ingresos, debido a que las ventas internacionales dependen de la cotización de la moneda en los mercados extranjeros.

Además, se incluye el crecimiento económico mundial, es decir, la expansión de la actividad económica global, como un factor que afecta la demanda de café; ya que a mayor índice de crecimiento mundial, mayor es la demanda.

- Factores del mercado del café

Por último, se consideran los factores del mercado del café, los cuales reflejan las dinámicas del mercado internacional, así como la oferta y la demanda. Primero, la producción nacional es la cantidad total de café que se produce en el país cada año. Este factor tiene un impacto directo en la oferta del producto nacional.

En contraparte, también se considera la oferta mundial como la cantidad de café disponible en el mercado internacional durante un periodo. Este factor afecta el equilibrio entre oferta y demanda y, por ende, los precios que los microbeneficios pueden obtener en los mercados globales.

Por otro lado, la demanda mundial se refiere a la cantidad de café consumido en el ámbito global. Este factor se relaciona con las tendencias de consumo, el crecimiento de la población y las preferencias de los consumidores en diversos mercados.

El último factor que se considera es el precio del café según la bolsa de valores de Nueva York. Este valor sirve como referencia para el café en el mercado de futuros y representa la especulación sobre el futuro del producto, así como el valor percibido internacionalmente en un momento dado.

A modo de recopilación de los factores que se identifican mediante la revisión documental, se presenta la siguiente tabla.

Tabla 11. Factores que se identifican mediante revisión documental

Factor
Factores locales
Cantidad de café beneficiado (factor por predecir)
Costos de beneficiado
Variedad del café
Capacidad de beneficiado
Factores climáticos
Temperatura media
Precipitaciones
Factores económicos
Tipo de cambio
Crecimiento económico mundial
Factores del mercado de café
Producción nacional
Oferta mundial
Demanda mundial
Precio según la bolsa de valores de Nueva York

Validación y observaciones del experto

Seguidamente, se aplicó una entrevista a una persona experta (ver Apéndice Ñ), para conocer su criterio acerca de los factores que se identifican. Entre las observaciones más relevantes, se destacan las siguientes.

Es válido utilizar registros comunes para todos los beneficios en un mismo año, como la producción nacional anual de café. Sin embargo, al hacerlo, existe el riesgo de que dicha columna no aporte un impacto significativo en el modelo. A pesar de esto, los valores de la producción nacional varían de 1 año a otro, lo cual puede resultar útil para identificar tendencias a través del tiempo y contribuir a la construcción del modelo.

Por otra parte, la persona experta identificó una limitación en los factores relevantes seleccionados: al contar con un único registro anual por cada beneficio. Debido a la naturaleza estacional de la actividad es probable que no existan suficientes datos o filas para entrenar adecuadamente cada modelo.

A continuación, se destacó la necesidad de realizar un análisis de correlación para cada uno de los factores clave. Este análisis permite identificar y descartar columnas que presentan una correlación muy fuerte, ya sea positiva o negativa, con otras variables, lo que ayuda a reducir la cantidad de variables en el modelo final.

El experto indicó que contar con 12 factores es un número óptimo para construir el modelo, ya que proporciona suficientes variables para respaldar las predicciones sin complicar los resultados con un exceso de columnas.

Finalmente, el experto no identificó factores adicionales relevantes, pero señaló la importancia de que todas las unidades de medida sean consistentes entre los registros. Algunas columnas pueden estar en quintales, mientras que otras pueden estar en fanegas, lo que evidencia las diferentes unidades que se utilizan en la caficultura. Por lo tanto, es necesario realizar un proceso de conversión o transformación de los datos para asegurar la uniformidad.

4.1.3 Identificación de los objetivos de minería de datos

Durante la entrevista con la persona experta (ver el Apéndice Ñ), se le consultó sobre las métricas de evaluación más adecuadas para medir la calidad de los modelos con base en series de tiempo. La experta destacó que, aunque es difícil establecer criterios de aceptación precisos, es válido utilizar aproximaciones o criterios que se ajusten a los resultados óptimos de cada métrica.

El propósito de definir estos objetivos en la minería de datos es, posteriormente, evaluar los modelos utilizando las métricas correspondientes para determinar su calidad y seleccionar los óptimos. A continuación, se procede a precisar y explicar cada objetivo.

- Objetivo de minería de datos: lograr un coeficiente de determinación del 80 %.

  Este objetivo establece que el modelo debe alcanzar un coeficiente de determinación (R^2) de al menos 80 %. Esto significa que el modelo debe ser capaz de explicar el 80 % de la variabilidad observada en los datos. Un R^2 de 80 % indica que el modelo posee un buen poder

predictivo, capturando una parte significativa de las relaciones subyacentes en los datos y dejando solo un 20 % de la variabilidad sin explicar.

- Objetivo de minería de datos: obtener un MAPE (error porcentual absoluto medio) máximo del 10 %

Este objetivo busca que el error promedio, expresado como un porcentaje de los valores reales, no supere el 10 %. Es decir, las predicciones pueden desviarse, en promedio, hasta un 10 % de los valores reales, independientemente de si se encuentran por encima o por debajo de estos.

- Objetivo de minería de datos 3: identificar al menos dos factores importantes del modelo.

Este objetivo se enfoca en identificar, mediante el análisis de los coeficientes del modelo, al menos dos factores que tienen un impacto significativo en las predicciones. Esto ayuda a entender cuáles variables son clave para el modelo y cómo influyen en los resultados.

4.1.4 Identificación de las fuentes de datos

A continuación, se presentan los hallazgos más relevantes de la revisión documental realizada para identificar las fuentes de datos correspondientes a cada factor clave (ver el Apéndice O). Estas fuentes permiten contextualizar los factores que se identifican previamente dentro del marco de este estudio. Los datos recopilados constituyen la base para construir el modelo predictivo.

- **Cantidad de café beneficiado (factor por predecir)**

Para este análisis, se consideran los informes anuales acumulados que reflejan la cantidad total de café procesado, expresada en fanegas, por cada empresa. Estos informes ofrecen una visión general del volumen procesado durante todo el año o la cosecha, lo que permite realizar comparaciones entre diferentes periodos.

- **Costos de beneficiado**

En este análisis se toma como referencia la estructura de costo elaborada a partir de las recomendaciones técnicas del Icafé. Este informe, actualizado anualmente, presenta el costo promedio nacional para el beneficiado de café y permite visualizar los costos por volumen procesado.

Además, se incluye información sobre los beneficios que procesan menos de 1.000 fanegas, que es el volumen de referencia para este estudio. Cabe destacar que, al tratarse de un costo promedio, se aplica el mismo valor para todos los beneficios en cada año evaluado.

- **Variedad del café**

No se encontró información relacionada con este factor relevante, por lo tanto, no se incluye en el modelo.

- **Capacidad de beneficiado**

No se encontró información relacionada con este factor relevante, por lo tanto, no se incluye en el modelo.

- **Temperatura media**

Se toman como referencia las estadísticas de Meteoblue, un servicio meteorológico desarrollado en la Universidad de Basilea, Suiza y basado en el reanálisis atmosférico ERA5, una fuente confiable y ampliamente utilizada.

Este servicio permite obtener la temperatura media anual en los Cerros de Dota, a una altitud de 1,894 m s. n. m. Los datos presentados en grados Celsius, no reflejan los microclimas locales, lo que puede generar discrepancias con las temperaturas reales presentes en las plantaciones de cada microbeneficio.

- **Precipitaciones**

Al igual que en la estadística anterior, el servicio de Meteoblue proporciona la variación anual total de las precipitaciones en la región de los Cerros de Dota, ubicada a 1,894 m s. n. m. En este caso, la unidad de medida es el milímetro, donde un milímetro de lluvia equivale a un litro de agua por m^2.

- **Tipo de cambio**

Se cuenta con la estadística sobre el tipo de cambio respecto al dólar estadounidense a partir del cual cada beneficio logró vender el café, de forma anual y en promedio de todas las ventas registradas. Contar con este dato es de especial relevancia, ya que permite disponer de datos diferenciados para cada beneficio, lo que puede aportar un gran valor al modelo.

- **Crecimiento económico mundial**

Para este análisis, se utiliza como referencia el indicador de crecimiento económico, lo que representa la tasa de crecimiento anual del producto interno bruto (PIB), de acuerdo con los datos proporcionados por el Banco Mundial.

- **Producción nacional**

En este caso, se toma como referencia la producción nacional de café verde reportada anualmente por el Icafé a través del informe anual, expresada en quintales, equivalentes a sacos de 46 kg.

- **Oferta mundial**

En este caso, se presenta el dato anual de la producción mundial de café, expresado en millones de sacos de 60 kg. La estadística que se utiliza se basa en diversas fuentes recopiladas por la Organización Internacional del Café (OIC) y se muestra a través del informe anual de Icafé.

- **Demanda mundial**

La demanda de café se calcula en función del consumo global del grano, impulsada principalmente por los países importadores. En este caso, se presenta el dato anual de la demanda mundial, expresado en millones de sacos de 60 kg. Esta estadística se fundamenta en diversas fuentes recopiladas por la Organización Internacional del Café (OIC) y se muestra en el informe anual del Icafé.

- **Precio según la bolsa de valores de Nueva York**

El precio del café se presenta como el precio promedio anual por quintal, expresado en dólares estadounidenses. Esta estadística lo recopila el Icafé, utilizando como referencia los datos provenientes de la bolsa de valores de Nueva York.

- **Precio de liquidación**

Aunque inicialmente no se contempló, se decidió incluir el precio de liquidación como un factor relevante. Este valor, que proviene de la misma fuente que el tipo de cambio, correspondiente a la liquidación final, no fue considerado en un principio debido a que no se identificó durante la revisión documental de potenciales factores relevantes. Sin embargo, en este punto se rescata su importancia, ya que refleja el precio al que cada microbeneficio logró comercializar su café, permitiendo realizar un análisis más completo y preciso del comportamiento de los microbeneficios en relación con el mercado.

4.2 Comprensión de los datos

A continuación, se inicia la segunda fase del procedimiento metodológico, que consiste en un análisis de los datos que se identificaron en la fase previa. En primer lugar, se lleva a cabo la recolección de datos a partir de las fuentes que se identificaron.

Posteriormente, estos datos se describen y exploran mediante técnicas de visualización y análisis, con el fin de identificar tendencias y características clave. Por último, se evalúa la calidad de los datos, localizando posibles registros incompletos o inconsistentes, los cuales requieren un tratamiento específico.

4.2.1 Recolección de datos

Enseguida, se procede a consolidar la información proveniente de las diversas fuentes de datos que se identificaron en un único archivo de Microsoft Excel, con el objetivo de centralizar y facilitar el acceso a la información. Esto se debe a que se cuenta con múltiples fuentes, este proceso también implica la integración de los datos, lo que resultará en un archivo de Microsoft Excel unificado.

El archivo de Microsoft Excel que se emplea para la carga de datos en R se encuentra en el Apéndice P. Es importante señalar que la primera fila del documento contiene los nombres de las columnas, los cuales corresponden a los factores relevantes considerados en el análisis.

La información sobre cada microbeneficio ha sido recolectada de manera anual, abarcando desde la cosecha 2016-2017 hasta la 2023-2024. Sin embargo, algunos datos sobre la cantidad de café beneficiado no están disponibles, debido a que ciertos microbeneficios no existían en los años más antiguos.

Además, es importante tener en cuenta que, debido a que el año 2024 no ha finalizado en el momento de este análisis, los datos de producción nacional, así como la oferta y la demanda mundiales correspondientes a la cosecha 2023-2024 son aproximaciones proporcionadas por las fuentes de datos empleadas.

Además, se creó una columna para el *microbeneficio* y otra para el *año*, con el fin de registrar la cantidad de café beneficiado por cada microbeneficio en cada periodo anual. Estas columnas son esenciales para organizar los datos y asegurar una correcta identificación de los registros a través del tiempo. Posteriormente, se procede a cargar los datos en la plataforma R Studio, utilizando el lenguaje de programación R para llevar a cabo la exploración y el análisis de estos.

4.2.2 Descripción de los datos

En primer lugar, el conjunto de datos consta de 96 filas, que representan los registros individuales y 13 columnas, que corresponden a los distintos factores o características asociadas a cada registro (ver el Apéndice Q). En la siguiente tabla se describen las unidades de medida de los factores relevantes. Este proceso es necesario, ya que asegura que cada variable se interprete y utilice correctamente de acuerdo con su contexto.

Las unidades de medida permiten comprender, tanto la magnitud como la naturaleza de los datos, lo que facilita su análisis posterior y la conversión a unidades uniformes cuando sea necesario, garantizando la coherencia en las comparaciones y cálculos entre diferentes variables. Además, en la tabla se indica el tipo de dato al que pertenece cada factor, cuya información detallada se puede visualizar en el Apéndice Q.

Tabla 12. Unidades de medida y tipos de datos

Factor relevante	Unidad de medida	Tipo de dato
Microbeneficio	No aplica	Character
Año	No aplica	Numérico
Cantidad de café beneficiado	Fanegas	Numérico
Costo de beneficiado	Colones por quintal	Numérico
Temperatura media	Grados Celsius	Numérico
Precipitación	Milímetros	Numérico
Tipo de cambio	Colones con respecto al dólar	Numérico
Precio liquidación	Colones por fanega	Numérico
Crecimiento económico	Porcentaje	Numérico
Producción nacional	Miles de fanegas	Numérico
Oferta mundial	Millones de sacos de 60 kg	Numérico
Demanda mundial	Millones de sacos de 60 kg	Numérico
Precio según la bolsa de NY	Dólares por quintal	Numérico

4.2.3 Exploración de los datos

La exploración de datos abarca técnicas de minería de datos que pueden abordarse mediante consultas o visualizaciones. Enseguida, se presentan estadísticas descriptivas del conjunto de datos, junto con visualizaciones de los atributos clave y un análisis de correlación.

Estadísticas básicas

Seguidamente, en el Apéndice R, se destacan algunas estadísticas básicas presentes en el conjunto de datos, las cuales se interpretan a través de la Tabla 13.

Tabla 13. Estadísticas básicas de los datos

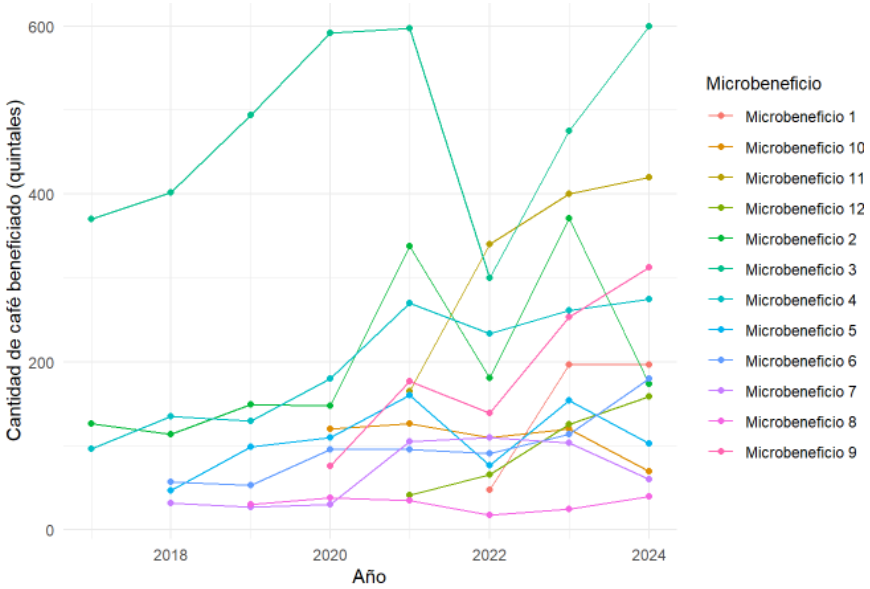
Factor relevante	Descripción
Año	Los datos abarcan de 2017 a 2024.
Cantidad de café beneficiado	Varía de 18 a 600 fanegas, con una media de 177.31 fanegas.
Costo de beneficiado	Oscila entre ₡26.569 y ₡37.488 por quintal, con una media de ₡32.833.
Temperatura media	Rango de 17.60 a 18.30 °C, lo que indica un clima constante que puede influir en la calidad del café producido.
Precipitación	Varía entre 2905 mm y 3924 mm, con una media de 3409 mm, estos valores sugieren una variabilidad marcada en la cantidad de lluvia cada año.
Tipo de cambio	Oscila entre ₡531.5 y ₡684.3, con una media de ₡593.3, indicando variabilidad que afecta costos y rentabilidad.
Precio liquidación	Varía de ₡75,444 a ₡397.656, con una media de ₡173.142, sugiriendo variaciones significativas en los precios de venta.
Crecimiento económico	Rango de -2.9 % a 6.3 %, con una media de 2.657 %, indica un crecimiento moderado con periodos de recesión (pandemia de la COVID-19) que pueden impactar el mercado.
Producción nacional	Oscila entre 1,673 y 2,018 miles de fanegas, con una media de 1,867 miles, esto muestra una producción nacional constante.
Oferta mundial	Varía de 160.6 a 176 200 000 de sacos de 60 kg. La media de 168 500 000 sugiere una oferta estable, pero cercana a la demanda, lo que puede generar fluctuaciones en los precios.
Demanda mundial	Oscila entre 159.5 y 169 600 000 de sacos. La media de 165 900 000 sugiere una demanda relativamente alta y sostenida, lo que puede llevar a una competencia intensa en el mercado.
Precio según la bolsa de NY	Varía entre $101.2 USD y $214.7 USD por quintal. La media de $141.6 USD indica fluctuaciones considerables, influenciadas por la oferta y demanda global, así como factores externos, como la economía y la especulación.

Visualizaciones

A continuación, se presentan las visualizaciones clave del análisis:

- Comportamiento de la variable por predecir

Ilustración 5. Cantidad de café beneficiado por año por cada microbeneficio



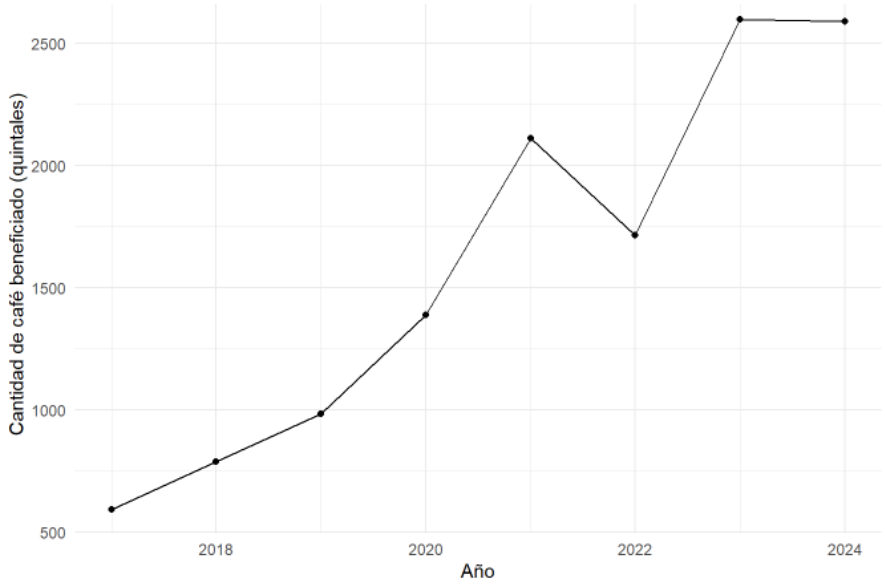
El gráfico muestra cómo varía la cantidad de café beneficiado para distintos microbeneficios a través de los años, revelando diferencias significativas entre ellos. Algunos microbeneficios procesan volúmenes de café de manera constante, mientras que otros presentan fluctuaciones más marcadas. Esta variabilidad sugiere que ciertos microbeneficios poseen una mayor capacidad o demanda sostenida.

Por otro lado, el análisis del gráfico revela que, inicialmente, todos los microbeneficios mostraron una tendencia de crecimiento en la cantidad de café beneficiado. Sin embargo, en 2022, la mayoría experimentó una notable disminución en su producción, lo que sugiere la existencia de un evento o factor que afectó de manera generalizada el volumen procesado en ese año.

A pesar de esta caída en 2022, la mayoría de los microbeneficios logró recuperarse en los años posteriores, alcanzando sus máximos históricos en 2023 o 2024. Este repunte puede estar relacionado con mejores condiciones de producción o con un aumento en la demanda de café. Así, aunque enfrentaron 1 año difícil, los microbeneficios demostraron capacidad de recuperación, logrando niveles de producción nunca vistos en los últimos 2 años del análisis.

- Cantidad total de café en el tiempo

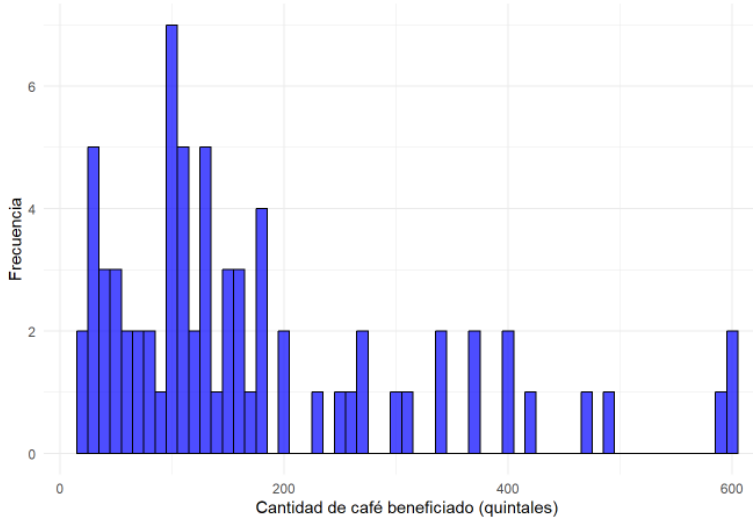
Ilustración 6. Cantidad de café total en el tiempo



La Ilustración 6 denota la evolución de la cantidad total de café beneficiado en quintales a través del tiempo. En 2017, la producción inicia con un valor aproximado de 600 quintales, presentando un aumento constante hasta 2021, cuando se superan los 2000 quintales. Sin embargo, en 2022, se registra una caída notable, descendiendo a alrededor de 1600 quintales. A partir de 2023, la producción se recupera rápidamente y alcanza más de 2500 quintales y se mantiene estable en ese nivel durante 2024.

- Distribución de los datos

Ilustración 7. Distribución de la cantidad de café beneficiado



El histograma ilustra la distribución de la cantidad de café beneficiado en quintales. Se puede observar que la frecuencia de registros es notablemente mayor en los rangos de 100 a 200 quintales, lo que indica que estas cantidades son las más comunes dentro del conjunto de datos.

La distribución se presenta como asimétrica, lo que indica que no sigue un comportamiento normal y muestra un sesgo hacia la derecha. Esto sugiere la existencia de algunos microbeneficios que procesan cantidades significativamente mayores de café, aunque estos casos son menos comunes. Además, se pueden identificar valores atípicos en los extremos del histograma, en especial en las cantidades que superan los 500 quintales, lo que señala la presencia de casos inusuales en la producción de café.

Análisis de correlación

En este apartado se realiza un análisis de correlación entre los factores relevantes del estudio. Este análisis permite identificar si existen relaciones significativas entre las variables, así como la fuerza y dirección de dichas correlaciones. Al comprender cómo se relacionan los factores, se pueden descartar aquellas variables que presenten una correlación importante entre sí, lo que evita redundancias y simplifica el modelo predictivo. Esto contribuye a una mayor eficiencia en el análisis y asegura que solo las variables más importantes sean incluidas en el modelo final. El *script* del examen de correlación se presenta en el Apéndice S.

Para interpretar los resultados del análisis de correlación mostrado en el mapa de calor del Apéndice T, se emplea la siguiente escala:

- Los colores rojos indican correlaciones positivas fuertes, mientras que los colores azules reflejan correlaciones negativas significativas.
- Cuanto más cercano al blanco, más débil resulta la correlación entre las variables.
- Valores cercanos a 1 representan una relación positiva fuerte entre dos factores, es decir, cuando un factor aumenta, el otro también tiende a aumentar.
- Valores cercanos a -1 sugieren una relación negativa fuerte, lo que implica que cuando un factor aumenta, el otro tiende a disminuir.

Una vez aclarado el proceso de interpretación de los resultados del análisis de correlación, se procede a interpretar los resultados más relevantes a través de la siguiente tabla. Es decir, únicamente se interpretan aquellos resultados que implican una relación fuerte, ya sea positiva o negativa (±0.7).

Tabla 14. Interpretación del análisis de correlación

Correlación	Factor 1	Factor 2	Valor	Interpretación
Positiva fuerte	Año	Costo de beneficiado	0.9	Hay una correlación muy fuerte entre el año y el costo de beneficiado, lo que sugiere que a medida que pasa el tiempo, el costo de beneficiado tiende a aumentar.
	Año	Demanda mundial	0.93	La demanda aumentó con el tiempo. Esto puede estar relacionado con tendencias de

Correlación	Factor 1	Factor 2	Valor	Interpretación
				consumo, crecimiento poblacional o cambios en los hábitos de consumo.
	Demanda mundial	Costo de beneficiado	0.97	Una correlación extremadamente alta, lo que sugiere que a medida que la demanda mundial de café aumenta, el costo de beneficiado también tiende a incrementarse.
	Costo de beneficiado	Precio según la bolsa de valores de NY	0.9	Existe una correlación muy fuerte entre el costo de beneficiado y el precio en la bolsa de Nueva York, indicando que a medida que el costo de beneficiado aumenta, también lo hace el precio del café en el mercado.
	Demanda mundial	Precio según la bolsa de valores de NY	0.78	Una fuerte correlación positiva indica que cuando la demanda mundial de café aumenta, el precio en la bolsa de Nueva York tiende a aumentar, reflejando la dinámica del mercado.
Negativa fuerte	Precipitación	Oferta mundial	-0.87	La fuerte correlación negativa sugiere que a medida que la precipitación aumenta, la oferta mundial de café tiende a disminuir. Sin embargo, esta interpretación debe ajustarse, ya que los datos de precipitación corresponden específicamente a la región de Dota, no al contexto global.
	Precipitación	Crecimiento económico mundial	-0.76	Esta correlación negativa sugiere que a medida que la precipitación aumenta, el crecimiento económico mundial disminuye. Sin embargo, al igual que en el caso anterior, esta conclusión no es válida, ya que los datos de precipitación son específicos de Dota.
	Temperatura	Tipo de cambio	-0.93	Esta correlación extremadamente fuerte sugiere que a medida que la temperatura aumenta, el tipo de cambio tiende a disminuir. Sin embargo, esta conclusión también es incorrecta cuando se considera el contexto. Los datos de temperatura corresponden solo a la región de Dota y no deben utilizarse para interpretar movimientos en el tipo de cambio en el ámbito nacional.
	Temperatura	Producción nacional	0.77	Indica que a medida que la temperatura media aumenta, la producción nacional de café también tiende a aumentar, posiblemente

Correlación	Factor 1	Factor 2	Valor	Interpretación
				debido a un clima más favorable para el cultivo.
	Tipo de cambio	Producción nacional	-0.73	Sugiere que una apreciación en el tipo de cambio puede estar asociada con una disminución en la producción nacional, posiblemente debido a la dificultad en la exportación de café.
	Precio según la bolsa de valores de NY	Producción nacional	-0.89	Una correlación negativa fuerte indica que a medida que el precio en la bolsa de Nueva York aumenta, la producción nacional tiende a disminuir.
	Costo de beneficiado	Producción nacional	-0. 7	Indica que a medida que el costo de beneficiado aumenta, la producción nacional tiende a disminuir, lo que sugiere que mayores costos pueden limitar la capacidad de producción.

4.2.4 Comprobación de la calidad de los datos

En el conjunto de datos se identifican 180 valores nulos distribuidos a lo largo de diferentes columnas, lo que indica que existen campos incompletos que deben tratarse antes del análisis. No se detectaron registros duplicados, lo que sugiere que los datos han sido depurados de manera adecuada.

Además, los datos son consistentes, ya que no se encontraron discrepancias lógicas o contradicciones, debido a que su recolección proviene de fuentes oficiales. Aunque existe una discrepancia en las unidades de medida entre diferentes columnas, cada columna mantiene uniformidad en sus valores, es decir, todos los datos dentro de una misma columna están expresados en la misma unidad de medida, lo que facilita su análisis posterior.

Lo expresado se recolectó a través del *script* en R del Apéndice U y se consolidó mediante el instrumento de hoja de recogida de datos del Apéndice W.

4.3 Preparación de los datos

En este apartado se aborda la preparación de datos mediante una serie de pasos que incluyen la selección de variables, la limpieza, la transformación y la normalización, asegurando la consistencia y calidad del conjunto de datos. Además, se consideran técnicas para reducir la multicolinealidad y realizar una correcta codificación de factores categóricos.

4.3.1 Seleccionar los datos

Siguiendo los resultados del análisis de correlación realizado en la fase anterior, se descartarán aquellos factores que presenten alta multicolinealidad, es decir, aquellos con un valor de correlación superior a ±0,9. Esta decisión es importante, ya que la multicolinealidad, además de no aportar valor, puede distorsionar los resultados del modelo.

Se espera que las variables independientes actúen de manera independiente, sin embargo, si su correlación es alta, se dificulta el aislamiento de la relación entre cada variable independiente y la variable dependiente. Esto puede llevar a una interpretación errónea de los coeficientes y a que disminuya la precisión de las predicciones. Al considerar lo anterior, en la siguiente tabla se rescatan aquellos factores con una alta correlación.

Tabla 15. Selección de factores con correlación

Factor 1	Factor 2	Valor de correlación	Factor para conservar	Justificación
Año	Costo de beneficiado	0.9	Año	El año se conserva como variable temporal relevante, ya que aporta información sobre la tendencia temporal.
Año	Demanda mundial	0.93	Año	El año se conserva por su rol temporal, brindando contexto histórico que es crucial para el análisis.
Demanda mundial	Costo de beneficiado	0.97	Ninguno	Ambos factores se eliminaron, ya que no aportan suficiente valor independiente en el análisis.
Costo de beneficiado	Precio según la bolsa de valores de NY	0.9	Precio según la bolsa de valores de NY	El costo de beneficiado se descartó en la elección anterior.

Finalmente, se conservarán el resto de los factores relevantes. El *script* para eliminar el costo de beneficiado y la demanda mundial se muestra en el Apéndice V.

4.3.2 Limpieza de los datos

En primer lugar, se define una estrategia para gestionar los valores nulos presentes en el conjunto de datos, considerando la columna de origen de cada uno. Se mencionan los valores nulos porque, de todas las métricas de calidad evaluadas en la sección anterior, este fue el único problema identificado.

Tabla 16. Estrategia de limpieza de datos

Columna	Valores nulos	Estrategia	Justificación
Cantidad de café beneficiado	24	Eliminar las filas	Se decidió eliminar las filas con valores nulos en esta columna porque estos registros probablemente corresponden a periodos en los que el microbeneficio no existía.
Costo de beneficiado	48	No aplica	Este factor se excluyó del conjunto de datos.
Temperatura media	12	Imputación con la media	La media es una estimación neutral y adecuada que no introduce sesgos en los datos climáticos
Precipitación	12	Imputación con la media	Al igual que en la temperatura media, la imputación con la media es apropiado para mantener la neutralidad en los datos.
Tipo de cambio	36	Imputación con la media	De los 36 valores faltantes, 12 corresponden al año 2024, mientras que los otros 24 se eliminaron junto con las filas que tenían valores nulos en la columna de cantidad *de café beneficiado*. Se utilizó la media histórica del tipo de cambio para imputar los valores faltantes del año en curso.
Precio de liquidación	36	Imputación con la media	Los 36 valores faltantes, incluidos 12 del año 2024, fueron tratados de manera similar al tipo de cambio, usando la media histórica para el año en curso.
Crecimiento económico mundial	12	Imputación con valores predictivos	Los valores faltantes corresponden al año 2024, para lo cual se utiliza la proyección del Banco Mundial (2.6 %) 34.

En total, se identificaron 180 valores nulos en las diversas columnas. Para llevar a cabo la limpieza de los datos, se utilizó el *script* presentado en el Apéndice X, el cual detalla el código empleado para este proceso.

4.3.3 Transformación y normalización de los datos

Enseguida, se realiza la estandarización de unidades y la normalización de los datos para asegurar la consistencia y la comparabilidad entre los factores relevantes en el análisis.

Transformación de los datos

En el proceso de transformación de los datos, se estandarizan las unidades de medida de los datos numéricos para garantizar su coherencia. Esto implica utilizar las mismas unidades para factores comparables, como en el caso de las cantidades de café.

Tabla 17. Estandarización de unidades de medida de los factores relevantes

Factor relevante	Unidad de medida original	Unidad de medida de destino
Cantidad de café beneficiado	Fanegas	Fanegas
Temperatura media	Grados Celsius	Grados Celsius
Precipitación	Milímetros	Milímetros
Tipo de cambio	Colones con respecto al dólar	Colones con respecto al dólar
Precio liquidación	Colones por fanega	Colones por fanega
Crecimiento económico	Porcentaje	Porcentaje
Producción nacional	Miles de fanegas	Fanegas
Oferta mundial	Millones de sacos de 60 kg	Fanegas
Demanda mundial	Millones de sacos de 60 kg	Fanegas
Precio según la bolsa de NY	Dólares por quintal	Dólares por fanega

Para esta actividad es necesario remitirse a la Microbeneficiado del café del presente documento, donde se explica la terminología propia del proceso de beneficiado del café, lo que incluye las unidades de medida empleadas y sus respectivas conversiones. En consecuencia, se consideran las siguientes equivalencias:

Tabla 18. Equivalencias de unidades de medida en el proceso de beneficiado de café

Unidad de medida	Equivale a
Una fanega	Un quintal
Una fanega	46 kg de café oro

Al considerar lo anterior, se procede a transformar únicamente los datos que se relacionan con la cantidad de café, ya que son los únicos que se refieren a una misma magnitud, pero con diferentes unidades de medida. Se toma la unidad de medida fanega como base para la transformación, debido a que las fanegas son la unidad del factor por predecir.

La producción nacional se multiplicará por 1,000 para convertirla de miles de fanegas a fanegas. En cuanto a la demanda y oferta mundial, se transformarán de millones de sacos de 60 kg a fanegas utilizando las siguientes equivalencias: 1 saco de 60 kg de café oro equivale a 1.30 fanegas (60/46). Por lo tanto, la ecuación para la conversión es:

$$Demanda\ u\ oferta\ en\ fanegas = X \times 10^6 \times 1.30$$

Donde X es la demanda u oferta en millones de sacos de 60 kg.

Una vez realizado este análisis, se ejecuta el *script* en R que permite transformar los datos, el cual se presenta en el Apéndice Y.

Normalización de los datos

Mediante este proceso, se busca transformar los valores de los factores numéricos para que se ubiquen en un rango común. Tradicionalmente, se utiliza la técnica de normalización min-max para escalar los datos entre 0 y 1, lo cual mantiene las proporciones originales de los valores dentro de este rango definido. Sin embargo, en este caso, se opta por normalizar entre 1 y 10 para evitar posibles problemas de división entre cero en el cálculo de las métricas. La fórmula ajustada es la siguiente:

$$X_{normalizado} = 1 + \frac{X - X_{min}}{X_{max} - X_{min}} * 9$$

Donde:

- X es el valor original del factor.
- X_{min} es el valor mínimo del factor.
- X_{mac} es el valor máximo del factor.

El proceso de normalización min-max implica lo siguiente: para cada valor numérico, se resta el valor mínimo del conjunto de datos y luego se divide por la diferencia entre el valor máximo y el valor mínimo, escalando en un rango de 1 a 10. Esto asegura que se mantengan las relaciones entre los datos, ajustando la escala. El *script* en R que implementa la normalización de los datos se detalla en el Apéndice Z.

La normalización es crucial, ya que se trabaja con variables que tienen diferentes escalas o unidades de medida. Si no se realiza este ajuste, las variables con valores más grandes pueden dominar el comportamiento de los modelos predictivos, lo que impactaría negativamente en su rendimiento. Al normalizar entre 1 y 10, se garantiza que las diferencias de magnitud entre los datos no influyan de manera desproporcionada en el modelo, lo que permite que todas las variables tengan una escala comparable.

4.3.4 Codificación one hot

El uso de la codificación *one hot* para el factor *microbeneficio* se llevó a cabo con el objetivo de generar predicciones específicas para cada tipo de microbeneficio. Esto se debe a que los algoritmos de predicción no pueden procesar directamente variables categóricas, por lo que es necesario transformar este factor en un formato que permita al modelo interpretar de manera correcta la información asociada a cada categoría.

La codificación *one hot* consiste en convertir una variable categórica en varias variables binarias, donde cada columna representa una categoría y toma el valor de 1 cuando un registro pertenece a esa categoría y 0 cuando no. En este caso, el factor *microbeneficio* incluía distintos tipos de microbeneficios, por lo que se generó una columna binaria para cada uno, lo que le permitió al modelo analizar el efecto de cada uno de forma independiente. Seguidamente, se eliminó una de las columnas binarias que se generan para evitar la colinealidad, es decir, una relación lineal entre los factores independientes y se eliminó la columna original. El detalle de este *script* se puede encontrar en el Apéndice AA.

4.3.5 Integración de los datos

Esta actividad ya se realizó en la fase 1: entendimiento del negocio.

4.4 Modelado predictivo

En esta sección se presentan los modelos predictivos que se utilizan para pronosticar la cantidad de café beneficiado, con base en técnicas que se seleccionaron previamente. Además, se describen las técnicas elegidas, justificando por qué algunas han sido descartadas y se evalúan los resultados a través de tres iteraciones.

4.4.1 Selección de técnicas de modelado

En la Métodos de minería se exploraron diversas técnicas de modelado predictivo, algunas de ellas con base en el análisis de series temporales y otras en modelos multivariables. Estas técnicas se seleccionaron para modelar y predecir la cantidad de café procesado a través del tiempo. Esto se debe a que cada microbeneficio opera en condiciones distintas y maneja volúmenes de producción diferentes, por lo que es fundamental capturar estas variaciones para obtener predicciones precisas y útiles para la planificación.

A continuación, se justifican brevemente los modelos seleccionados y los que se descartaron.

Modelos descartados

Los modelos de media móvil, media móvil ponderada, suavizamiento exponencial y Arima han sido descartados en este contexto porque se basan exclusivamente en una serie temporal univariada, es decir, solo utilizan los datos históricos de una sola variable para realizar las predicciones. Si bien estos modelos pueden resultar útiles para ciertos tipos de predicciones, no son adecuados en este caso, ya que no permiten incorporar factores auxiliares importantes, como la diferenciación para cada microbeneficio y otros factores relevantes que se identificaron.

Por otro lado, si bien es posible intentar crear un modelo individual para cada microbeneficio, la falta de datos suficientes para cada uno hace que esta opción tampoco sea viable. Lo anterior limita la posibilidad de utilizar estos modelos de forma puntual y los descarta como herramientas efectivas en este estudio.

Modelos que se conservan

- Arimax: este modelo es una extensión de Arima que permite incorporar variables exógenas, es decir, factores externos al sistema que pueden influir en la producción. Esto lo convierte en una opción válida, ya que ofrece la flexibilidad necesaria para capturar la influencia de variables externas importantes, lo que permite ajustar las predicciones de acuerdo con las diferencias observadas entre los microbeneficios.
- Sarimax: este modelo extiende Arimax al incorporar un componente de estacionalidad. Debido a la naturaleza estacional de ciertos procesos productivos, como los ciclos agrícolas, Sarimax permite modelar y ajustar los patrones estacionales.
- Redes neuronales: estos modelos son capaces de capturar relaciones complejas no lineales entre múltiples variables. Esto se debe a que el sistema de producción agrícola involucra varios factores interdependientes, las redes neuronales se utilizan en virtud

de su capacidad para aprender patrones complejos a partir de los datos y realizar predicciones más precisas.

- Random forest: este modelo es un método de ensamble que combina múltiples árboles de decisión, lo que permite manejar grandes cantidades de variables y sus interacciones sin requerir una normalización previa de los datos. Su robustez ante el sobreajuste lo convierte en una alternativa confiable para predecir la demanda en los microbeneficios.
- KNN: este modelo clasifica o predice valores con base en la proximidad de un punto a sus *vecinos* en el espacio de características. En contextos de producción, KNN puede resultar útil para identificar patrones similares en datos históricos y aplicar esta información para realizar predicciones.
- Gradient *boosting*: este es un método de aprendizaje que construye modelos secuenciales, donde cada nuevo modelo intenta corregir los errores del modelo anterior.

4.4.2 Desarrollo del modelo predictivo

A continuación, se desarrollan los distintos modelos seleccionados en la actividad previa. Se realiza un total de tres iteraciones, en las cuales se prueban diferentes configuraciones para optimizar los modelos y mejorar su precisión. Asimismo, es importante destacar algunas consideraciones clave que se tienen en cuenta para construir estos modelos.

En todas las iteraciones se trabaja con un conjunto de datos que incluye variables *dummy*. Estas variables se generan para convertir las categorías de variables cualitativas (como *microbeneficio*) en valores numéricos, lo cual es necesario para que los modelos puedan procesar y analizar correctamente dicha información. Aunque las variables *dummy* son fundamentales en el proceso de modelización, no se consideran relevantes para evaluar la importancia en cada modelo, ya que representan meras codificaciones de las categorías presentes en el conjunto de datos.

Para evaluar el rendimiento de los modelos que se implementan, se utilizan dos métricas principales: R^2 (coeficiente de determinación) y MAPE (error porcentual absoluto medio), que son las métricas establecidas en los objetivos de minería de datos. Estas métricas permiten medir el grado de ajuste de los modelos a los datos y la precisión de sus predicciones.

- El R^2 es una métrica que indica la proporción de la variabilidad en la variable dependiente (en este caso, la cantidad de café beneficiado) que el modelo es capaz de explicar. Cuanto más alto sea el valor de R^2, mejor ajustado está el modelo a los datos. Un valor de $R^2 = 1$ dicta un ajuste perfecto, mientras que un valor cercano a 0 sugiere que el modelo no explica la variabilidad de los datos. Para calcular el R^2, se compara la suma de los errores al cuadrado (la diferencia entre los valores reales y predichos) con la variabilidad total en los datos.
- El MAPE se calcula tomando la diferencia absoluta entre los valores reales y los valores predichos por el modelo, dividiendo este valor por los valores reales y luego promediando esos errores. Finalmente, se multiplica por 100 para obtener el error en forma de porcentaje.

En cuanto a la importancia de las variables, se sigue un enfoque diferenciado según el tipo de modelo:

- Para los modelos de redes neuronales, *random forest*, *gradient boosting* y *k-nearest neighbors* (KNN), se calcula la importancia de las variables automáticamente utilizando la función varImp() del paquete *caret*. Esta función evalúa la relevancia de las variables para el rendimiento del modelo.
- En el caso del modelo KNN, aunque no se calcula directamente la importancia de las variables debido a la naturaleza del algoritmo, que no maneja particiones ni coeficientes, se incluye este modelo en el análisis junto con sus respectivas métricas de rendimiento.
- Para los modelos con base en series temporales con variables exógenas, como Arimax y Sarimax, la importancia de las variables se determina a partir de los coeficientes asociados a dichas variables exógenas. En estos casos, se extraen los coeficientes del modelo ajustado y se identifican como más relevantes las variables con los coeficientes de mayor magnitud absoluta.

Por otro lado, los recursos para construir cada modelo se detallan a continuación:

Redes neuronales

El modelo utiliza la librería *caret*, la cual es muy popular en R para entrenar diversos modelos de *machine learning*. La función clave es train(), que permite configurar y entrenar el modelo utilizando redes neuronales artificiales, en este caso a través del método *nnet*.

Random forest

También se construye utilizando la librería *caret*, con el método *rf*. Random forest es un algoritmo de aprendizaje automático basado en la construcción de múltiples árboles de decisión para mejorar la precisión y evitar el sobreajuste. La función train() facilita nuevamente el proceso de entrenamiento, por lo que automatiza la selección de hiperparámetros y permite un control del modelo.

Gradient boosting

Este modelo también utiliza *caret* con el método *gbm*, que implementa *gradient boosting*. Este algoritmo entrena múltiples árboles de decisión de manera secuencial, corrigiendo los errores del modelo anterior en cada iteración. La elección de *caret* se debe a su capacidad para manejar este tipo de modelos con facilidad, ajustar automáticamente los parámetros y permitir la evaluación rápida del desempeño.

K-Nearest Neighbors (KNN)

El modelo de *k-nearest neighbors* (KNN) utiliza la misma librería *caret*, mediante el método *knn*. Este enfoque se basa en encontrar las observaciones más cercanas en términos de distancia en el espacio de características. Se elige este paquete porque permite experimentar de manera eficiente con diferentes configuraciones de KNN.

Arimax

Para el modelo Arimax, se utiliza la librería *forecast*, que incluye la función auto.arima(), la cual selecciona automáticamente los mejores parámetros para el modelo Arima con variables exógenas (xreg). En este contexto, se modela una serie temporal y *forecast* se presenta como una librería adecuada, ya que está optimizada para manejar modelos de series temporales y automatiza el proceso de ajuste de parámetros.

Sarimax

Al igual que Arimax, este modelo utiliza *forecast* para ajustar un modelo Sarimax. La diferencia principal es que dicho modelo incluye un componente estacional, lo cual lo gestiona el argumento seasonal = TRUE en la función auto.arima(). La librería *forecast* se elige porque maneja de manera eficiente los modelos que incluyen, tanto componentes estacionales como variables exógenas, haciendo que sea más sencillo ajustar el modelo Sarimax.

Es importante mencionar que a todos los modelos que utilizan el paquete *caret* se les aplicó una validación cruzada de 10 pliegues. Esto significa que los datos se dividen en 10 subconjuntos o *pliegues*, de manera que en cada iteración se utiliza uno de estos subconjuntos como conjunto de validación y los otros nueve como conjunto de entrenamiento. Este proceso se repite 10 veces, cada vez con un subconjunto diferente como conjunto de validación.

Con estas consideraciones, se procede a ejecutar la primera iteración de los modelos. El *script* de esta iteración y de las posteriores se puede visualizar en el documento de R consolidado (Apéndice BB). Además, los resultados de las iteraciones se presentan en el Apéndice CC.

Iteración 1

Redes neuronales (MAPE: 18.24 %, R^2: 88.57 %)

Este modelo muestra un buen ajuste (R^2 alto) y un MAPE relativamente bajo. Esto sugiere que las redes neuronales han capturado de manera adecuada la relación no lineal entre las variables y han realizado un buen trabajo al predecir la cantidad de café por beneficiar. Las variables más importantes en dicho modelo fueron la oferta mundial y el año.

El modelo se configuró de la siguiente manera.

- linout = TRUE: este parámetro se utiliza en redes neuronales para establecer que la salida es un valor continuo, es decir, se realiza una regresión. Si se estuviera llevando a cabo una clasificación, este valor sería FALSE.
- size = 1: indica que se utiliza una única neurona en la capa oculta. Este es el valor por defecto si no se especifica otro.

Random forest (MAPE: 13.96 %, R^2: 93.64 %)

Random forest ha tenido el mejor rendimiento en esta iteración en términos de precisión predictiva, evidenciado por el MAPE más bajo y el R^2 más alto. Esto sugiere que este modelo ha logrado captar mejor las relaciones entre las variables, siendo el tipo de cambio y el año las que han generado un mayor impacto.

El modelo se configuró de la siguiente manera:

- ntree = 500: esto indica que el modelo utiliza 500 árboles, que es el valor por defecto en *random forest* si no se especifica otro.

Gradient boosting (MAPE: 54.84 %, R^2: 19.95 %)

Este modelo no ha funcionado adecuadamente, presentando un MAPE muy alto y un R^2 muy bajo, lo que sugiere que apenas ha capturado las relaciones entre las variables. Las variables más relevantes, el precio de liquidación y el tipo de cambio, parecen no haber sido modeladas de manera adecuada.

Los parámetros del modelo se configuraron de la siguiente manera:

- n.trees = 50: el modelo se ha entrenado con 50 árboles. Este valor especifica el número total de árboles que se generaron en el proceso de *boosting*. Utilizar un número bajo de árboles puede dar lugar a un modelo subajustado, como parece ser el caso en esta situación.
- interaction.depth = 1: este parámetro controla la profundidad máxima de cada árbol. En este caso, los árboles son muy poco profundos (solo 1 nivel), lo que significa que cada uno es extremadamente simple y puede no ser capaz de capturar interacciones complejas entre las variables. Esto puede explicar, en parte, el bajo rendimiento del modelo.
- La tasa de aprendizaje (*shrinkage*) fue establecida en 0.1. Este valor controla el impacto de cada nuevo árbol en el modelo final. Una tasa de aprendizaje baja hace que el modelo aprenda más lentamente, lo que puede ayudar a evitar el sobreajuste, sin embargo, también puede haber limitado la capacidad del modelo para mejorar su ajuste en este caso.

K-nearest neighbors (KNN) (MAPE: 71.56 %, R^2: 4.03 %)

Este modelo ha mostrado el peor desempeño, con un MAPE muy alto y un R^2 muy bajo, lo que indica que no ha sido capaz de capturar correctamente las relaciones en los datos. Esto puede ser un indicio de que KNN no es adecuado para este tipo de problema, debido a que este algoritmo asume que los datos cercanos en el espacio de características son similares. Sin embargo, en este caso, las relaciones entre las variables son complejas o no lineales, lo cual puede ocasionar problemas para capturarlas de manera efectiva, lo que limita la capacidad predictiva.

El modelo se configuró de la siguiente manera:

- k = 23: este parámetro indica que el modelo utiliza 23 vecinos más cercanos para realizar las predicciones. Este valor se determinó por el propio modelo. En este caso parece no ser adecuado, porque no existen suficientes observaciones para cada microbeneficio.

Arimax y Sarimax (MAPE: 21.39 %, R^2: 87.10 %)

Estos dos modelos, Arimax y Sarimax, se analizan de manera conjunta, ya que ambos arrojaron los mismos resultados. Al incorporar el componente de estacionalidad en Sarimax, no se logró un impacto significativo, lo que indica que la estacionalidad no desempeña un papel relevante en este conjunto de datos. Esto sugiere que el conjunto de datos no presenta un patrón estacional claro, motivo por el cual los resultados de Sarimax son prácticamente idénticos a los de Arimax.

A pesar de esta falta de estacionalidad, los resultados son relativamente buenos, con un MAPE bajo y un coeficiente de determinación (R^2) alto, lo que sugiere que el modelo captura de manera adecuada las relaciones subyacentes en los datos, aunque la estacionalidad no sea un factor destacado. Por otra parte, las variables exógenas con mayor impacto en el modelo son el año y la oferta mundial.

Los parámetros de los modelos se configuraron de la siguiente manera:

- Seasonal = FALSE para Arimax: esto indica que no se utiliza un componente estacional en este modelo.
- Seasonal = TRUE para Sarimax: aunque Sarimax incluye un componente estacional, este no logró mejorar el rendimiento, lo que reafirma la falta de estacionalidad en los datos.

Por otro lado, al utilizar la función auto.arima, los valores de p, d y q se seleccionan automáticamente:

- p (autorregresivo): el número de términos de valores previos (autorregresivos) que se utilizan.
- d (diferenciación): el número de diferencias necesarias para que la serie sea estacionaria.
- q (promedio móvil): el número de términos del promedio móvil, con base en los errores previos.

En este caso, el modelo Arimax seleccionó una configuración Arima (0,0,0), lo que indica que no se encontró una estructura autorregresiva ni de promedio móvil relevante en los datos. Esto significa que el modelo no aplica términos de autorregresión, diferenciación o promedio móvil, sino que se basa únicamente en las variables exógenas para realizar las predicciones.

Iteración 2

Para este análisis se decidió descartar el modelo Sarimax, ya que incorporar un componente estacional no demostró agregar un valor significativo al rendimiento del modelo. También se descartó el modelo Arimax, debido a que al utilizar la función auto.arima, los valores de p, d y q se seleccionan automáticamente y no se encontraron patrones relevantes para la configuración.

En vista de lo anterior, se realiza una segunda iteración con los modelos restantes.

Redes neuronales (MAPE: 5.69 %, R^2: 99.14 %)

En la segunda iteración, el modelo de redes neuronales mejoró de forma notable, con un MAPE significativamente más bajo y un R^2 casi perfecto, lo que indica un ajuste excelente y una alta precisión en las predicciones. Esto sugiere que las redes neuronales capturaron con mayor exactitud las relaciones complejas en los datos. Las variables más importantes en esta iteración fueron el crecimiento económico mundial y el precio según la Bolsa de Nueva York.

Configuración del modelo:

- linout = TRUE: indica que la salida es un valor continuo, es decir, se trata de un problema de regresión.
- size = 5: en esta iteración se aumentó el número de neuronas en la capa oculta a 5, lo cual parece haber mejorado la capacidad del modelo para captar patrones complejos y contribuyó con el aumento en la precisión.

Random forest (MAPE: 13.99 %, R^2: 93.71 %)

El modelo de *random forest* mantuvo un rendimiento similar al de la iteración anterior, con un MAPE y R^2 muy cercanos, lo que demuestra que es una alternativa estable y confiable. Las variables más relevantes en esta iteración también fueron el tipo de cambio y el año, reafirmando su importancia en el modelo.

Configuración del modelo:

- ntree = 100: en esta iteración, el número de árboles se redujo de 500 a 100. A pesar de la disminución en el número de árboles, el rendimiento no se vio afectado significativamente, lo que sugiere que 100 árboles son suficientes para capturar las relaciones en los datos.

Gradient boosting (MAPE: 51.72 %, R^2: 33.42 %)

El rendimiento del modelo de *gradient boosting* mejoró ligeramente en comparación con la primera iteración, pero continúa mostrando un MAPE alto y un R^2 bajo, lo que indica que aún no captura adecuadamente las relaciones en los datos. Las variables más relevantes son el precio de liquidación y el tipo de cambio.

Configuración del modelo:

- n.trees = 200: se incrementó el número de árboles a 200, lo cual permite que el modelo tenga una mayor capacidad para aprender patrones complejos.
- interaction.depth = 3: la profundidad de los árboles se incrementó a 3, lo que permite al modelo captar interacciones más complejas entre las variables.
- *shrinkage* = 0.05: la tasa de aprendizaje se redujo a 0.05, lo que permite que el modelo aprenda de manera controlada.

K-nearest neighbors (KNN) (MAPE: 46.51 %, R^2: 46.59 %)

En esta iteración, el modelo KNN mejoró en comparación con la primera, presentando un MAPE y un R^2 más altos, sin embargo, aún no logra un rendimiento satisfactorio. Este resultado sugiere que el ajuste del parámetro k tiene un impacto significativo en el desempeño del modelo, aunque KNN es menos efectivo que otros enfoques.

Configuración del modelo:

- k = 2: en esta iteración, se ajustó el modelo para utilizar únicamente los 2 vecinos más cercanos. Esta reducción parece haber mejorado el rendimiento.

Iteración 3

En este análisis se intenta mejorar el rendimiento de los modelos con tasas de precisión moderadas y, al mismo tiempo, reducir el sobreajuste en aquellos modelos con tasas de exactitud excepcionalmente altas. El sobreajuste ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento, capturando, tanto patrones relevantes como detalles específicos (ruido), lo que afecta su capacidad de generalización en datos nuevos.

Para controlar el sobreajuste, se implementó una validación cruzada de 10 pliegues, la cual permite evaluar el rendimiento del modelo en distintas particiones de los datos. Esto ayuda a verificar si el modelo mantiene un buen rendimiento en múltiples subconjuntos y no solo en un conjunto fijo, lo que indica que el modelo ha aprendido patrones generalizables. Si el modelo obtiene buenos resultados en todos los pliegues de validación, es menos probable que esté sobreajustando.

A pesar de esta medida, se ajustarán los parámetros en los modelos de redes neuronales y *random forest* para reforzar la prevención del sobreajuste. Adicionalmente, se busca optimizar los modelos de *gradient boosting* y KNN con el objetivo de mejorar su rendimiento general.

Redes neuronales (MAPE: 12.39 %, R^2: 95.31 %)

En la tercera iteración, el rendimiento del modelo de redes neuronales disminuyó en la precisión respecto a la iteración anterior. Aunque el MAPE aumentó y el R^2 se redujo levemente, sigue mostrando una buena capacidad predictiva, lo que sugiere que los ajustes realizados ayudaron a controlar el sobreajuste. Las variables más relevantes fueron la producción nacional y las precipitaciones.

Configuración del modelo:

- linout = TRUE: indica que la salida es continua y adecuada para regresión.
- size = 3: se disminuyó el número de neuronas en la capa oculta, lo que puede haber simplificado el modelo.
- decay = 0.2: se incrementó la regularización, lo que penaliza los pesos altos y ayuda a mejorar la generalización.

Random forest (MAPE: 14.29 %, R^2: 93.93 %)

En esta iteración, *random forest* mostró un leve cambio en el rendimiento, con una pequeña reducción en el MAPE y un R^2 similar al de la iteración anterior, lo que indica que el modelo es robusto. La reducción en el número de árboles parece no haber afectado significativamente la capacidad predictiva. Las variables más relevantes fueron el tipo de cambio y la oferta mundial.

Configuración del modelo:

- ntree = 50: el número de árboles se redujo a 50, lo que simplificó el modelo y redujo el riesgo de sobreajuste.

Gradient boosting (MAPE: 57.83 %, R^2: 20.50 %)

El rendimiento del modelo de *gradient boosting* disminuyó en esta iteración, con un MAPE más alto y un R^2 más bajo, lo que indica que el ajuste de los hiperparámetros no ocasionó mejoras significativas en la capacidad predictiva. Es posible que el modelo aún no capture adecuadamente las relaciones en los datos.

Configuración del modelo:

- n.trees = 300: se aumentó el número de árboles a 300, lo que permite una mayor capacidad para capturar patrones complejos.
- interaction.depth = 3: mantener la profundidad de los árboles permite captar interacciones complejas.
- *shrinkage* = 0.01: se redujo la tasa de aprendizaje para lograr un ajuste más controlado, aunque el modelo aún no logra generalizar adecuadamente.

K-nearest neighbors (KNN) (MAPE: 59.86 %, R^2: 12.49 %)

En esta iteración el modelo de KNN disminuyó en el rendimiento comparado con la iteración anterior. El incremento en el número de vecinos (k) parece haber resultado en una reducción de la precisión, lo que sugiere que KNN es menos efectivo en comparación con otros modelos.

Configuración del modelo:

- k = 8: se ajustó el modelo para utilizar los 8 vecinos más cercanos, lo que ocasionó un rendimiento inferior.

Para finalizar, se presenta la siguiente tabla como resumen de los resultados obtenidos por cada modelo en las distintas iteraciones.

Tabla 19. Resultados de los modelos evaluados

Modelo	Iteración	MAPE	R2
Redes neuronales	1	18,24%	88,57%
Random Forest	1	13,96&	93,64%
Gradient Boosting	1	54,84%	19,95%
K-Nearest Neighbors (KNN)	1	71,56%	4,03%

Modelo	Iteración	MAPE	R2
ARIMAX	1	21,39%	87,10%
SARIMAX	1	21,39%	87,10%
Redes neuronales	2	5,69%	99,14%
Random Forest	2	13,99	93,71%
Gradient Boosting	2	51,72%	33,42%
K-Nearest Neighbors (KNN)	2	46,51%	46,59%
Redes neuronales	3	12,39%	95,31%
Random Forest	3	14,29%	93,93%
Gradient Boosting	3	57.83%	20,50%
K-Nearest Neighbors (KNN)	3	59,86%	12,49%

4.5 Evaluación

En esta fase se evalúa la efectividad de los modelos predictivos que se generan en la fase de desarrollo. Los modelos se comparan según los objetivos de minería de datos establecidos. Finalmente, se selecciona el modelo que mejor se adapte a los requisitos del proyecto.

4.5.1 Validación y evaluación del modelo

La validación cruzada, inicialmente planeada para esta etapa, se realizó en la fase de desarrollo del modelo predictivo. Esta acción se llevó a cabo para asegurar la robustez de los modelos entrenados, por lo que no es necesario implementarla de nuevo en esta etapa. Además, los modelos ya se evaluaron mediante las métricas obtenidas y el ajuste de los parámetros en cada iteración se llevó a cabo con la intención de mejorar los resultados previos. A continuación, se procede a comparar los diversos modelos con los objetivos de minería de datos establecidos, para evaluar el cumplimiento de cada uno y seleccionar el modelo más adecuado. Los resultados de dicho análisis se presentan en el Apéndice DD, el cual corresponde al instrumento de hoja de comprobación.

Objetivo n.º 1: lograr un coeficiente de determinación del 80 %

- Cumplen**:** redes neuronales (iteraciones 1, 2 y 3), *random forest* (iteraciones 1, 2 y 3), Arimax (iteración 1), Sarimax (iteración 1).
- No cumplen**:** *gradient boosting* y *k-nearest neighbors* (todas las iteraciones).

Objetivo n.º 2: obtener un MAPE (error porcentual absoluto medio) máximo del 10 %

- Cumplen**:** redes neuronales (iteración 2).
- No cumplen: todos los demás modelos en todas las iteraciones, ya que solo las redes neuronales, en su segunda iteración, logran cumplir con el umbral de MAPE del 10 %.

Objetivo n.º 3: identificar al menos dos factores importantes del modelo

- Cumplen**:** redes neuronales (todas las iteraciones), *random forest* (todas las iteraciones), *gradient boosting* (todas las iteraciones), Arimax y Sarimax (iteración 1).
- No cumple: *k-nearest neighbors* en todas las iteraciones, ya que no identifica los factores de relevancia.

Selección del mejor modelo

Aunque el modelo de redes neuronales en la segunda iteración cumple con todos los objetivos, se decidió seleccionar el modelo de la tercera iteración de redes neuronales. Esto se debe a que en esta iteración se tomaron medidas adicionales para evitar el sobreajuste, lo que mejora su generalización y robustez, a pesar de que el MAPE es ligeramente superior al objetivo. Este modelo es el más adecuado para proceder con la implementación.

4.5.2 Próximos pasos

Tal como se indicó en las exclusiones de este proyecto, la fase de despliegue de CRISP-DM queda fuera del alcance. Esto significa que no se contempla el traslado del modelo a un entorno operativo. Sin embargo, se considera que el modelo de redes neuronales seleccionado puede presentar resultados que satisfagan los objetivos del negocio, aunque es importante seguir entrenando el modelo con datos actualizados, debido a que la base de datos actual es pequeña, en virtud de la naturaleza anual de los datos.

Capítulo 5. Limitaciones y problemas que se encontraron

Durante el desarrollo de este proyecto se identificaron diversas limitaciones y problemas que influyeron en los procedimientos, resultados y alcance del modelo predictivo propuesto para la planificación de la demanda y producción de café en los microbeneficios de Santa María de Dota. Enseguida, se detallan las dificultades encontradas organizadas en función de los aspectos metodológicos y operativos que impactaron en el estudio.

Inicialmente, el proyecto contemplaba crear un *software* para predecir la producción. Sin embargo, en la fase de anteproyecto, se determinó que un enfoque basado en un modelo predictivo es más adecuado. Aunque este cambio de dirección no representó un obstáculo significativo, debido a que se realizó en una etapa temprana requirió un ajuste en los objetivos y métodos de investigación, adaptando el enfoque a una metodología basada en CRISP-DM.

La participación de múltiples actores en el sector cafetalero, como propietarios de microbeneficios, instituciones reguladoras y clientes finales, representa una barrera para la comunicación directa y frecuente, lo que complica las instancias de validación y retroalimentación. Esta multiplicidad de partes interesadas genera desafíos para lograr un consenso o una visión unificada, lo que puede afectar la alineación de los objetivos del modelo con las expectativas y necesidades de todos los involucrados.

Por otra parte, aunque el modelo predictivo se desarrolló con los factores disponibles y considerados críticos, es posible que algunos factores relevantes no se hayan identificado debido a la complejidad del sector cafetalero y sus fluctuaciones. Además, la creación de un modelo predictivo eficiente en la práctica depende de un volumen considerable de datos históricos para su entrenamiento adecuado. Sin embargo, debido a la naturaleza anual de los datos y a la implementación relativamente reciente de microbeneficios en la región, la cantidad de registros disponibles resultó ser limitada.

Por último, el proceso iterativo característico de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) representó un reto considerable debido a las limitaciones de tiempo en el desarrollo del proyecto. La aplicación de este enfoque iterativo requería ajustes continuos en el modelo y en el conjunto de datos, aunque este último no se modificó una vez que se consolidó.

Capítulo 6. Discusión y conclusiones

6.1 Discusión

En este apartado, se analizan los resultados obtenidos a partir de los modelos predictivos aplicados en este proyecto, así como las oportunidades de mejora que podrían fortalecer tanto los modelos como sus conclusiones.

En primer lugar, es importante señalar que los modelos fueron entrenados modificando únicamente sus parámetros y no el conjunto de datos utilizado. Esto significa que las conclusiones extraídas podrían cambiar considerablemente si se incorporaran nuevos datos o si se modificara el enfoque en la selección de variables. Por ejemplo, capturar información adicional sobre procesos específicos de tratamiento del café, como las variedades empleadas, las prácticas agrícolas o aspectos detallados de trazabilidad, podría haber enriquecido el análisis. Este tipo de variables no solo mejorarían el poder predictivo del modelo, sino que también contribuirían a una mejor comprensión de la trazabilidad y la gestión operativa de los microbeneficios.

Asimismo, los criterios de aceptación de las métricas empleadas pueden no ser aplicables de manera universal. Cada empresa u organización opera en un contexto único, lo que podría requerir ajustar las métricas de evaluación a las necesidades específicas de cada caso. Sin embargo, en este estudio se procuró utilizar métricas que fueran aplicables a una variedad de contextos y que, además, permitieran realizar comparaciones en términos relativos. Por ejemplo, se empleó el MAPE (Error Absoluto Medio Porcentual), que mide la tasa de error en términos relativos, a diferencia de otras métricas como el MAE (Error Absoluto Medio), que se expresan en valores absolutos y pueden ser menos útiles para ciertos escenarios comparativos.

Adicionalmente, podrían incorporarse métricas complementarias que permitan analizar aspectos alternativos del desempeño del modelo, lo cual contribuiría a obtener una perspectiva más amplia sobre sus fortalezas y limitaciones.

Dicho esto, los resultados técnicos obtenidos ofrecen valiosas perspectivas sobre el desempeño de los modelos evaluados. A diferencia de estudios previos mencionados en la revisión de trabajos similares, que emplearon series de tiempo univariables, este proyecto requirió un análisis más complejo debido a la naturaleza multifactorial de los datos.

El uso de redes neuronales en este estudio ocasionó una capacidad predictiva notable. Aunque en la segunda iteración de este modelo se obtuvo un desempeño satisfactorio y, de hecho, fue el único en cumplir en su totalidad con los criterios de minería de datos, se optó por una tercera iteración que incorporó medidas adicionales para reducir el riesgo de sobreajuste. No obstante, el sobreajuste es una posibilidad debido a lo complejo del modelo y la naturaleza limitada de los datos disponibles.

Cada modelo aplicado mostró un comportamiento característico que refleja sus fortalezas y limitaciones en el contexto de la predicción del café. Por ejemplo, los modelos con base en series de tiempo tradicionales, como Arima, no se consideraron, finalmente, debido a la dificultad para capturar variaciones impulsadas por factores externos.

Por otro lado, el modelo de *random forest* también mostró un rendimiento robusto y constante a lo largo de las iteraciones, con un desempeño ligeramente inferior al de las redes neuronales, pero igual de confiable.

En contraste, los modelos de *gradient boosting* y *k-nearest neighbors* (KNN) presentaron un rendimiento considerablemente inferior. Gradient boosting mostró dificultades para capturar patrones complejos en los datos, lo que puede estar relacionado con una configuración de parámetros que no optimizó el modelo de manera efectiva, a pesar de que se aumentó el número de árboles y se redujo la tasa de aprendizaje.

KNN, por otra parte, fue el modelo con el peor desempeño debido a su limitación para capturar relaciones no lineales entre las variables. Lo anterior sugiere que este método no es adecuado para el contexto multifactorial de este proyecto.

Por último, los modelos Arimax y Sarimax, al basarse en series de tiempo y ajustados con variables exógenas, mostraron un desempeño intermedio. Ambos modelos arrojaron resultados similares, lo que indica que la estacionalidad no es un factor relevante en este conjunto de datos. Aunque capturaron adecuadamente las relaciones de las variables exógenas, como el año y la oferta mundial, su rendimiento fue limitado en comparación con los modelos más complejos, como las redes neuronales.

Desde una perspectiva teórica, este proyecto plantea la relevancia de los modelos multifactoriales en la predicción de la demanda agrícola. En el ámbito práctico, el modelo seleccionado, aunque aún se encuentra en una fase preliminar, es potencialmente útil para los microbeneficios del café, ya que permite prever la demanda con mayor precisión y planificar la producción de manera informada. Si bien el modelo actual proporciona una herramienta básica de predicción, puede servir como base para un desarrollo más sofisticado, integrando incluso datos adicionales en el futuro para mejorar su exactitud.

Por último, al abordar la problemática central de esta investigación, relacionada con "la incertidumbre en la cantidad de café por procesar que enfrentan los microbeneficios del cantón de Dota", es fundamental reconocer que esta situación trasciende el ámbito técnico y requiere un enfoque más integral. Se necesitan iniciativas institucionales que impulsen la digitalización del sector, fomenten la adopción de tecnologías disruptivas y promuevan la toma de decisiones basadas en datos, antecedentes y tendencias del mercado.

Además, dado el carácter multifacético del sector cafetalero, con múltiples actores y encadenamientos involucrados, este esfuerzo debe ser colectivo y coordinado. Solo a través de esta colaboración será posible anticipar las variaciones en el mercado y garantizar que los microbeneficios puedan mantenerse competitivos en un entorno cada vez más globalizado y complejo.

6.2 Conclusiones

A continuación, se presentan las conclusiones del estudio, organizadas de acuerdo con cada objetivo.

Con respecto al primer objetivo específico: "Determinar los factores cualitativos y cuantitativos necesarios para alimentar el modelo predictivo, identificando también las fuentes de información pertinentes, mediante la recolección de datos a través de informes y revisión documental", se concluye:

- Se identificaron 12 factores relevantes a partir de investigaciones previas, los cuales se consideran útiles como insumo para el entrenamiento de cada modelo, tal como se menciona en la Determinación de factores.
- Se recopilaron 11 factores relevantes, seleccionados al extrapolar los factores que se identifican y ajustarlos a la disponibilidad específica del contexto del proyecto, tal como se detalla en la Identificación de las fuentes de datos.

El objetivo específico n.º 2 es: "Preparar los datos mediante un proceso de limpieza y transformación para que se asegure su calidad, consistencia y adecuación para su uso en el modelo", del cual se proponen las siguientes conclusiones:

- Se identificaron 180 registros con problemas de calidad en el conjunto de datos, los cuales correspondían completamente a registros nulos, como se puede observar en la Comprobación de la calidad
- Se descartaron los factores relevantes de costo de beneficiado y demanda mundial del conjunto de datos, al encontrarse una fuerte colinealidad con el factor año, como se denota en la sección correspondiente a Seleccionar los datos.
- Se llevó a cabo un proceso de limpieza para los 180 registros con problemas de calidad, empleando técnicas de imputación y eliminación de registros. Esto se aprecia en la Limpieza

Por último, el objetivo específico n.º 3 establece: "Construir el modelo predictivo, por medio de un proceso iterativo de evaluación de métricas, con el fin de proporcionar a los microbeneficios de café una herramienta que les permita predecir la demanda". Las conclusiones son las siguientes:

- Se seleccionaron seis técnicas de modelado que se consideraron las más funcionales en relación con el conjunto de datos disponibles y los objetivos del proyecto, tal como se describe en la Selección de técnicas
- El modelo de redes neuronales, en su tercera iteración, alcanzó un MAPE de 12.39 % y un R^2 de 95.31 %, posicionándose como el modelo más preciso para la predicción de demanda en los microbeneficios de café, tal como se denota en el apartado Desarrollo del modelo
- Aunque el modelo seleccionado no alcanzó un MAPE inferior al 10 %, sí cumple con un R^2 superior al 80 %. Por lo tanto, se considera una herramienta potencialmente viable para predecir la demanda de café y permitir una planificación de producción informada, apoyada en datos validados y métodos de evaluación cruzada, como se observa en la Validación .

Capítulo 7. Cumplimiento de objetivos

A continuación, se presenta un cuadro comparativo que detalla los objetivos establecidos en el proyecto, junto con los objetivos alcanzados durante su ejecución, lo que incluye el porcentaje de cumplimiento correspondiente para cada uno.

Tabla 20. Cumplimiento de objetivos

Objetivo general: Diseñar, durante el segundo semestre de 2024, un modelo predictivo que apoye a los microbeneficios de café en Santa María de Dota para que se prevenga la demanda, lo que reduce la incertidumbre y optimiza sus operaciones para adaptarse a las fluctuaciones del mercado.			
Objetivo específico	**Productos**	**% de logro**	**Comentarios**
Determinar los factores cualitativos y cuantitativos necesarios para que se alimente el modelo predictivo, identificando también las fuentes de información pertinentes, mediante la recolección de datos a través de informes y revisión documental.	Sección 4.1.2 Determinación de factores Sección 4.1.4 Identificación de las fuentes de datos	100	N/A
Preparar los datos mediante un proceso de limpieza y transformación para asegurar su calidad, consistencia y adecuación para su uso en el modelo.	Sección 4.2 Comprensión de los datos Sección 4.3. Preparación de los datos	100	N/A
Construir el modelo predictivo, por medio de un proceso iterativo de evaluación de métricas, con el fin de proporcionar a los microbeneficios de café una herramienta que les permita predecir la demanda.	Sección 4.4Modelado predictivo Sección 4.5 Evaluación	100	N/A

Capítulo 8. Recomendaciones para futuras investigaciones

En este capítulo se presentan las recomendaciones para futuros investigadores interesados en profundizar en el tema o en realizar estudios similares. Estas sugerencias contemplan aspectos que no se abordaron en este proyecto, pero que pueden enriquecer investigaciones futuras sobre la predicción de la demanda y la planificación de la producción en microbeneficios de café.

Se recomienda incluir factores adicionales que puedan influir en la precisión del modelo predictivo, como condiciones climáticas específicas o fluctuaciones económicas, ya que estos pueden mejorar la exactitud de las predicciones. Además, a medida que surjan nuevos microbeneficios en la región es importante integrarlos en el modelo para obtener un panorama más amplio y actualizado de la producción cafetalera.

En caso de utilizar este modelo como base, es beneficioso agregar los datos más recientes para mejorar su capacidad predictiva y adecuarlo a las condiciones actuales del mercado. Asimismo, resulta pertinente explorar otros modelos predictivos, lo que incluye técnicas avanzadas como redes neuronales o modelos de aprendizaje profundo, para comparar su efectividad y, potencialmente, obtener resultados más precisos en determinados contextos.

Para completar el ciclo de la metodología CRISP-DM, se sugiere desplegar el modelo en un ambiente operativo. Sin embargo, es aconsejable esperar hasta contar con un volumen de datos más extenso, lo que permite asegurar la robustez del modelo en aplicaciones prácticas. Por último, debido a que los factores que afectan la producción y la demanda de café pueden variar con el tiempo, es importante realizar un seguimiento continuo de estos factores y actualizar los parámetros del modelo periódicamente para mantener su precisión y relevancia.

Capítulo 9. Referencias

1. Icafé. *Informe Actividad Cafetalera de Costa Rica 2023*. Heredia, 2023. https://www.icafe.cr/wp-content/uploads/informes_gestion/actividad_cafetalera/Informe%20Actividad%20Cafetalera%20de%20Costa%20Rica%202023.pdf
2. TEC. *Qué es el TEC*. 2024. https://www.tec.ac.cr/que-es-tec
3. TEC. *Administración de Tecnología de Información*. 2002. https://www.tec.ac.cr/administracion-tecnologia-informacion
4. Icafé. *Estructura del sector*. 2015. https://www.icafe.cr/nuestro-cafe/estructura-del-sector/
5. Saharrea, Francisco Aguirre. Producción, beneficiado e industrialización del café en México. *Revista Vinculando* (1999).
6. Icafé. Los Santos. 2015. https://www.icafe.cr/nuestro-cafe/regiones-cafetaleras/lossantos/
7. TEC. *Ejes de Conocimiento Estratégicos 2023 a 2032*. 2023. https://www.tec.ac.cr/ejes-conocimiento-estrategicos-2023-2032
8. ONU. *Objetivo 2: Poner fin al hambre*. 2024. https://www.un.org/sustainabledevelopment/es/hunger/
9. ONU. *Objetivo 12: Garantizar modalidades de consumo y producción sostenibles*. 2024. https://www.un.org/sustainabledevelopment/es/sustainable-consumption-production/
10. Garzón, David. *Diseño de un modelo predictivo de la oferta de aguacate en el municipio de Herveo Tolima*. Universidad Santo Tomás de Aquino.
11. Suhardi, A, Amalia, S, Oktafien, S, Adiyanti, Komariah, S, and Rohendra, T. Time Series Analysis to Predicting Demand of Roasted Coffee. *International Journal of Financial Research*, 10 (2019), 26-31.
12. Vijayan, K, Vennila, J, Sebastian, L, and Rita, S. Predictive Modelling for Coffee Production Using R Programming. In *2022 3rd International Conference on Communication, Computing and Industry 4.0* (2022), IEEE, 1-6.
13. Tolentino, R and Hernández, A. Assessment of Predictive Models for Coffee Production in the Philippines. In *2018 16th International Conference on ICT and Knowledge Engineering (ICT&KE)* (2018), IEEE, 1-6.
14. Kittichotsatsawat, Y, Tippayawong, N, and Tippayawong, K. Prediction of arabica coffee production using artificial neural network and multiple linear regression techniques. *Scientific Reports* (2022).
15. Khumaidi, A. Data mining for predicting the amount of coffee production using CRISP-DM method. *Journal Techno Nusa Mandiri*, 17 (2020), 1-8.
16. Jiawei, Han, Kamber, Michelline, and Pei, Jian. *Data mining: concepts and techniques*. Morgan Kaufmann, 2012.
17. Shearer, Colin. The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5 (2000), 13-22.
18. Shafique, Umair and Qaiser, Haseeb. A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, 12 (2014), 217-222.
19. Luna, Zipporah. *Understanding CRISP-DM and its importance in Data Science projects*. 2021.
20. Hotz, Nick. *What is CRISP DM?* 2024. https://www.datascience-pm.com/crisp-dm-2/
21. IBM. *Guía de CRISP-DM de IBM SPSS Modeler*. 2018. https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf
22. Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth. *CRISP-DM 1.0 Step-by-step data mining guide*. 2000. https://mineracaodedados.wordpress.com/wp-content/uploads/2012/12/crisp-dm-1-0.pdf
23. Luqueño and Peña. Autoregressive integrated moving average (Arima) modeling. *Encyclopedia of Statistics in Quality and Reliability*. (2008).

24. Yadav, Dinesh Kumar and Soumya, K and Goswami, Laxmi. Autoregressive Integrated Moving Average Model for Time Series Analysis. (2024), IEEEE, 1-6.

25. Rueda, Karen and Cardozo, Shirley. Aplicación de redes neuronales artificiales para el pronóstico de precios de café. *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, 1 (2022), 113-117.

26. Rollón, Álvaro. *Experimentos computacionales en un estudio de simulación de modelos de regresión para una mejor comprensión de las herramientas Random Forests y Conditional Trees*. Industriales. 2016.

27. Andrade, Vinicio and Flores, Pablo. Comparativa entre classification trees, random forest y *gradient boosting*; en la predicción de la satisfacción laboral en Ecuador. *Ciencia Digital*, 2 (2018), 42-54.

28. Gallego, Antonio, Rico, Juan, and Valero, Jose. Efficient -nearest neighbor search based on clustering and adaptive values. *Pattern recognition*, 122 (2022).

29. Paz, Guillermina Baena. *Metodología de la investigación*. Grupo Editorial Patria, 2017.

30. Ileana Ulate, Elizarda Vargas. *Metodología para elaborar una tesis*. 2016.

31. Hernández Sampieri, Fernández Collado, Baptista Lucio. *Metodología de la investigación*. 2014.

32. Aranda, Tómas Campoy and Aráujo, Elda Gómez. Técnicas e instrumentos cualitativos de recogida de datos. In *Manual básico para la realización de tesinas, tesis y trabajos de investigación*. Editorial EOS, 2009.

33. Fundibeq. *Hojas de comprobación y hojas de recogida de datos*.

34. Banco Mundial. *El crecimiento mundial se estabiliza por primera vez en tres años*. 2024.

35. López, Olga Lucía Ocampo and Herrera, Lina María Álvarez. Tendencia de la producción y el consumo del café en Colombia. *Apuntes del CENES*, 36 (2017), 139-165.

36. Donnet, Laura, Weatherspoon, Dave D., and Hoehn, John P. Price determinants in top-quality e-auctioned specialty coffees. *Agricultural Economics*, 38 (2008), 267-276.

37. Rotta, Neil McCandless. *Operations and Mass Flows in Postharvest Processing of Coffee: Comprehensive Case Study in Washed Coffee*. University of California, Davis. 2020.

38. Cazorla, Mario Obando. Factores productivos asociados a la exportación de café de la Región Cusco, periodo 2013-2018. *Revista Científica INTEGRACIÓN*, 6 (2023), 61-68.

39. Dia, Lucielma de Oliveira and Silva, Marcelo dos Santos da. Determinantes da demanda internacional por café brasileiro. *Revista de Política Agrícola*, 24 (2015), 86-98.

40. Barreto, Ricardo Candéa Sá and Zugaib, Antônio César Costa. Dynamics of the international coffee market and instrumental in price formation. *Economia & Região*, 4 (2016), 7-27.

41. Icafé. *20 de Setiembre de 2024 - Café informado - Cosecha 23-24*. Icafé, 2024. https://www.icafe.cr/wp-content/uploads/comercializacion/cafe_recibido/actual/18%20de%20Setiembre%20de%202024%20-%20Caf%C3%A9%20informado%20-%20Cosecha%2023-24.xlsx

42. Icafé. *18 de Setiembre de 2023 - Café informado - Cosecha 22-23*. Icafé, 2923. https://www.icafe.cr/wp-content/uploads/comercializacion/cafe_recibido/historico/2023-09.zip

43. Icafé. *30 de Junio de 2022 - Café informado cosecha 21-22*. Icafé, 2022. https://www.icafe.cr/wp-content/uploads/comercializacion/cafe_recibido/historico/2022-06.zip

44. Icafé. *02 de julio de 2021 Café informado - Cosecha 20-21*. Icafé, 2021. https://www.icafe.cr/wp-content/uploads/comercializacion/cafe_recibido/historico/2021-07.rar

45. Icafé. *30 de julio de 2020 Café informado - Cosecha 19-20*. Icafé, 2020. https://www.icafe.cr/wp-content/uploads/comercializacion/cafe_recibido/historico/2020-07.rar

46. Icafé. *25 de julio de 2019 Café informado - Cosecha 18-19*. Icafé, 2019. https://www.icafe.cr/wp-content/uploads/comercializacion/cafe_recibido/historico/2019-07.rar

47. Icafé. *15 de agosto 2018 Café informado - Cosecha 17-18*. Icafé, 2018. https://www.icafe.cr/wp-content/uploads/comercializacion/cafe_recibido/historico/2018-08.rar

48. Icafé. *24 de agosto de 2017 Café informado - Cosecha 16-17*. Icafé, 2017. https://www.icafe.cr/wp-content/uploads/comercializacion/cafe_recibido/historico/2017-08.rar

49. Molina, Marco Antonio Araya. *Costos de beneficiado de café aceptados por ley no 2762 cosecha 2022-2023*. Instituto del café de Costa Rica, San José, 2023. https://www.icafe.cr/wp-content/uploads/informacion_mercado/costos_actividad/beneficiado/ECBC2223.pdf

50. Molina, Marco Antonio Araya. *Costos de beneficiado de café aceptados por ley no 2762 cosecha 2021-2022*. Instituto del café de Costa Rica (Icafé), San José, 2022. https://www.icafe.cr/wp-content/uploads/informacion_mercado/costos_actividad/beneficiado/ECBC2122.pdf

51. Molina, Marco Antonio Araya. *Costos de beneficiado de café aceptados por ley no 2762 cosecha 2020-2021*. Instituto del café de Costa Rica (Icafé), San José, 2021. https://www.icafe.cr/wp-content/uploads/informacion_mercado/costos_actividad/beneficiado/ECBC2021.pdf

52. Molina, Marco Antonio Araya. *Costos de beneficiado de café aceptados por ley no 2762 cosecha 2019-2020*. Instituto del café de Costa Rica (Icafé), San José, 2020. https://www.icafe.cr/wp-content/uploads/informacion_mercado/costos_actividad/beneficiado/ECBC1920.pdf

53. Meteoblue. *Cambio climático Cerros de Dota*. 2024. https://www.meteoblue.com/es/tiempo/historyclimate/change/cerros-de-dota_costa-rica_3623891

54. Icafé. *Precio de liquidación final cosecha 2022-2023*. Icafé, San José, 2023. https://www.icafe.cr/wp-content/uploads/liquidaciones_beneficios/Cosecha%202022-2023%20Publicación%20Liquidación%20Final.pdf

55. Icafé. *Precio de liquidación final cosecha 2021-2022*. Icafé, San José, 2022. https://www.icafe.cr/wp-content/uploads/liquidaciones_beneficios/Cosecha%202021-2022%20Publicación%20Liquidación%20Final.PDF

56. Icafé. *Precio de liquidación final cosecha 2020-2021*. Icafé, San José, 2021. https://www.icafe.cr/wp-content/uploads/liquidaciones_beneficios/Cosecha%202021%20Publicacion%20Liquidacion%20Final.pdf

57. Icafé. *Precio de liquidación final cosecha 2019-2020*. Icafé, San José, 2020. https://www.icafe.cr/wp-content/uploads/liquidaciones_beneficios/Cosecha%201920%20Publicacion%20Liquidacion%20Final.pdf

58. Icafé. *Precio de liquidación final cosecha 2018-2019*. Icafé, San José, 2019. https://www.icafe.cr/wp-content/uploads/liquidaciones_beneficios/Cosecha%201819%20Publicacion%20Liquidacion%20Final.pdf

59. Icafé. *Precio de liquidación final cosecha 2017-2018*. Icafé, San José, 2018. https://www.icafe.cr/wp-content/uploads/liquidaciones_beneficios/Cosecha%201718%20Publicacion%20Liquidacion%20Final.pdf

60. Icafé. *Precio de liquidación final cosecha 2016-2017*. Icafé, San José, 2017. https://www.icafe.cr/wp-content/uploads/liquidaciones_beneficios/Cosecha%201617%20Publicacion%20Liquidacion%20Final.pdf

61. Grupo Banco Mundial. *Crecimiento del PIB (% anual)*. World Bank Open Data. 2023. https://datos.bancomundial.org/indicador/NY.GDP.MKTP.KD.ZG

62. Icafé. *Informe sobre la actividad cafetalera de Costa Rica 2021*. Icafé, San José, 2021. https://www.icafe.cr/wp-content/uploads/informes_gestion/actividad_cafetalera/Informe%20Actividad%20Cafetalera%20de%20Costa%20Rica%202021.pdf

63. Icafé. *Precios históricos del Café en Nueva York – 2024*. Icafé, San José, 2024. https://www.icafe.cr/wp-content/uploads/informacion_mercado/estadisticas_precios/precios/historico/2024.xls

64. Icafé. *Precios históricos del Café en Nueva York - 2023*. Icafé, San José, 2024. https://www.icafe.cr/wp-content/uploads/informacion_mercado/estadisticas_precios/precios/historico/2023.xls

65. Icafé. *Precios históricos del Café en Nueva York - 2022*. Icafé, San José, 2023. https://www.icafe.cr/wp-content/uploads/informacion_mercado/estadisticas_precios/precios/historico/2022.xls

66. Icafé. *Precios históricos del Café en Nueva York - 2021*. Icafé, San José, 2022. https://www.icafe.cr/wp-content/uploads/informacion_mercado/estadisticas_precios/precios/historico/2021.xls

67. Icafé. *Precios históricos del Café en Nueva York - 2020*. Icafé, San José, 2021. https://www.icafe.cr/wp-content/uploads/informacion_mercado/estadisticas_precios/precios/historico/2020.xls

68. Icafé. *Precios históricos del Café en Nueva York – 2019*. Icafé, 2019. https://www.icafe.cr/wp-content/uploads/informacion_mercado/estadisticas_precios/precios/historico/2019.xls

69. Icafé. *Precios históricos del Café en Nueva York – 2018*. Icafé, 2018. https://www.icafe.cr/wp-content/uploads/informacion_mercado/estadisticas_precios/precios/historico/2018.xls

70. Icafé. *Precios históricos del Café en Nueva York – 2017*. Icafé, 2017. https://www.icafe.cr/wp-content/uploads/informacion_mercado/estadisticas_precios/precios/historico/2017.xls

71. Fundibeq. *Diagrama de flechas*.

72. Fundibeq. *Diagrama matricial*.

Capítulo 10. Apéndices

Apéndice A. Plantilla para minutas de reunión

Minuta número	Minuta-XX	Fecha	
Medio de comunicación		Hora de inicio	
		Hora de finalización	
Motivo de reunión			
Personas convocadas			
Presentes			
Ausentes			
Temas tratados			
Número de tema	Asunto		Comentarios
Acuerdos			
Número de tema	Detalle		
Próxima reunión			
Tema por abordar	Fecha	Comentario	

Apéndice B. Plantilla de firma de minutas del profesor tutor

Número de reunión	Firma del profesor tutor
El profesor tutor valida la participación en las siguientes minutas: Reu-XX Reu-N	

Apéndice C. Plantilla para la gestión del cambio

Solicitud de cambio	
ID de solicitud	
Fecha de solicitud	
Responsable	
Prioridad del cambio	Urgencia: • Alta • Media • Baja
Fecha de realización	
Descripción de la solicitud	
Estado	

Apéndice D. Perfil del macroproceso de beneficiado de café

Nombre del macroproceso: *Beneficiado y venta del café*

Ítem por evaluar	Anotaciones
1. Visión: *el objetivo del proceso de beneficiado del café es transformar la fruta del café en un producto final de alta calidad (café oro) que esté listo para su comercialización, asegurando que todas las etapas del procedimiento se ejecuten eficientemente y con los estándares de calidad requeridos para satisfacer las demandas del mercado nacional e internacional*	
2. Dueño del proceso: *el beneficiador de café, quien se encarga de recibir, procesar, financiar y vender el café*	
3. Cliente(s) del proceso: • Exportadores • Torrefactores • Consumidores nacionales e internacionales.	4. **Expectativas del cliente:** • Provisión de café oro de alta calidad. • Procesamiento eficiente y oportuno del café.

	• Transparencia y confiabilidad en el proceso de beneficiado.

5. Resultados: *café oro listo para la venta, satisfacción de las expectativas de los productores y exportadores en términos de calidad y tiempo de entrega.*

6. Disparador: *la recepción de la fruta madura de café de los productores.*

7. Actividades del proceso:

1. *Recepción de la fruta.*
2. *Proceso de limpieza y flote (los granos que flotan son de mala calidad, por lo tanto, se separan)*
3. *Chancado o despulpado de la fruta.*
4. *Secado del café (15 días aproximadamente).*
5. *Almacenamiento del café pergamino (2 meses aproximadamente).*
6. *Pelado y transformación en café oro.*
7. *Envío de muestras a compradores.*
8. *Comercialización del café.*

8. Interfaces de entrada: *proceso de producción del café en las plantaciones (productores), procedimiento de recolección y transporte del café, desde las fincas hasta el beneficiador.*

9. Interfaces de salida: *entregar a entregado.*

10. Recursos requeridos:

a. Recursos humanos:

- *Encargados de transporte de la fruta.*
- *Operarios de las máquinas.*
- *Jornaleros*

b. Entorno de trabajo, materiales, infraestructura:

- *Equipos de protección personal para los trabajadores.*

- *Maquinaria para el despulpado, lavado y pelado del café.*
- *Equipos de protección personal.*
- *Instalaciones para almacenamiento.*

11. Medidas de rendimiento del proceso: *tiempo de ciclo, cantidad de flotes (café de mala calidad), porcentaje de volumen perdido durante el procedimiento.*

12. Observaciones adicionales:

La calidad del café es crítica y se evalúa después del procesamiento completo.

La planificación adecuada de la producción es esencial para evitar la sobreproducción o la falta de producto

Apéndice E. Minuta 1 Reunión inicial con un propietario del microbeneficio Tributos del Ota

Minuta número	Minuta-01	Fecha	29-05-2024
Medio de comunicación	Comunicación personal	Hora de inicio	12: 00
		Hora de finalización	13:00
Motivo de reunión	Desafíos y problemáticas que enfrentan los microbeneficios de café en la región		
Personas convocadas	Víctor Romero Chacón Julio Romero Chacón		
Presentes	Víctor Romero Chacón Julio Romero Chacón	VICTOR JAVIER ROMERO CHACON (FIRMA) Firmado digitalmente por VICTOR JAVIER ROMERO CHACON (FIRMA) Fecha: 2024.11.24 13:18:43 -08'00'	
Ausentes			
Temas tratados			
Número de tema	Asunto	Comentarios	
1	Contexto	El propietario explicó que el proceso de beneficiado del café en su microbeneficio incluye la recolección de la fruta madura, el chancado, el secado y la posterior comercialización.	

		Destacó que realizan todas las etapas del proceso para eliminar intermediarios y maximizar las ganancias. Comentó que este enfoque permite tener un mejor control sobre la calidad del producto final y ofrecer un café de mayor valor agregado a compradores locales e internacionales.
2	Problemática	El propietario manifestó que uno de los mayores problemas que enfrentan es la incertidumbre en la cantidad de café que deben procesar. Esta falta de previsión afecta la planificación y la eficiencia operativa.
3	Causas	El propietario mencionó que la volatilidad del mercado internacional contribuye significativamente a esta incertidumbre, ya que los precios y la demanda fluctúan de manera constante debido a factores climáticos y económicos globales. Indicó que la calidad del café es un factor crítico en la comercialización, pero la evaluación de esta calidad solo se puede realizar después del procesamiento completo, lo que añade una capa de incertidumbre en la etapa de planificación. Explicó que la falta de información confiable sobre la demanda y los precios del mercado impide una planificación adecuada de la producción.
4	Efectos	El propietario señaló que la falta de previsión puede llevar a la sobreproducción, lo que ocasiona un exceso de café beneficiado que no se puede vender y genera costos adicionales por almacenamiento y depreciación. Por otro lado, no procesar suficiente café para satisfacer la demanda puede llevar a la pérdida de oportunidades de venta y, en consecuencia, a una disminución de los ingresos.
Acuerdos		
Número de tema	Detalle	
Próxima reunión		

Tema por abordar	Fecha	Comentario
Por definir	Por definir	N/A

Apéndice F. Plantilla de entrevista

Entrevista # N		
Fecha de la entrevista	XX-XX-2024	
Hora de inicio		
Hora de finalización		
Objetivo de la entrevista		
Participantes		
Preguntas		
1	Pregunta n.º 1	Respuesta 1
2	Pregunta n.º 2	Respuesta 2
n	Pregunta n	Respuesta n

Apéndice G. Plantilla de revisión documental – Factores relevantes.

ID	**Fecha**	**Factor**	**Estudio**	**Justificación**
1		Factor 1		
2		Factor 2		
n		Factor n		

Apéndice H. Plantilla de revisión documental-fuentes de información

ID	Fecha	Factor	Fuente de información de la variable	Hallazgo
1		Factor 1		
2		Factor 2		
n		Factor n		

Apéndice I. Plantilla de diagrama matricial

Factor	Grado de influencia del factor en la proyección de la demanda del café				
	Muy bajo	Bajo	Moderado	Alto	Muy alto
Factor 1					
Factor 2					
Factor n					

Apéndice J. Plantilla de hoja de recogida de datos para métricas de calidad de datos

Indicador de calidad de los datos	Definición de indicador	Número de registros
Valores nulos	Indica cuántos valores faltan en un campo específico del conjunto de datos.	
Registros duplicados	Se refiere a la presencia de registros idénticos en los datos.	
Datos inconsistentes	Indican la presencia de discrepancias lógicas o contradicciones dentro de los datos, como fechas que no siguen una secuencia cronológica.	
Datos no uniformes	Indican que los datos no siguen un formato estándar o presentan variaciones en las unidades de medida o en la representación de la información.	

Apéndice K. Plantilla de hoja de recogida de datos para parámetros de los modelos

Modelo	Iteración	Parámetros			Resultados		
		Parámetro 1	Parámetro 2	Parámetro n	Métrica de evaluación 1	Métrica de evaluación 2	Métrica de evaluación n
Modelo 1	1						
Modelo 1	2						
Modelo 1	n						
Modelo 2	1						
Modelo 2	2						
Modelo 2	n						
Modelo n	1…n						

Apéndice L. Plantilla de script de limpieza de datos

```
> # Cargar dataset
Cargar dataset desde la fuente

# Aplicar reglas de limpieza de datos
Aplicar Regla 1: Identificación y tratamiento de valores nulos
Aplicar Regla 2: Corrección de inconsistencias en formato
Aplicar Regla 3: Reemplazo de valores erróneos
Aplicar Regla 4: Conversión de tipos de datos
Aplicar Regla 5: Creación de nuevas variables
Aplicar Regla 6: Filtrado de datos no relevantes

# Validación final
Validar calidad de los datos

# Guardar dataset limpio
Guardar dataset limpio en la ubicación especificada
```

Apéndice M. Plantilla de hoja de comprobación

Modelos	Iteración	Objetivos de minería de datos		
		Objetivo de minería 1	Objetivo de minería 2	Objetivo de minería n
Modelo 1				
Modelo 2				
Modelo n				

Apéndice N. Revisión documental de factores relevantes

ID	Fecha	Factor	Estudio	Hallazgo
Factores locales				
1	09-09-2024	Datos históricos propios.		Debido a que esencialmente esta es la variable que se desea predecir, es necesario tener un registro histórico de la variable.
2	09-09-2024	Costos de producción	Tendencia de la producción y el consumo del café en Colombia 35	El estudio sostiene que el costo de producción es una de las causas estructurales de las tendencias decrecientes de producción.
3	10-09-2024	Variedad del café	Price determinants in top-quality e-auctioned specialty coffees 36	En la producción de café de especialidad el precio está determinado por la variedad del café y otros atributos de calidad.
4	12-09-2024	Capacidad de beneficiado	Operations and Mass Flows in Postharvest Processing of Coffee: Comprehensive Case Study in Washed Coffee 37	La investigación señala que la capacidad de procesamiento influye directamente en el volumen de café procesado, ya que este último no puede exceder su capacidad.
Factores climáticos				
5	11-09-2024	Temperatura media	Factores productivos asociados a la exportación de café de la región Cusco, periodo 2013-2018 38	El estudio señala que un aumento en la temperatura ambiental tiene un impacto negativo en la capacidad de producción de café.
6	12-09-2024	Precipitaciones	Data mining for predicting the amount of coffee production using CRISP-DM method 15	El análisis muestra que las precipitaciones influyen en la producción de café, junto con otros factores ambientales.
Factores económicos				

ID	Fecha	Factor	Estudio	Hallazgo
7	11-09-2024	Tipo de cambio	Determinantes da demanda internacional por café brasileiro 39	Se considera que el tipo de cambio influye en la demanda internacional del café. Considerándola como una de las variables que se utilizan en la ecuación de demanda y oferta del café.
8	12-09-2024	Crecimiento económico mundial	Dynamics of the international coffee market and instrumental in price formation 40.	Según el estudio, la demanda de café está estrechamente ligada al crecimiento de la economía global. A mayor crecimiento, se espera un aumento en la demanda y, por ende, un incremento en los precios del café
Factores del mercado de café				
9	11-09-2024	Producción nacional	Determinantes da demanda internacional por café brasileiro 39	El estudio destaca la producción nacional como una de las variables de la ecuación de la oferta.
10	12-09-2024	Oferta mundial	Dynamics of the international coffee market and instrumental in price formation 40.	La oferta depende de factores como la producción y los inventarios existentes.
11	12-09-2024	Demanda mundial	Dynamics of the international coffee market and instrumental in price formation 40.	Este factor se encuentra íntimamente relacionado con el consumo global de café.
12	11-09-2024	Precio según la Bolsa de valores de Nueva York	Determinantes da demanda internacional por café brasileiro 39	El precio del café en el mercado de futuros de Nueva York influye directamente en el valor del café en el comercio internacional.

Apéndice Ñ. Entrevista con experto

Entrevista #1		
Fecha de la entrevista	13-09-2024	
Hora de inicio	4:00 p. m.	
Hora de finalización	4:50 p. m.	
Objetivo de la entrevista	Obtener el criterio de una persona experta en relación con los factores clave que se identifican mediante la revisión documental, que se consideran relevantes para construir un modelo predictivo en el contexto del beneficiado de café. A través de esta entrevista, se busca validar dichos factores, identificar posibles limitaciones y sugerencias para optimizar el modelo, así como definir las métricas más adecuadas para evaluar su calidad.	
Participantes	Lorena Zúñiga Segura (experta) Julio Romero Chacón	
Preguntas		
1	¿Es válido utilizar registros comunes para todos los beneficios, como la producción nacional anual de café?	**Respuesta:** Sí, es válido utilizar registros comunes como la producción nacional anual de café. Sin embargo, al hacerlo, existe el riesgo de que esta columna no tenga un impacto significativo en el modelo. A pesar de esto, por ejemplo, los valores de la producción nacional varían de 1 año a otro, lo cual puede ser útil para identificar tendencias a lo largo del tiempo y contribuir a la construcción del modelo.
2	¿Cuáles limitaciones observa con los factores seleccionados?	**Respuesta:** Una limitación importante es que, al contar con un único registro anual por cada beneficio debido a la naturaleza estacional de la actividad, es probable que no haya suficientes datos o filas para entrenar adecuadamente cada modelo.
3	¿Es necesario realizar un análisis de correlación para los factores relevantes?	**Respuesta**: Sí, es necesario realizar un análisis de correlación para cada uno de los factores relevantes. Este análisis permite descartar columnas que presenten una correlación muy fuerte, positiva o negativa, con otras variables. Esto ayuda a reducir la cantidad de variables en el modelo final.

Entrevista #1		
4	¿Cuántos factores considera que son óptimos para la construcción del modelo?	**Respuesta:** Contar con 12 factores es un número óptimo para construir el modelo, ya que proporciona suficientes variables para respaldar las predicciones sin complicar los resultados con un exceso de columnas.
5	¿Identifica algún factor adicional que debe ser considerado en el modelo?	**Respuesta:** No se identificaron factores adicionales relevantes, pero es fundamental asegurar que todas las unidades de medida sean consistentes entre los registros. Algunas columnas podrían estar en quintales y otras en fanegas, lo cual refleja las diferentes unidades utilizadas en la caficultura. Por lo tanto, será necesario realizar un proceso de conversión o transformación de los datos para asegurar la uniformidad.
6	¿Cuáles métricas considera más adecuadas para evaluar la calidad de los modelos con base en series de tiempo?	**Respuesta:** En este caso las métricas de evaluación más adecuadas dependen de los resultados óptimos de cada métrica. Aunque es difícil establecer criterios de aceptación precisos, es válido utilizar aproximaciones o criterios que se ajusten a los resultados óptimos de las métricas seleccionadas.

Apéndice O. Revisión documental fuentes de información

ID	Fecha	Factor	Fuente de información del factor relevante	Hallazgo
1	21-09-2024	Cantidad de café beneficiado (factor por predecir)	20 de septiembre de 2024-Café-informado Cosecha 23-24 41	Se presenta la cantidad de café procesado por cada empresa en los diferentes periodos, expresada en fanegas. Para el análisis, se optó por considerar los informes finales acumulados de cada año, ya que estos reflejan el total reportado durante todo el año o la cosecha.
			18 de septiembre de 2023-Café informado-Cosecha 22-23 42	
			30 de junio de 2022-Café informado-cosecha 21-22 43	

ID	Fecha	Factor	Fuente de información del factor relevante	Hallazgo
			02 de julio de 2021-Café informado-Cosecha 20-21 44	
			30 de julio de 2020-Café informado-Cosecha 19-20 45	
			25 de julio de 2019 Café informado-Cosecha 18-19 46	
			15 de agosto de 2018-Café informado-Cosecha 17-18 47	
			24 de agosto de 2017-Café informado-Cosecha 16-17 48	
2	19-09-2024	Costo de beneficiado	Costos de beneficiado de café aceptados por ley no 2762 cosecha 2022-2023 49	Los informes estudiados presentan el costo promedio nacional para el beneficiado de café, para cada cosecha. No se encontraron informes de las cosechas 2018-2019, 2017-2018 y 2023-2024, este último debido a que el informe se presenta cada diciembre.
			Costos de beneficiado de café aceptados por ley no 2762 cosecha 2021-2022 50	
			Costos de beneficiado de café aceptados por ley no 2762 cosecha 2020-2021 51	
			Costos de beneficiado de café aceptados por ley no 2762 cosecha 2019-2020 52	
3	25-09-2024	Variedad del café	No se encontró	
4	25-09-2024	Capacidad de beneficiado	No se encontró	
5	20-09-2024	Temperatura media	Cambio climático Cerros de Dota 53	La estadística permite obtener la temperatura media anual en los Cerros

ID	Fecha	Factor	Fuente de información del factor relevante	Hallazgo
				de Dota, con una altitud de 1894 m s. n. m.
6	20-09-2024	Precipitaciones	Cambio climático Cerros de Dota 53	La estadística proporciona la variación total anual de las precipitaciones en la región de los Cerros de Dota, situada a 1894 m s. n. m.
7	20-09-2024	Tipo de cambio	Precio de liquidación final cosecha 2022-2023 54	Se dispone de la estadística que muestra el tipo de cambio al que cada beneficio logró vender su café, calculado anualmente como un promedio de todas las ventas. Este tipo de cambio está expresado en relación con el dólar estadounidense.
			Precio de liquidación final cosecha 2021-2022 55	
			Precio de liquidación final cosecha 2020-2021 56	
			Precio de liquidación final cosecha 2019-2020 57	
			Precio de liquidación final cosecha 2018-2019 58	
			Precio de liquidación final cosecha 2017-2018 59	
			Precio de liquidación final cosecha 2016-2017 60	
8	20-09-2024	Crecimiento económico mundial	Crecimiento del PIB (% anual) 61	Se presenta la tasa de crecimiento anual del producto interno bruto (PIB), según los datos proporcionados por el Banco Mundial. La estadística está disponible desde el año 1961 hasta 2023.
9	19-09-2024	Producción nacional	Informe sobre la actividad cafetalera de Costa Rica 2023 1	Se presentan registros anuales de la cantidad de café producido anualmente en miles de fanegas.
			Informe sobre la actividad cafetalera de Costa Rica 2021 62	

ID	Fecha	Factor	Fuente de información del factor relevante	Hallazgo
10	19-09-2024	Oferta mundial	Informe sobre la actividad cafetalera de Costa Rica 2023 1	Se presenta el dato anual de la producción mundial de café, expresado en millones de sacos de 60 kg.
			Informe sobre la actividad cafetalera de Costa Rica 2021 62	
11	19-09-2024	Demanda mundial	Informe sobre la actividad cafetalera de Costa Rica 2023 [1]	Se presenta el dato anual del consumo mundial de café, expresado en millones de sacos de 60 kg.
			Informe sobre la actividad cafetalera de Costa Rica 2021 62	
12	19-09-2024	Precio según la bolsa de valores de Nueva York	Precios históricos del café en Nueva York – 2024 63	Se cuenta con la estadística del precio promedio anual por quintal de café en dólares estadounidenses, que recopiló el Icafé.
			Precios históricos del café en Nueva York – 2023 64	
			Precios históricos del café en Nueva York – 2022 65	
			Precios históricos del café en Nueva York – 2021 66	
			Precios históricos del café en Nueva York – 2020 67	
			Precios históricos del café en Nueva York – 2019 68	
			Precios históricos del café en Nueva York – 2018 69	
			Precios históricos del café en Nueva York – 2017 70	

ID	Fecha	Factor	Fuente de información del factor relevante	Hallazgo
13	19-09-2024	Precio de liquidación	Precio de liquidación final cosecha 2022-2023 54	Se presenta el registro anual promedio del precio de comercialización de cada beneficiador.
			Precio de liquidación final cosecha 2021-2022 55	
			Precio de liquidación final cosecha 2020-2021 56	
			Precio de liquidación final cosecha 2019-2020 57	
			Precio de liquidación final cosecha 2018-2019 58	
			Precio de liquidación final cosecha 2017-2018 59	
			Precio de liquidación final cosecha 2016-2017 60	

Apéndice P. Datos consolidados

Datos consolidados

Apéndice Q. Cantidad de registros y tipos de datos

```
dim(datos)
```

```
## [1] 96 13
```

Hide

```
str(datos)
```

```
## tibble [96 × 13] (S3: tbl_df/tbl/data.frame)
##  $ Microbeneficio               : chr [1:96] "Microbeneficio 1" "Microbeneficio 10" "Microbeneficio 11" "
##  $ Año                          : num [1:96] 2024 2024 2024 2024 2024 ...
##  $ Cantidad de café beneficiado : num [1:96] 197 70 420 158 174 ...
##  $ Costo de beneficiado         : num [1:96] NA NA NA NA NA NA NA NA NA NA ...
##  $ Temperatura media            : num [1:96] NA NA NA NA NA NA NA NA NA NA ...
##  $ Precipitación                : num [1:96] NA NA NA NA NA NA NA NA NA NA ...
##  $ Tipo de cambio               : num [1:96] NA NA NA NA NA NA NA NA NA NA ...
##  $ Precio liquidación           : num [1:96] NA NA NA NA NA NA NA NA NA NA ...
##  $ Crecimiento económico mundial: num [1:96] NA NA NA NA NA NA NA NA NA NA ...
##  $ Producción nacional          : num [1:96] 1912 1912 1912 1912 1912 ...
##  $ Oferta mundial               : num [1:96] 173 173 173 173 173 ...
##  $ Demanda mundial              : num [1:96] 169 169 169 169 169 ...
##  $ Precio según la bolsa de NY  : num [1:96] 118 118 118 118 118 ...
```

Apéndice R. Estadísticas básicas de los datos

```
##  Microbeneficio          Año       Cantidad de café beneficiado
##  Length:96          Min.   :2017   Min.   : 18.00
##  Class :character   1st Qu.:2019   1st Qu.: 76.72
##  Mode  :character   Median :2020   Median :126.34
##                     Mean   :2020   Mean   :177.31
##                     3rd Qu.:2022   3rd Qu.:238.83
##                     Max.   :2024   Max.   :600.00
##                                    NA's   :24
##  Costo de beneficiado Temperatura media Precipitación  Tipo de cambio
##  Min.   :26569        Min.   :17.60     Min.   :2905   Min.   :531.5
##  1st Qu.:29809        1st Qu.:17.70     1st Qu.:3054   1st Qu.:569.8
##  Median :33638        Median :18.00     Median :3401   Median :581.0
##  Mean   :32833        Mean   :17.99     Mean   :3409   Mean   :593.3
##  3rd Qu.:36661        3rd Qu.:18.30     3rd Qu.:3860   3rd Qu.:624.4
##  Max.   :37488        Max.   :18.30     Max.   :3924   Max.   :684.3
##  NA's   :48           NA's   :12        NA's   :12     NA's   :36
##  Precio liquidación Crecimiento económico mundial Producción nacional
##  Min.   : 75444     Min.   :-2.900                Min.   :1673
##  1st Qu.:131434     1st Qu.: 2.600                1st Qu.:1810
##  Median :155427     Median : 3.100                Median :1899
##  Mean   :173142     Mean   : 2.657                Mean   :1867
##  3rd Qu.:202931     3rd Qu.: 3.500                3rd Qu.:1930
##  Max.   :397656     Max.   : 6.300                Max.   :2018
##  NA's   :36         NA's   :12
##  Oferta mundial  Demanda mundial Precio según la bolsa de NY
##  Min.   :160.6   Min.   :159.5   Min.   :101.2
##  1st Qu.:165.9   1st Qu.:163.9   1st Qu.:113.1
##  Median :167.8   Median :166.5   Median :125.7
##  Mean   :168.5   Mean   :165.9   Mean   :141.6
##  3rd Qu.:172.5   3rd Qu.:169.2   3rd Qu.:169.3
##  Max.   :176.2   Max.   :169.6   Max.   :214.7
##
```

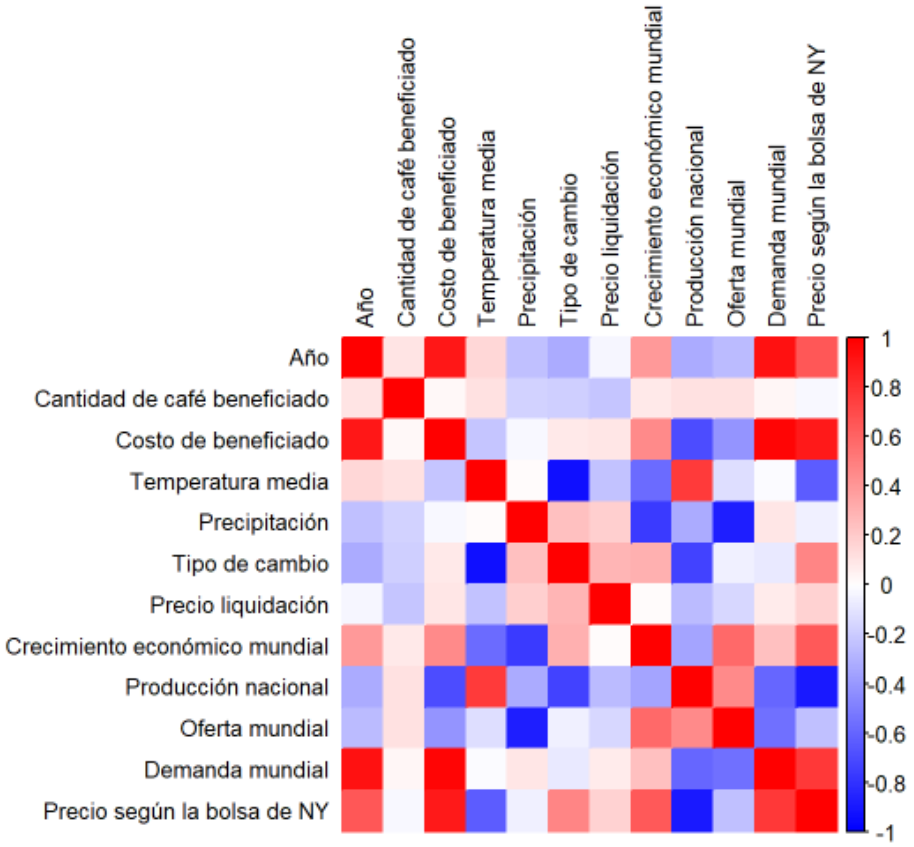
Apéndice S. Script de análisis de correlación

```
# Análisis de correlación
  # Filtrar solo las columnas numéricas
  numerical_data <- datos[, sapply(datos, is.numeric)]

  # Calcular la matriz de correlación solo con las variables numéricas
  correlation_matrix <- cor(numerical_data, use = "complete.obs")

  corrplot(correlation_matrix, method = "color",
           col = colorRampPalette(c("blue", "white", "red"))(200),
           tl.cex = 0.8, number.cex = 0.8,
           tl.col = "black")  # Cambia el color de las etiquetas a negro
```

Apéndice T. Mapa de calor del análisis de correlación



Apéndice U. Script de recolección de métricas de calidad en los datos

```
# Comprobación de la calidad de los datos
  # Valores nulos
   sum(is.na(datos))
```

```
## [1] 180
```

Hide

```
  # valores duplicados
   filasDuplicadas <- duplicated(datos)
   num_duplicados <- sum(filasDuplicadas)
   num_duplicados
```

```
## [1] 0
```

```r
# datos no uniformes
datosNoUniformes <- function(datos) {
totalInconsistencias <- 0  # Variable para contar los valores no uniformes
for (nombreColumna in names(datos)) {
  columna <- datos[[nombreColumna]]
  columnaSinNulos <- columna[!is.na(columna)]
  # Si es una columna numérica, buscar valores que no son numéricos
  if (is.numeric(datos[[nombreColumna]])) {
    # Detectar valores no numéricos
    totalInconsistencias <- totalInconsistencias + sum(!sapply(columnaSinNulos, is.numeric))
  # Si es de tipo fecha, buscar valores que no pueden convertirse a fecha
  } else if (inherits(datos[[nombreColumna]], "Date")) {
    # Detectar valores que no son fechas válidas
    totalInconsistencias <- totalInconsistencias + sum(is.na(as.Date(columnaSinNulos, format = "%Y-%m-%d
  # Si es texto, verificar que no contenga datos no alfabéticos o alfanuméricos
  } else if (is.character(datos[[nombreColumna]])) {
    # Detectar valores que no sean exclusivamente texto alfabético o alfanumérico (permitiendo números e
    totalInconsistencias <- totalInconsistencias + sum(grepl("[^a-zA-Z0-9 ]", columnaSinNulos))
  }
}
return(totalInconsistencias)  # Retornar el número total de datos no uniformes
}
# Aplicar la función al dataset.
totalDatosNoUniformes <- datosNoUniformes(datos)
totalDatosNoUniformes  # Ver el número total de inconsistencias
```

```
## [1] 0
```

Apéndice V. Amputación de factores

```r
```{r}
  # Se descarta la variable Costo de beneficiado.y demanda mundial
  datos <- datos[, -which(names(datos) == "Costo de beneficiado")]
  datos <- datos[, -which(names(datos) == "Demanda mundial")]
  datos
```
```

Apéndice W. Hoja de recogida de datos para métricas de calidad de datos

Indicador de calidad de los datos	Definición de indicador	Número de registros
Valores nulos	Indica cuántos valores faltan en un campo específico del conjunto de datos.	180
Registros duplicados	Se refiere a la presencia de registros idénticos en los datos.	0
Datos inconsistentes	Indican la presencia de discrepancias lógicas o contradicciones dentro de los datos.	0
Datos no uniformes	Indican que los datos no siguen un formato estándar o presentan variaciones en las unidades de medida o en la representación de la información.	0

Apéndice X. Script para el tratamiento de valores nulos

```
  # 1. Eliminar las filas con valores nulos en "Cantidad de café beneficiado"
datos <- datos %>% filter(!is.na(`Cantidad de café beneficiado`))



# 2. Imputación de los valores faltantes con la media:
# Imputar valores en "Temperatura media"
datos$`Temperatura media`[is.na(datos$`Temperatura media`)] <- mean(datos$`Temperatura media`, na.rm = T

# Imputar valores en "Precipitación"
datos$`Precipitación`[is.na(datos$`Precipitación`)] <- mean(datos$`Precipitación`, na.rm = TRUE)

# Imputar "Tipo de cambio" con la media histórica (excepto los valores del 2024)
# Primero, calculamos la media histórica para años anteriores al 2024
mean_historico_tipo_cambio <- mean(datos$`Tipo de cambio`[datos$Año < 2024 & !is.na(datos$`Tipo de cambi

# Ahora imputamos los valores faltantes del tipo de cambio para el 2024 con la media histórica
datos$`Tipo de cambio`[is.na(datos$`Tipo de cambio`) & datos$Año == 2024] <- mean_historico_tipo_cambio

# Imputación similar para "Precio de liquidación"
mean_historico_precio_liquidacion <- mean(datos$`Precio liquidación`[datos$Año < 2024 & !is.na(datos$`Pr
datos$`Precio liquidación`[is.na(datos$`Precio liquidación`) & datos$Año == 2024] <- mean_historico_prec


# 3. Imputación de "Crecimiento económico mundial" con el valor predictivo para 2024 (2.6%)
datos$`Crecimiento económico mundial`[is.na(datos$`Crecimiento económico mundial`) & datos$Año == 2024]

# Verificar el resultado de las imputaciones y tratamiento de valores nulos
sum(is.na(datos))
```

Apéndice Y. Transformación de los datos

```
# Transformar la producción nacional de miles de fanegas a fanegas
datos$`Producción nacional` <- datos$`Producción nacional` * 1000
# Transformar la oferta y demanda mundial de millones de sacos de 60 kg a fanegas
conversion_sacos_a_fanegas <- 1.30

# Multiplicar los valores de demanda y oferta mundial por 1,000,000 y por la conversión a fanegas
datos$`Demanda mundial` <- datos$`Demanda mundial` * 1e6 * conversion_sacos_a_fanegas
datos$`Oferta mundial` <- datos$`Oferta mundial` * 1e6 * conversion_sacos_a_fanegas
```

Apéndice Z. Normalización de los datos

```
  # Seleccionar solo las columnas numéricas
numerical_columns <- sapply(datos, is.numeric)

# Normalizar todas las columnas numéricas (valores entre 1 y 10)

datos_normalizados <- datos
datos_normalizados[numerical_columns] <- lapply(datos[numerical_columns], function(x) (x - min(x)) * (10 - 1) / (max(x) - min(x)) + 1)


# Guardar el archivo normalizado
write_xlsx(datos_normalizados, path = "datos2_normalizados.xlsx")
```

Apéndice AA. Codificación one-hot

```
#Crear dummies y eliminar la columna original de Microbeneficio
# Al usar drop_first = TRUE, eliminamos una columna dummy por cada conjunto de variables categóricas para evitar colinealidad
datos_con_dummies <- dummy_cols(datos_normalizados, select_columns = "Microbeneficio", remove_first_dummy = TRUE)
datos_con_dummies <- datos_con_dummies %>% select(-Microbeneficio)
```

Apéndice BB. Documento de R

Documento de R

Apéndice CC. Hoja de recogida de datos para los modelos entrenados

Modelo	Iteración	Parámetros	Resultados		
			MAPE	R2 – Coeficiente de determinación	Variables más relevantes
Redes neuronales	1	Linout=TRUE Size = 1	18.24 %	88.57 %	Oferta mundial y año
Random Forest	1	Ntree = 500	13,96y	93.64 %	Tipo de cambio y año
Gradient boosting	1	n.trees = 50 Interation.depth = 1 shrinkage = 0.1	54.84 %	19.95 %	Precio de liquidación y tipo de cambio
K-nearest neighbors (KNN)	1	K=23	71.56 %	4.03 %	No aplica
Arimax	1	Seasonal=False p = 0 d = 0 q = 0	21.39 %	87.10 %	Año y oferta mundial
Sarimax	1	Seasonal=TRUE p = 0 d = 0 q = 0	21.39 %	87.10 %	Año y oferta mundial
Redes neuronales	2	Linout=TRUE Size = 5 Decay = 0,1	5.69 %	99.14 %	Crecimiento económico mundial y precio según la bolsa de Nueva York.

Modelo	Iteración	Parámetros	Resultados		
			MAPE	R2 – Coeficiente de determinación	Variables más relevantes
Random Forest	2	Ntree = 100	13,99	93.71 %	Tipo de cambio y año
Gradient boosting	2	n.trees = 200 Interation.depth = 3 shrinkage = 0.05	51.72 %	33.42 %	Precio de liquidación y tipo de cambio.
K-nearest neighbors (KNN)	2	K=2	46.51 %	46.59 %	No aplica
Redes neuronales	3	Linout=TRUE Size = 5 Decay = 0,2	12.39 %	95.31 %	Producción nacional y precipitaciones.
Random Forest	3	Ntree=50	14,29%	93,93%	Tipo de cambio y oferta mundial
Gradient boosting	3	n.trees = 300 Interation.depth = 3 shrinkage = 0.01	57.83 %	20.50 %	Precio liquidación y tipo de cambio
K-nearest neighbors (KNN)	3	K=8	59.86 %	12.49 %	No aplica

Apéndice DD. Hoja de comprobación

Modelos	Iteración	Objetivos de minería de datos		
		Lograr un coeficiente de determinación del 80 %	Obtener un MAPE (error porcentual absoluto medio) máximo de 10 %	Identificar al menos dos factores de relevancia del modelo
Redes neuronales	1	X		X
Random forest	1	X		X
Gradient boosting	1			X

Modelos	**Iteración**	**Objetivos de minería de datos**		
		Lograr un coeficiente de determinación del 80 %	**Obtener un MAPE (error porcentual absoluto medio) máximo de 10 %**	**Identificar al menos dos factores de relevancia del modelo**
K-nearest neighbors (KNN)	1			
Arimax	1	X		X
Sarimax	1	X		X
Redes neuronales	2	X	X	X
Random forest	2	X		X
Gradient boosting	2			X
K-nearest neighbors (KNN)	2			
Redes neuronales	3	X		X
Random forest	3	X		X
Gradient boosting	3			X
K-nearest neighbors (KNN)	3			

Apéndice EE Minuta reunión 2

Minuta número	Minuta-02	Fecha	23-05-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	15: 20
		Hora de finalización	15:50
Motivo de reunión	Acercamiento inicial con el profesor tutor		
Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón		
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón		
Ausentes			
Temas tratados			

Número de tema	Asunto	Comentarios
1	Dinámica	Se propuso realizar una agenda antes de cada reunión con los puntos para tratar.
2	Problemática	Recomendaciones para plantear el problema.
3	Libro de referencia	Recomendaciones acerca del libro base para realizar la metodología.
4	Temática	Se plantearon dos posibles temáticas para el trabajo de investigación
Acuerdos		
Número de tema	Detalle	
3	Se acordó trabajar con el libro: Metodología para elaborar una tesis, Ileana Ulate y Elizarda Vargas UNED	
4	Se acordó tener la redacción del problema y el tema para la siguiente reunión	
Próxima reunión		
Tema por abordar	Fecha	Comentario
Seguimiento del tema y el problema	28-05-2024	N/A

Apéndice FF Minuta reunión 3

Minuta número	Minuta-03	Fecha	28-05-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	16: 16
		Hora de finalización	16:55
Motivo de reunión	Seguimiento del tema y el problema		
Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón		
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón		
Ausentes			

Temas tratados		
Número de tema	Asunto	Comentarios
1	Avances	Se presentaron los avances en cuanto al problema y la temática y se brindaron recomendaciones.
2	Fuentes de información	Se incentivó a buscar fuentes de información primaria e investigación para respaldar el problema y el anteproyecto en general.
Acuerdos		
Número de tema	Detalle	
1	Se acordó reducir el alcance de la investigación.	
Próxima reunión		
Tema por abordar	Fecha	Comentario
Avances del anteproyecto	2-06-2024	N/A

Apéndice GG Minuta reunión 4

Minuta número	Minuta-04	Fecha	2-06-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	11:10
		Hora de finalización	11:25
Motivo de reunión	Avances del anteproyecto		
Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón		
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón		
Ausentes			
Temas tratados			
Número de tema	Asunto	Comentarios	
1	Presentación de avances	Se presentaron los avances del anteproyecto (beneficios, objetivos, problemática depurada).	

Acuerdos		
Número de tema	Detalle	
1	Se acordó enviar el documento al profesor para comentar las observaciones en la siguiente reunión.	
Próxima reunión		
Tema por abordar	Fecha	Comentario
Observaciones acerca del avance enviado al profesor.	3-06-2024	N/A

Apéndice HH Minuta reunión 5

Minuta número	Minuta-05	Fecha	3-06-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	16:00
		Hora de finalización	16:30
Motivo de reunión	Observaciones acerca del avance enviado al profesor.		
Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón		
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón		
Ausentes			
Temas tratados			
Número de tema	Asunto	Comentarios	
1	Observaciones del avance	Se aprobó el problema, objetivos y beneficios que se plantearon.	
Acuerdos			
Número de tema	Detalle		
1	Se acordó enviar al profesor la versión terminada del anteproyecto.		
Próxima reunión			
Tema por abordar	Fecha	Comentario	
Por definir	Por definir	N/A	

Apéndice II Minuta reunión 6

Minuta número	Minuta-06	Fecha	7-06-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	9:30
		Hora de finalización	10:00
Motivo de reunión	Criterio de experto de la profesora Lorena con respecto a la temática de la investigación.		
Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón Lorena Zúñiga Segura		
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón Lorena Zúñiga Segura		
Ausentes			
Temas tratados			
Número de tema	Asunto	Comentarios	
1	Título	La profesora recomendó cambiar el título de la propuesta, de manera que se oriente al diseño de un modelo predictivo, en lugar de un prototipo de *software* funcional.	
2	Fases	Se recomendó que la primera fase del proyecto sea determinar los factores cuantitativos y cualitativos que pueden ayudar a construir el modelo predictivo y mediante qué fuente puede ser que se recopiló.	
3	Observación	La profesora mencionó que un posible problema en el proyecto es la disponibilidad de la información necesaria, principalmente relacionado con registros históricos y consistencia de los datos, por lo tanto, recomendó utilizar 1 año base.	
4	Marco de referencia	Se recomendó orientar el proyecto hacia el cumplimiento del marco de referencia CRISP-DM.	
Acuerdos			

Número de tema	Detalle	
1	Se modificó el título del proyecto.	
2	Se siguió la recomendación.	
4	Se orientaron las fases del proyecto hacia el cumplimiento del marco de referencia.	
Próxima reunión		
Tema por abordar	Fecha	Comentario
Por definir	Por definir	N/A

Apéndice JJ Minuta reunión 7

Minuta número	Minuta-07	Fecha	28-08-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	11:00
		Hora de finalización	12:10
Motivo de reunión	Definición de aspectos logísticos y forma de trabajo para seguir durante el desarrollo del TFG.		
Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón		
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón		
Ausentes			
Temas tratados			
Número de tema	Asunto	Comentarios	
1	Cronograma	Se estableció un cronograma de reuniones semanales.	
2	Asignación de tareas	Se estableció la herramienta Todolis para asignar las tareas de cada semana y monitorear su avance.	
3	Anteproyecto	El profesor recomendó incluir en el apartado de antecedentes una sección orientada al sector cafetalero y sus principales implicados.	

4	Marco teórico	Se acordó comenzar con este capítulo durante la semana 2 e incluir la metodología CRISP-DM y métodos estadísticos potencialmente útiles en el estudio.
Acuerdos		
Número de tema	Detalle	
1	Se acordó mantener una reunión cada semana, los miércoles.	
2	Se asignaron las primeras tareas.	
Próxima reunión		
Tema por abordar	Fecha	Comentario
Avances de la semana 2	31-07-2024	N/A

Apéndice KK Minuta reunión 8

Minuta número	Minuta-08	Fecha	31-08-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	4:00
		Hora de finalización	4:40
Motivo de reunión	Presentación de avances de semana 2.		
Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón		
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón		
Ausentes			
Temas tratados			
Número de tema	Asunto	Comentarios	
1	Formato	Se recomendó trabajar con tres niveles de títulos en el documento final y agregar siempre una introducción después de cada título.	

2	Antecedentes	Se presentó el trabajo realizado relacionado con los involucrados del sector.
Acuerdos		
Número de tema	Detalle	
2	Se acordó terminar para la próxima semana los antecedentes del macroproceso de beneficiado.	
Próxima reunión		
Tema por abordar	Fecha	Comentario
Avances de semana 3	7-08-2024	

Apéndice LL Minuta reunión 9

Minuta número	Minuta-09	Fecha	9-08-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	11:00
		Hora de finalización	11:40
Motivo de reunión	Presentación de avances de semana 3.		
Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón		
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón		
Ausentes			
Temas tratados			
Número de tema	Asunto	Comentarios	
1	Estado del arte	Se discutieron las secciones por incluir en el estado del arte.	
2	Teoría de conceptos y teorías del estado del arte.	Se recomendó incluir aspectos técnicos del beneficiado del café, es decir, la jerga que se utiliza en el sector.	

3	Metodología	Se recomendó empezar el desarrollo de este capítulo, además de identificar las posibles fases y actividades.
Acuerdos		
Número de tema	Detalle	
2	Se acordó incluir proyectos similares que se relacionan, preferiblemente con la automatización y la minería de datos en las actividades de beneficiado de café o, en su defecto, en otros cultivos.	
Próxima reunión		
Tema por abordar	Fecha	Comentario
Avances de semana 4	14-08-2024	

Apéndice MM Minuta reunión 10

Minuta número	Minuta-10	Fecha	21-08-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	4:15 p. m.
		Hora de finalización	5:15 p. m.
Motivo de reunión	Presentación de avances de semana 5.		
Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón		
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón		
Ausentes			
Temas tratados			
Número de tema	Asunto	Comentarios	
1	Estado del arte	Se presentaron los avances que se relacionan con el estado del arte, siendo los trabajos similares y conceptos relevantes para el estudio.	

2	Metodología	Se presentó el diagrama de las fases del procedimiento metodológico.
3	Sujetos de investigación	Se conversó acerca de un contacto que potencialmente puede funcionar como sujeto de investigación por parte del Icafé.
Acuerdos		
Número de tema	Detalle	
1	Se acordó incluir tendencias de los temas tratados y conceptos adicionales como el de minería de datos y de forma general mencionar otras metodologías.	
2	Se acordó replantear el diagrama de acuerdo con la metodología explicada en el Capítulo 2.	
Próxima reunión		
Tema por abordar	Fecha	Comentario
Avances de semana 6	28-08-2024	

Apéndice NN Minuta reunión 11

Minuta número	Minuta-11	Fecha	1-09-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	3:00 p. m.
		Hora de finalización	4:00 p. m.
Motivo de reunión	Presentación de avances de semana 6.		
Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón		
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón		
Ausentes			
Temas tratados			
Número de tema	Asunto	Comentarios	

1	Metodología	Se presentaron los avances correspondientes al Capítulo 3 y metodología.
2	Muestra de estudio	Se explicó al profesor que no se encontró alguna fuente que señalara la población del estudio, es decir, la cantidad de microbeneficios en Santa María de Dota.
3	Fuentes de información	El profesor explicó la diferencia entre fuentes y sujetos de investigación, los cuales no estaban correctamente diferenciados en el documento.
4	Variables de la investigación	Se presentaron las variables de cada objetivo al profesor.
5	Procedimiento metodológico	Se presentó el procedimiento metodológico
Acuerdos		
Número de tema	Detalle	
2	Se acordó tratar de contactar con personas de Icafé, para que ayuden a identificar la población.	
3	Corregir las fuentes de información	
4	Se acordó tratar de coordinar una reunión con la profesora Lorena para validar las fuentes de información.	
5	Se acordó simplificar el diagrama.	
Próxima reunión		
Tema por abordar	Fecha	Comentario
Avances de semana 7	4-09-2024	

Apéndice ÑÑ Minuta reunión 12

Minuta número	Minuta-12	Fecha	27-09-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	11:00 a. m.
		Hora de finalización	12:00 p. m.
Motivo de reunión	Presentación de avances del Capítulo 4 y discusión del artículo científico.		

Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón	
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón	
Ausentes		
Temas tratados		
Número de tema	Asunto	Comentarios
1	Estado general del proyecto	Se brindaron los avances de la reunión con la persona experta y otros aspectos del proyecto.
2	Resultados	Se brindaron los avances de la fase 1 del proyecto que es la comprensión del negocio.
3	Artículo científico	Se presentó una propuesta de tema, sin embargo, fue descartada debido a su complejidad.
Acuerdos		
Número de tema	Detalle	
2	Se decidió incluir las citas de las fuentes de información directamente en el cuerpo del capítulo, además de mencionarlas en el instrumento de revisión documental presente en los Apéndices.	
3	Se acordó revisar las recomendaciones del profesor sobre la estructuración correcta del artículo. Además, el docente consultará con Coordinación la posibilidad de desarrollar un artículo basado en los resultados completos del TFG	
Próxima reunión		
Tema por abordar	Fecha	Comentario
Por definir	Por definir	

Apéndice OO Minuta reunión 13

Minuta número	Minuta-13	Fecha	10-10-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	5:00 p. m.
		Hora de finalización	6:00 p. m.
Motivo de reunión	Presentación de avances del Capítulo 4 y discusión del artículo científico.		
Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón		
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón		
Ausentes			
Temas tratados			
Número de tema	Asunto	Comentarios	
1	Estado general del proyecto	Se brindaron los avances del Capítulo 4.	
2	Artículo científico	Se discutió acerca de la temática del artículo.	
Acuerdos			
Número de tema	Detalle		
2	Se decidió usar el tercer objetivo específico como insumo para la redacción del artículo.		
Próxima reunión			
Tema por abordar	Fecha	Comentario	
Por definir	Por definir		

Apéndice PP Minuta reunión 14

Minuta número	Minuta-13	Fecha	27-10-2024
Medio de comunicación	Reunión de Google Meet.	Hora de inicio	5:00 p. m.
		Hora de finalización	5:20 p. m.
Motivo de reunión	Presentación de avances del Capítulo 4.		
Personas convocadas	Pedro Leiva Chinchilla Julio Romero Chacón		
Presentes	Pedro Leiva Chinchilla Julio Romero Chacón		
Ausentes			
Temas tratados			
Número de tema	Asunto	Comentarios	
1	Estado general del proyecto	Se brindaron los avances finales del Capítulo 4.	
2	Próximos pasos	Se acordó enviar el documento final al tutor el lunes, para su posterior envío al filólogo y empezar con la redacción del artículo.	
Acuerdos			
Número de tema	Detalle		
Próxima reunión			
Tema por abordar	Fecha	Comentario	
Por definir	Por definir		

Apéndice QQ. Firma de minutas del profesor tutor

Número de reunión	Firma del profesor tutor
El profesor tutor valida la participación en las siguientes minutas: Minuta-1, Minuta-2, Minuta-3, Minuta-4, Minuta-5, Minuta-6, Minuta-7, Minuta-8, Minuta-9, Minuta-10, Minuta-11, Minuta-12, Minuta-13 y Minuta-14	

Apéndice RR solicitud de cambio 1

Solicitud de cambio	
ID de solicitud	1
Fecha de solicitud	7-06-2024
Responsable	Julio Romero Chacón
Prioridad del cambio	Urgencia: • Alta
Fecha de realización	21-06-2024
Descripción de la solicitud	Cambiar el título de la propuesta para orientarse al diseño de un modelo predictivo, en lugar de un prototipo de *software* funcional, lo que implica cambiar objetivos y metodología.
Estado	Realizado

Capítulo 11. Anexos

Carta de aprobación del filólogo

Cartago, 11 de noviembre de 2024

Los suscritos, Elena Redondo Camacho, mayor, casada, filóloga, incorporada a la Asociación Costarricense de Filólogos con el número de carné 0247, portadora de la cédula de identidad número 3-0447-0799 y, Daniel González Monge, mayor, casado, filólogo, incorporado a la Asociación Costarricense de Filólogos con el número de carné 0245, portador de la cédula de identidad número 1-1345-0416, ambos vecinos de Quebradilla de Cartago, revisamos el trabajo final de graduación que se titula: *Diseño de un modelo predictivo para la demanda y planificación de la producción de café en microbeneficios de Santa María de Dota*, sustentado por Julio César Romero Chacón.

Hacemos constar que se corrigieron aspectos de ortografía, redacción, estilo y otros vicios del lenguaje que se pudieron trasladar al texto. A pesar de esto, la originalidad y la validez del contenido son responsabilidad directa de la persona autora.

Esperamos que la participación de Filólogos Bórea Costa Rica satisfaga los requerimientos del Tecnológico de Costa Rica.

X ANA ELENA REDONDO CAMACHO (FIRMA)
Firmado digitalmente por ANA ELENA REDONDO CAMACHO (FIRMA)
Fecha: 2024.11.11 18:02:42 -06'00'

Elena Redondo Camacho
Filóloga - Carné ACFIL n.º 0247

X DANIEL ALBERTO GONZALEZ MONGE (FIRMA)
Firmado digitalmente por DANIEL ALBERTO GONZALEZ MONGE (FIRMA)
Fecha: 2024.11.11 18:03:04 -06'00'

Daniel González Monge
Filólogo - Carné ACFIL n.º 0245

Desarrollo de un modelo predictivo de demanda y producción para microbeneficios de café en Santa María de Dota

Un enfoque basado en la selección de técnicas predictivas

Julio Romero Chacón*
Tecnológico de Costa Rica
Cartago, Cartago, Costa Rica
julioromero@estudiantec.cr

Pedro Leiva Chichilla*
Tecnológico de Costa Rica
Cartago, Cartago, Costa Rica
peleiva@itcr.ac.cr

RESUMEN

La producción y demanda de café en los microbeneficios de Santa María de Dota, Costa Rica, es vulnerable a variaciones en el mercado y a condiciones externas, lo que genera incertidumbre al planificar la producción. Esto puede ocasionar pérdidas financieras debido a la acumulación de excedentes o, por el contrario, a oportunidades de venta desaprovechadas. Para abordar este problema, se desarrolló un modelo predictivo basado en la metodología CRISP-DM, que incluye las fases de análisis, preparación de datos, modelado y evaluación. El objetivo es reducir la incertidumbre y mejorar la toma de decisiones en los microbeneficios, promoviendo prácticas agrícolas sostenibles.

El estudio evaluó diversas técnicas de modelado, como redes neuronales, *random forest*, *gradient boosting*, *k-nearest neighbors* (KNN) y Arimax, para identificar el modelo más eficaz en términos de precisión y capacidad explicativa, medidas mediante el error absoluto medio (MAE) y el coeficiente de determinación (R^2). A través de iteraciones, se optimizó el rendimiento de los modelos y se identificaron los factores más relevantes para la demanda de café. Los resultados preliminares indican que el modelo de redes neuronales, con tres neuronas en la capa profunda, es capaz de capturar una alta variabilidad en los datos de demanda, lo que proporciona predicciones precisas y confiables para la planificación en los microbeneficios.

KEYWORDS

Modelo predictivo, CRISP-DM, café.

ACM Reference format:

Julio Romero Chacón y Pedro Leiva Chinchilla. 2024. "Desarrollo de un modelo predictivo de demanda y producción para microbeneficios de café en Santa María de Dota: un enfoque basado en la selección de técnicas predictivas". En *Proceedings of ACM Woodstock conference* (WOODSTOCK'18). ACM, New York, NY, USA, 9 pages. Disponible en: https://doi.org/10.1145/nnnnnnn.nnnnnnn

1. INTRODUCCIÓN

La producción y demanda de café en microbeneficios constituye una actividad altamente susceptible de variaciones en el mercado y a condiciones externas, las cuales afectan la estabilidad económica de los beneficiadores en Santa María de Dota, Costa Rica. Sin una herramienta adecuada basada en datos, los microbeneficios enfrentan incertidumbre al planificar su producción, lo que puede resultar en una producción excesiva o insuficiente. Este problema provoca, por un lado, pérdidas financieras debido a la acumulación de excedentes y, por otro, oportunidades de venta desaprovechadas.

Para abordar este problema, se diseñó un modelo predictivo que actúa como una herramienta basada en datos para la planificación de la producción y demanda de café. Para esto, se ha seguido la metodología CRISP-DM, estructurada en fases de análisis, preparación de datos, modelado y evaluación. Se espera que el modelo predictivo resultante no solo reduzca la incertidumbre y mejore la toma de decisiones al planificar la producción de café, sino que también contribuya a prácticas agrícolas más sostenibles.

El presente estudio evalúa diversas técnicas de modelado, tales como redes neuronales, *random forest*, *gradient boosting*, *k-nearest neighbors* (KNN) y Arimax, con el objetivo de identificar el modelo más efectivo en términos de error absoluto medio (MAE) y coeficiente de determinación (R^2). La pregunta de investigación se centra en determinar cuál de estos enfoques es capaz de proporcionar predicciones útiles y fiables, con base en los datos históricos y en variables exógenas o predictoras relevantes.

A través de una serie de iteraciones y configuraciones en los modelos, se busca no solo mejorar la precisión de las predicciones, sino también evaluar la estabilidad y robustez de cada método. Al final de este proceso, se espera identificar el modelo que mejor satisfaga los requisitos de exactitud y explicabilidad, lo que facilita una planificación informada y sostenible en el contexto específico de los microbeneficios del café.

Por lo tanto, se parte de la siguiente hipótesis de investigación: "La aplicación de técnicas de minería de datos multifactoriales permite captar patrones complejos en la demanda de café en

microbeneficios, lo que proporciona una herramienta predictiva precisa y robusta para planificar la producción".

Los resultados preliminares indican que el modelo de redes neuronales, con tres neuronas en la capa profunda es capaz de explicar una parte de la variabilidad en los datos de demanda, alcanzando un nivel de precisión suficiente para apoyar decisiones clave en los microbeneficios. Al mismo tiempo, identifica factores relevantes como la producción nacional y las precipitaciones.

2. TRABAJOS RELACIONADOS

En los últimos años, la predicción de la demanda y la planificación de la producción agrícola han ganado relevancia debido a su impacto directo en la optimización de los recursos y la reducción de riesgos en la industria. Varios estudios previos han explorado técnicas de modelado predictivo aplicadas a diferentes cultivos, lo que ofrece herramientas valiosas para el análisis del comportamiento de los mercados agrícolas.

Garzón [1] realizó un estudio sobre la predicción de la oferta de aguacate Hass en Herveo, Tolima, utilizando modelos de series temporales, como promedios móviles y suavizamiento exponencial. Estas técnicas permitieron identificar patrones en los datos históricos y proyectar el comportamiento futuro de la producción. El estudio concluyó que la aplicación de estos métodos mejora la toma de decisiones agrícolas y reduce los riesgos asociados con la variabilidad en la oferta de productos agrícolas.

Por otro lado, Suhardi y colaboradores [2] analizaron la demanda de café tostado mediante técnicas como medias móviles, medias móviles ponderadas y suavizamiento exponencial y mostraron que estos métodos pueden prever la demanda de café con alta precisión. En otro estudio relacionado, Vijayan *et al.* [3] aplicaron el modelo Arima para predecir la producción de café arábica y robusta en la India, utilizando el lenguaje de programación R para realizar sus predicciones. Este enfoque facilitó el análisis de tendencias futuras en la producción de café y subrayó el valor de los modelos con base en series temporales para la planificación en la industria cafetalera.

Tolentino y Hernández [4], en su investigación en Filipinas, compararon diferentes técnicas predictivas, como el suavizamiento exponencial, la media móvil y la regresión, en la producción de café. Los resultados indicaron que la media móvil era el método más preciso, ya que presentaba la tasa de error más baja en las predicciones, lo que evidenció su utilidad en contextos con demanda fluctuante.

Además, se han explorado técnicas avanzadas como las redes neuronales artificiales (RNA) y la regresión lineal múltiple (MLR) para predecir el rendimiento del café arábica. Kittichotsatsawat *et al.* [5] lograron una alta precisión en sus predicciones ($R^2 = 0{,}9524$) al utilizar redes neuronales artificiales (RNA) con datos recolectados sobre variables como el área de cultivo, las precipitaciones y la temperatura, lo cual demostró la capacidad de estos modelos para capturar relaciones complejas y no lineales en los datos.

Finalmente, Khumaidi [6] empleó la metodología CRISP-DM junto con la regresión lineal múltiple para desarrollar un modelo predictivo en la producción de café. De esta forma, evaluó la calidad del modelo mediante la raíz del error cuadrático medio (RMSE) y demostró la viabilidad de esta metodología en contextos agrícolas.

El valor agregado de este estudio frente a las investigaciones previas radica en la inclusión de una mayor cantidad de variables en el entrenamiento de los modelos predictivos. Por otro lado, la mayoría de los estudios existentes se enfoca en modelos básicos de series de tiempo, específicamente en modelos univariables que solo capturan tendencias en una única serie temporal, el presente trabajo explora modelos multivariables capaces de captar las variaciones específicas de cada microbeneficio de café, así como de incorporar otras variables de entorno y de mercado.

Además de los modelos de series de tiempo, este estudio evalúa técnicas avanzadas de modelado, como redes neuronales y *random forest*. Estas pueden detectar patrones complejos y no lineales en los datos, lo que ofrece potencialmente una mayor precisión en la predicción de la demanda.

Además, esta investigación utiliza la metodología CRISP-DM, que proporciona un marco estructurado para proyectos de minería de datos. Esto asegura que el proceso sea robusto y que los datos que llegan a la etapa de modelado sean de alta calidad y útiles para maximizar el rendimiento de los modelos. Asimismo, se adopta un enfoque iterativo para ajustar los parámetros de los modelos. Lo anterior tiene el fin de identificar aquellos que ofrecen los mejores resultados en términos de precisión y capacidad explicativa.

3. METODOLOGÍA

La investigación utiliza la metodología CRISP-DM (CRoss Industry Standard Process for Data Mining), la cual estructura el desarrollo en cinco fases para construir un modelo predictivo eficiente. A continuación, se detallan las etapas que se implementan en el proyecto; nótese que la fase de despliegue no se desarrolló:

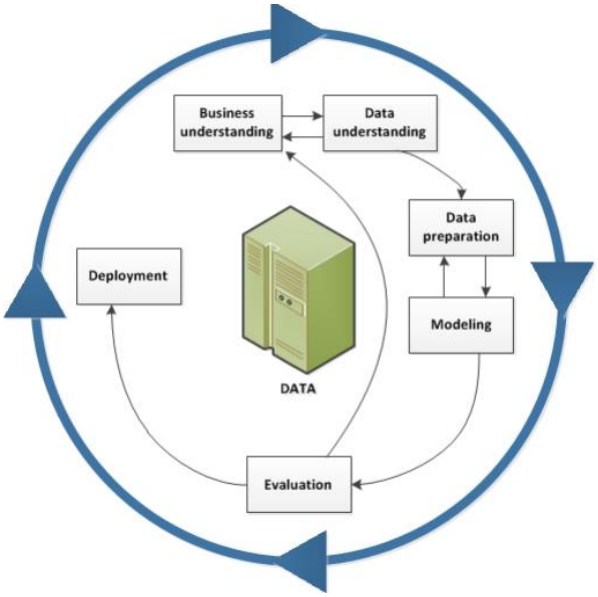


Figura 1**:** fases de CRISP-DM

3.1 Entendimiento del negocio

La primera fase consistió en comprender el negocio del microbeneficiado de café en Santa María de Dota. Se establecieron los objetivos empresariales para optimizar cómo se planifica la producción y reducir la incertidumbre en la demanda. Esto incluyó la definición de factores clave que afectan el negocio y la precisión de fuentes de datos relevantes para la construcción del modelo.

Cabe resaltar que los sujetos de investigación en este estudio fueron los microbeneficios de café en la región de Santa María de Dota, Costa Rica, los cuales se caracterizan por procesar volúmenes inferiores a 1,000 fanegas de café anualmente.

3.2 Comprensión de los datos

En esta fase se llevó a cabo la recolección y exploración inicial de los datos. El conjunto de datos incluyó variables históricas propias y del entorno que son relevantes para la predicción de la producción y la demanda de café. Además, se evaluó la calidad de los datos y se detectaron valores atípicos y faltantes.

3.3 Preparación de los datos

Los datos recolectados se sometieron a un proceso de limpieza y transformación. Se trataron los valores inconsistentes y se normalizaron las variables numéricas. Además, se desarrolló un procedimiento de codificación *one-hot* para transformar las variables categóricas en datos numéricos manejables. Finalmente, se integraron los datos en un único repositorio para su uso en el modelado predictivo.

3.4 Modelado predictivo

La fase de modelado predictivo se centró en construir modelos capaces de prever la demanda de café en microbeneficios mediante un enfoque de análisis y evaluación. El proceso se estructuró en dos pasos principales: la evaluación de técnicas potenciales y la configuración del modelo a través de tres iteraciones.

3.4.1. Evaluación de técnicas de modelado potenciales

El primer paso consistió en investigar y seleccionar técnicas de modelado que se pudieran adaptar a los datos disponibles y a los objetivos específicos de predicción de demanda. Esto incluyó la revisión de modelos, tanto de series de tiempo como multifactoriales, que permitieran incorporar múltiples variables relevantes para la producción de café. Se definieron ciertos criterios iniciales de selección, tales como la capacidad del modelo para manejar datos no lineales, su flexibilidad para incluir variables exógenas y la necesidad de una configuración que minimizara el sobreajuste, debido al carácter variado, pero limitado de los datos históricos.

3.4.2. Configuración y desarrollo del modelo.

En esta etapa, cada modelo potencial fue entrenado y configurado de acuerdo con las características específicas de los datos. Esto incluyó un proceso de ajuste iterativo en el que se probaron diferentes configuraciones de parámetros para optimizar el rendimiento. El entrenamiento de cada modelo se realizó utilizando un conjunto de datos segmentado en subconjuntos de entrenamiento y prueba, de manera que el modelo pudiera aprender patrones sin depender excesivamente de un grupo particular de datos.

Además, se empleó la validación cruzada para evaluar la generalización de los modelos, asegurando que su rendimiento se mantuviera estable independientemente de los datos específicos con los que se entrenan y prueban en cada iteración.

3.5 Evaluación

El desempeño de cada modelo se evaluó a través de métricas de error y precisión establecidas antes del proceso de modelado, tales como el error absoluto medio (MAE) y el coeficiente de determinación (R^2). Estas métricas permitieron medir de manera cuantitativa la exactitud de las predicciones y proporcionaron un marco objetivo para comparar el rendimiento de los diferentes modelos.

La selección final del modelo se basó en su capacidad para minimizar el error de predicción y maximizar la capacidad explicativa según el criterio de R^2, asegurando que el modelo

seleccionado ofreciera no solo precisión, sino también robustez y estabilidad para su aplicación en los microbeneficios del café.

4. CONCEPTOS

A continuación, se pretende brindar una contextualización de los modelos multifactoriales por evaluar y la importancia del uso de estos.

4.1 Redes neuronales

Las redes neuronales se inspiran en la estructura del cerebro humano, donde las neuronas están conectadas y trabajan en conjunto para procesar información. En el contexto de la predicción, las redes neuronales aprenden patrones complejos al ajustar sus conexiones internas (pesos) en función de los datos que reciben. Este proceso permite que las redes neuronales manejen relaciones no lineales entre las variables, lo cual resulta especialmente útil para problemas en los que los datos presentan interacciones complejas y múltiples influencias [7].

4.2 *Random forest*

Random forest es un modelo que combina múltiples árboles de decisión para mejorar la exactitud de las predicciones. Un árbol de resolución toma decisiones dividiendo los datos en *ramas* en función de ciertas reglas, hasta llegar a una *hoja* que representa una predicción. En lugar de utilizar un solo árbol, *random forest* genera varios árboles a partir de diferentes muestras del conjunto de datos y luego combina sus resultados.

Esto hace que el modelo sea más robusto, ya que reduce el riesgo de que un solo árbol influya demasiado en el resultado y es resistente al sobreajuste. La combinación de árboles ayuda a capturar mejor las relaciones entre múltiples variables. Sin embargo, debido a que *random forest* genera varios árboles, puede requerir un alto poder de cómputo y ser más lento al trabajar con grandes volúmenes de datos [8].

4.3 Gradient boosting

Gradient boosting es otro modelo que utiliza múltiples árboles de decisión, sin embargo, a diferencia de *random forest*, los árboles se construyen de forma secuencial. En este caso, cada árbol intenta corregir los errores cometidos por el árbol anterior, *aprendiendo* de sus fallos. Este proceso hace que *gradient boosting* sea muy preciso en la captura de patrones complejos en los datos, especialmente en relaciones no lineales.

Sin embargo, al agregar más árboles para mejorar la precisión, el modelo puede volverse propenso al sobreajuste, lo que significa que puede adaptarse demasiado a los datos de entrenamiento y no generalizar bien a datos nuevos. Por esto, es importante controlar cuidadosamente los parámetros del modelo para evitar este problema [9].

4.4 K-nearest neighbors (KNN).

KNN es un modelo basado en la proximidad o cercanía entre puntos de datos. Cuando se necesita realizar una predicción, KNN busca los *k* puntos de datos más cercanos al nuevo dato y utiliza sus valores para efectuar la predicción. Por ejemplo, si se está prediciendo la demanda de café y se encuentra que en años anteriores, en condiciones similares, la demanda fue alta, KNN empleará esos datos para llevar a cabo una predicción similar. Este modelo es fácil de entender e implementar, pero puede volverse ineficiente y lento al trabajar con grandes volúmenes de datos, ya que debe calcular la distancia entre el nuevo punto y todos los datos históricos. Además, KNN funciona mejor cuando las relaciones entre variables no son muy complejas y los datos están bien distribuidos [10].

4.5 Arimax

Arima es un modelo de predicción que se basa en tres componentes principales [11]:

- La autorregresión (AR) utiliza los valores pasados de la serie para predecir el futuro, asumiendo que existe una relación entre los datos anteriores y los futuros.
- Integración (I): ayuda a estabilizar la serie al eliminar tendencias, de manera que el modelo pueda trabajar con datos *más planos* o estacionarios.
- Media móvil (MA): captura la relación entre el valor actual y los errores de predicciones anteriores, ajustando los pronósticos en función de las diferencias entre las predicciones y los datos reales pasados.

En conjunto, estos tres componentes permiten que Arima utilice patrones históricos para realizar predicciones futuras. Pero, se limita a considerar únicamente los datos de esta serie de tiempo.

Por este motivo, se plantea el uso de Arimax, el cual extiende Arima al permitir la inclusión de *variables exógenas*. Es decir, factores externos que pueden afectar la serie de tiempo, como el clima o los precios del café en el mercado. Esto significa que, además de los componentes AR, I y MA, Arimax permite al modelo aprender cómo estos factores externos influyen en el resultado, lo que hace que la predicción sea más completa y esté adaptada a situaciones en las que no solo los datos históricos propios, sino también las condiciones externas, desempeñan un papel importante.

5. RESULTADOS

Para construir el conjunto de datos necesarios para desarrollar el modelo predictivo, se identificaron 10 factores relevantes que potencialmente pueden tener un impacto en la determinación de la

producción y la demanda de café. A continuación, se describen los factores importantes.

Tabla 1. Factores relevantes

Factor relevante	Descripción
Cantidad de café beneficiado	La variable objetivo, que representa la cantidad total procesada anualmente por cada beneficio.
Año	Variable temporal que permite analizar tendencias históricas a lo largo de diferentes periodos.
Microbeneficio	Identificador de cada microbeneficio, lo cual permite capturar la variabilidad entre distintas entidades en el análisis. Esta variable se codificó mediante *one-hot encoding* para su integración adecuada en el modelo.
Temperatura media	Refleja las condiciones climáticas locales promedio en grados Celsius.
Precipitaciones	Medida en milímetros, este factor representa el total de lluvia en cada periodo.
Tipo de cambio	Representa el tipo de cambio con respecto al dólar a partir del cual cada microbeneficio logró comercializar su café en cada año.
Crecimiento económico mundial	Expresado en porcentaje, refleja la situación económica global, influyendo en la demanda en el ámbito internacional.
Producción nacional	Muestra el volumen de producción en el ámbito nacional, lo que proporciona contexto sobre la oferta disponible en el mercado interno.
Oferta mundial	Cantidad de café disponible en el mercado internacional en un periodo.
Precio de café según la Bolsa de Nueva York	Valor de referencia para el café en el mercado de futuros y representa la especulación del futuro del producto y el valor percibido internacionalmente en un momento dado.
Precio de liquidación	Refleja el precio promedio por el cual cada microbeneficio logró comercializar su café.

Todos los factores considerados en el análisis provienen de fuentes confiables y se recolectaron mediante revisión documental (ver el Apéndice 2), consolidándose en un único archivo de Microsoft Excel que sirve como insumo para entrenar los diversos modelos (ver el Apéndice 1).

Además, los datos se encuentran normalizados en una escala de 1 a 10 y se expresan en una misma unidad de medida, si corresponde, para favorecer la comparabilidad entre ellos.

Por otra parte, el objetivo consiste en diseñar un modelo predictivo que estime la cantidad de café a procesar en futuros periodos, lo que optimiza el planeamiento de la producción y reduciendo la incertidumbre en la demanda del mercado.

Para lograr esto, se establecieron los siguientes objetivos de minería de datos. Es decir, criterios técnicos que ayudan a determinar la efectividad de cada modelo.

- Lograr un coeficiente de determinación del 80 %

Significa que el modelo debe ser capaz de explicar el 80 % de la variabilidad observada en los datos. Un R^2 del 80 % indica que el modelo posee un buen poder predictivo, capturando una parte significativa de las relaciones en los datos y dejando solo un 20 % de la variabilidad sin explicar.

- Obtener un MAPE (error porcentual absoluto medio) máximo del 10 %

Este objetivo establece que el error promedio, expresado como un porcentaje de los valores reales, debe mantenerse por debajo del 10 %. Es decir, las predicciones pueden diferir en promedio hasta un 10 % de los valores reales, sin importar si se encuentran por encima o por debajo de estos.

- Identificar al menos dos factores que sean importantes en el modelo.

Este objetivo se enfoca en identificar al menos dos factores que tienen un impacto significativo en las predicciones. Esto ayuda a entender qué variables son clave para el modelo y cómo influyen en los resultados.

Una vez que se conocen los resultados que se pretende alcanzar, se realiza cada una de las tres iteraciones, utilizando el lenguaje de programación R (ver el Apéndice 3).

5.1 Redes neuronales

Tabla 2. Resultados de redes neuronales

Iteración	Parámetros	MAPE	R2
1	Linout=TRUE, Size=1	18.24	88.57
2	Linout=TRUE, Size=5, Decay=0.1	5.69	99.14
3	Linout=TRUE, Size=3, Decay=0.2	12.39	95.31

En el modelo de redes neuronales, se configuró para realizar regresión en lugar de clasificación. En la primera iteración, se utilizó solo una neurona en la capa oculta, lo que aumenta a cinco en la segunda y luego a tres en la tercera, con el objetivo de capturar patrones más complejos. Además, se aplicó una regularización (*decay*) de 0.1 y, posteriormente, de 0.2, para evitar que el modelo se ajustara demasiado a los datos de entrenamiento y lograra una mejor generalización.

5.2 Random forest

Tabla 3. Resultados de random forest

Iteración	Parámetros	MAPE	R2
1	Ntree = 500	13.96	93.64
2	Ntree = 100	13.99	93.71
3	Ntree = 50	14.29	93.93

Para *random forest*, se definió el número de árboles mediante el parámetro Ntree. En la primera iteración, se utilizaron 500 árboles, disminuyéndolo a 100 y luego a 50 en las iteraciones siguientes, lo cual buscó equilibrar la precisión y el tiempo de procesamiento.

5.3 Gradient boosting

Tabla 4. Resultados de *gradient boosting*

Iteración	Parámetros	MAPE	R2
1	n.trees=50, interaction.depth=1, shrinkage=0.1	54.84	19.95
2	n.trees=200, interaction.depth=3, shrinkage=0.05	51.72	33.42
3	n.trees=300, interaction.depth=3, shrinkage=0.01	57.83	20.50

En *gradient boosting*, el número de árboles comenzó en 50, luego aumentó a 200 y, posteriormente, a 300, con el objetivo de mejorar la precisión. La profundidad de los árboles se incrementó a 3 en las iteraciones posteriores para capturar interacciones más complejas entre las variables.

La tasa de aprendizaje (*shrinkage*) se ajustó de 0.1 a valores más bajos (0.05 y 0.01) para que el modelo aprendiera de forma controlada y así evitar el sobreajuste.

5.4 KNN

Tabla 5. Resultados de KNN

Iteración	Parámetros	MAPE	R2
1	k = 23	71.56	4.03
2	k = 2	46.51	46.59
3	k = 8	59.86	12.49

Para *k-nearest neighbors* (KNN), el parámetro k comenzó en 23 vecinos para realizar una predicción más general, luego disminuyó a 2 en la segunda iteración y aumentó a 8 en la tercera, ajustando el grado de detalle en las predicciones.

5.5 Arimax

Tabla 6. Resultados de Arimax

Iteración	Parámetros	MAPE	R2
1	Arima(0,0,0)	21.39	87.1

En este caso, el modelo seleccionó de forma automática una configuración Arima (0,0,0), lo que indica que no se identificó una estructura autorregresiva ni de media móvil significativa en los datos. Esto significa que el modelo no incorpora términos de autorregresión, diferenciación o media móvil, sino que depende únicamente de las variables externas para realizar las predicciones. Por lo tanto, se optó por llevar a cabo únicamente una iteración del modelo.

6. DISCUSIÓN

A continuación, se interpretan los resultados en cada una de las tres iteraciones, contrastando los cambios realizados en los modelos y cómo estos ajustes impactaron en su rendimiento.

- Iteración 1

En la primera iteración, el modelo de *random forest* obtuvo el mejor rendimiento, con un MAPE del 13.96 % y un R^2 del 93.64 %. Esto indica que el modelo logró captar adecuadamente las relaciones entre las variables. Las redes neuronales también mostraron un buen rendimiento, con un MAPE de 18.24 % y un R^2 de 88.57 %, lo que sugiere una capacidad aceptable para capturar la relación no lineal entre las variables relevantes, como la oferta mundial y el año.

Por otro lado, *gradient boosting* y *k-nearest neighbors* (KNN) presentaron un desempeño significativamente inferior. *gradient boosting* tuvo un MAPE alto (54.84 %) y un R^2 bajo (19.95 %), lo que indica que no capturó adecuadamente las relaciones entre las variables. KNN también mostró un rendimiento deficiente, con un MAPE de 71.56 % y un R^2 de 4.03 %, lo que sugiere que este modelo no es adecuado para este tipo de predicción debido a la complejidad de las relaciones no lineales en los datos.

- Iteración 2

En la segunda iteración, se optimizaron algunos parámetros, lo que condujo a una mejora notable en el rendimiento de las redes neuronales. Con un MAPE de 5.69 % y un R^2 de 99.14 %, este modelo alcanzó el mejor rendimiento en esta iteración. El incremento en el número de neuronas en la capa oculta a 5 y un ajuste de regularización (*decay*) mejoraron la capacidad del modelo para captar patrones complejos. Las variables más importantes en esta iteración fueron el crecimiento económico mundial y el precio en la Bolsa de Nueva York, lo cual puede indicar una mayor sensibilidad a variables macroeconómicas.

Random forest mantuvo un rendimiento estable en esta iteración, con un MAPE del 13.99 % y un R^2 de 93.71 %, lo que demuestra robustez y consistencia en la captura de las relaciones entre las variables.

Gradient boosting mostró una leve mejora, alcanzando un MAPE de 51.72 % y un R^2 de 33.42 % tras aumentar el número de árboles y la profundidad de las interacciones. Sin embargo, el modelo sigue sin captar suficientemente bien las relaciones en los datos. KNN también mejoró ligeramente, con un MAPE de 46.51 % y un R^2 de 46.59 %, al reducir el número de vecinos a 2. Pero, aún se encuentra lejos de ser un modelo competitivo.

- Iteración 3

En la tercera iteración, se realizaron ajustes adicionales para reducir el riesgo de sobreajuste, en particular en las redes neuronales. A pesar de una leve disminución en el rendimiento predictivo, con un MAPE de 12.39 % y un R^2 de 95.31 %, el modelo mantuvo una buena capacidad predictiva, lo que indica que los ajustes de regularización y la reducción en el número de neuronas (size = 3) ayudaron a mejorar su capacidad de generalización.

Random forest mostró un leve cambio en el rendimiento, con una ligera reducción en el MAPE (14.29 %) y un R^2 similar al de iteraciones anteriores (93.93 %), lo cual confirma su estabilidad y robustez. Por otro lado, *gradient boosting* experimentó una disminución en el rendimiento, con un MAPE de 57.83 % y un R^2 de 20.50 %, lo que sugiere que los ajustes realizados en esta iteración no contribuyeron a mejorar su capacidad predictiva. En cuanto a KNN, el aumento en el número de vecinos (k = 8) ocasionó un rendimiento inferior, con un MAPE de 59.86 % y un R^2 de 12.49 %, lo que confirma que este modelo no es adecuado para este contexto.

A lo largo de las tres iteraciones, el modelo de redes neuronales demostró ser el más adecuado para el contexto de predicción de la demanda de café en microbeneficios, alcanzando el rendimiento más alto en términos de MAPE y R^2 en la segunda iteración. Sin embargo, para la selección del modelo, se optó por el modelo de la tercera iteración, debido a las medidas adicionales contra el sobreajuste, que incrementaron su robustez y capacidad de generalización, a pesar de una ligera reducción en la precisión.

7. CONCLUSIONES Y LIMITACIONES

A pesar de los avances logrados en el desarrollo del modelo predictivo, existen algunas limitaciones que pueden afectar su desempeño y aplicabilidad. En primer lugar, aunque el modelo se construyó considerando los factores disponibles y críticos para el sector cafetalero, es posible que algunos elementos relevantes no se hayan identificado. La complejidad del sector, así como las constantes fluctuaciones en los mercados nacionales e internacionales, dificultan la captura de todos los factores que pueden influir en la demanda de café en los microbeneficios de la región.

Además, un modelo predictivo eficiente depende de un volumen considerable de datos históricos para su entrenamiento adecuado. No obstante, la naturaleza anual de los datos disponibles y la implementación relativamente reciente de los microbeneficios en Santa María de Dota limitaron la cantidad de registros históricos, lo que puede afectar la precisión y robustez del modelo en su aplicación práctica.

Finalmente, la metodología CRISP-DM, caracterizada por su enfoque iterativo, representó un reto considerable debido a las restricciones de tiempo en el desarrollo del proyecto. Aunque este enfoque requiere ajustes continuos para mejorar el modelo, en este caso se utilizaron las mismas variables en cada iteración sin modificar el conjunto de variables que se seleccionaron inicialmente. No obstante, se aseguró que no existiera multicolinealidad en los datos mediante un análisis de correlación

previo, que permitió verificar que las variables que se seleccionaron fueran independientes entre sí.

Finalmente, se extraen conclusiones relevantes del análisis que permiten responder afirmativamente a la hipótesis que se planteó en la introducción: "La aplicación de técnicas de minería de datos multifactoriales permite captar patrones complejos en la demanda de café en microbeneficios, lo que proporciona una herramienta predictiva precisa y robusta para planificar la producción".

A lo largo de las tres iteraciones de pruebas, se observó que, aunque cada técnica evaluada tiene sus propias ventajas, el modelo basado en redes neuronales se destacó por su capacidad para reconocer patrones complejos y relaciones entre múltiples factores. Este modelo logró explicar más del 80 % de las variaciones en los datos, lo cual indica que es muy eficaz para representar cómo ciertos factores afectan la demanda de café.

En términos de precisión, el modelo seleccionado alcanzó un error promedio del 12.39 %, lo que significa que, en promedio, las predicciones del modelo estuvieron muy cerca de los valores reales, aunque un poco por encima del objetivo inicial de un 10 % de error. Aun así, se considera que el modelo es suficientemente preciso y confiable para ser útil al planificar la producción de café. Además, este identificó que la producción nacional y las precipitaciones son factores clave que influyen directamente en la producción de café.

En conclusión, el modelo predictivo seleccionado demuestra que el uso de técnicas avanzadas de análisis de datos es potencialmente efectivo y confiable para planificar la producción en el sector cafetalero. Esto ayuda a reducir la incertidumbre y a mejorar el uso de recursos en prácticas agrícolas más sostenibles.

REFERENCIAS

[1] Garzón, David. *Diseño de un modelo predictivo de la oferta de aguacate en el municipio de Herveo Tolima*. Universidad Santo Tomás de Aquino.

[2] Suhardi, A, Amalia, S, Oktafien, S, Adiyanti, Komariah, S, and Rohendra, T. Time Series Analysis to Predicting Demand of Roasted Coffee. *International Journal of Financial Research*, 10 (2019), 26-31.

[3] Vijayan, K, Vennila, J, Sebastian, L, and Rita, S. Predictive Modelling for Coffee Production Using R Programming. In *2022 3rd International Conference on Communication, Computing and Industry 4.0* (2022), IEEE, 1-6.

[4] Tolentino, R and Hernandez, A. Assessment of Predictive Models for Coffee Production in the Philippines. In *2018 16th International Conference on ICT and Knowledge Engineering (ICT&KE)* (2018), IEEE, 1-6.

[5] Kittichotsatsawat, Y, Tippayawong, N, and Tippayawong, K. Prediction of arabica coffee production using artificial neural network and multiple linear regression techniques. *Scientific Reports* (2022).

[6] Khumaidi, A. Data *mining* for predicting the amount of coffee production using CRISP-DM method. *Jurnal Techno Nusa Mandiri*, 17 (2020), 1-8.

[7] Rueda, Karen and Cardozo, Shirley. Aplicación de redes neuronales artificiales para el pronóstico de precios de café. *REVISTA COLOMBIANA DE TECNOLOGIAS DE AVANZADA (RCTA)*, 1 (2022), 113-117.

[8] Rollón, Álvaro. *Experimentos computacionales en un estudio de simulación de modelos de regresión para una mejor comprensión de las herramientas Random Forests y Conditional Trees*. Industriales. 2016.

[9] Andrade, Vinicio and Flores, Pablo. Comparativa entre classification trees, *random forest* y *gradient boosting*; en la predicción de la satisfacción laboral en Ecuador. *Ciencia Digital*, 2 (2018), 42-54.

[10] Gallego, Antonio, Rico, Juan, and Valero, Jose. Efficient -nearest neighbor *search* based *on clustering* and adaptive values. *Pattern recognition*, 122 (2022).

[11] Yadav, Dinesh Kumar and Soumya, K and Goswami, Laxmi. Autoregressive Integrated Moving Average Model for Time Series Analysis. (2024), IEEEE, 1-6.

[12] ICAFÉ. *20 de Setiembre de 2024 - Café informado - Cosecha 23-24*. ICAFÉ, 2024.

[13] ICAFÉ. *18 de Setiembre de 2023 - Café informado - Cosecha 22-23*. ICAFÉ, 2923.

[14] ICAFÉ. *30 de Junio de 2022 - Cafe informado cosecha 21-22*. ICAFÉ, 2022.

[15] ICAFÉ. *02 de julio de 2021 Cafe informado - Cosecha 20-21*. ICAFÉ, 2021.

[16] ICAFÉ. *30 de julio de 2020 Cafe informado - Cosecha 19-20*. ICAFÉ, 2020.

[17] ICAFÉ. *25 de julio de 2019 Cafe informado - Cosecha 18-19*. ICAFÉ, 2019.

[18] ICAFÉ. *15 de agosto 2018 Cafe informado - Cosecha 17-18*. ICAFÉ, 2018.

[19] ICAFÉ. *24 de agosto de 2017 Cafe informado - Cosecha 16-17*. ICAFÉ, 2017.

[20] Molina, Marco Antonio Araya. *Costos de beneficiado de café aceptados por ley no 2762 cosecha 2022-2023*. Instituto del café de Costa Rica, San José, 2023.

[21] Molina, Marco Antonio Araya. *Costos de beneficiado de café aceptados por ley no 2762 cosecha 2021-2022*. Instituto del café de Costa Rica (ICAFÉ), San José, 2022.

[22] Molina, Marco Antonio Araya. *Costos de beneficiado de café aceptados por ley no 2762 cosecha 2020-2021*. Instituto del café de Costa Rica (ICAFÉ), San José, 2021.

[23] Molina, Marco Antonio Araya. *Costos de beneficiado de café aceptados por ley no 2762 cosecha 2019-2020*. Instituto del café de Costa Rica (ICAFÉ), San José, 2020.

[24] METEOBLUE. *Cambio climático Cerros de Dota*. 2024.

[25] ICAFÉ. *Precio de liquidación final cosecha 2022-2023*. ICAFÉ, San José, 2023.

[26] ICAFÉ. *Precio de liquidación final cosecha 2021-2022*. ICAFÉ, San José, 2022.

[27] ICAFÉ. *Precio de liquidación final cosecha 2020-2021*. ICAFÉ, San José, 2021.

[28] ICAFÉ. *Precio de liquidación final cosecha 2019-2020*. ICAFÉ, San José, 2020.

[29] ICAFÉ. *Precio de liquidación final cosecha 2018-2019*. ICAFÉ, San José, 2019.

[30] ICAFÉ. *Precio de liquidación final cosecha 2017-2018*. ICAFÉ, San José, 2018.

[31] ICAFÉ. *Precio de liquidación final cosecha 2016-2017*. ICAFÉ, San José, 2017.

[32] GRUPO BANCO MUNDIAL. *Crecimiento del PIB (% anual)*. World Bank Open Data. 2023.

[33] ICAFÉ. *Informe Actividad Cafetalera de Costa Rica 2023*. Heredia, 2023.

[34] ICAFÉ. *Informe sobre la actividad cafetalera de Costa Rica 2021*. ICAFÉ, San José, 2021.

[35] ICAFÉ. *Precios Históricos del Café en Nueva York – 2024*. ICAFÉ, San José, 2024.

[36] ICAFÉ. *Precios Históricos del Café en Nueva York - 2023*. ICAFÉ, San José, 2024.

[37] ICAFÉ. *Precios Históricos del Café en Nueva York - 2022*. ICAFÉ, San José, 2023.

[38] ICAFÉ. *Precios Históricos del Café en Nueva York - 2021*. ICAFÉ, San José, 2022.

[39] ICAFÉ. *Precios Históricos del Café en Nueva York - 2020*. ICAFÉ, San José, 2021.

[40] ICAFÉ. *Precios Históricos del Café en Nueva York – 2019*. ICAFÉ, 2019.

[41] ICAFÉ. *Precios Históricos del Café en Nueva York – 2018*. ICAFÉ, 2018.

[42] ICAFÉ. *Precios Históricos del Café en Nueva York – 2017*. ICAFÉ, 2017.

APÉNDICES

1. Datos consolidados

Datos consolidados

2. Documento de R

Documento de R

3. Fuentes de información de los factores relevante

Tabla 6. Fuentes de información de los datos

Factor	**Fuente de información del factor relevante**
Cantidad de café beneficiado (factor a predecir)	20 de septiembre De 2024 Café Informado Cosecha 23-24 [12]
	18 de septiembre de 2023-Café informado-Cosecha 22-23 [13]
	30 de junio de 2022-Café informado cosecha 21-22 [14]
	02 de julio de 2021 Café informado-Cosecha 20-21 [15]
	30 de julio de 2020 Café informado-Cosecha 19-20 [16]
	25 de julio de 2019 Café informado-Cosecha 18-19 [17]
	15 de agosto de 2018 Café informado-Cosecha 17-18 [18]
	24 de agosto de 2017 Café informado-Cosecha 16-17 [19]
Costo de beneficiado	Costos de beneficiado de café aceptados por ley no 2762 cosecha 2022-2023 [20]
	Costos de beneficiado de café aceptados por ley no 2762 cosecha 2021-2022 [21]
	Costos de beneficiado de café aceptados por ley no 2762 cosecha 2020-2021 [22]
	Costos de beneficiado de café aceptados por ley no 2762 cosecha 2019-2020 [23]
Temperatura media	Cambio climático Cerros de Dota [24]
Precipitaciones	Cambio climático Cerros de Dota [24]
Tipo de cambio	Precio de liquidación final cosecha 2022-2023 [25]

Instituto Tecnológico de Costa Rica, noviembre 1,
2024, Cartago, Costa Rica

Factor	**Fuente de información del factor relevante**
	Precio de liquidación final cosecha 2021-2022 [26]
	Precio de liquidación final cosecha 2020-2021 [27]
	Precio de liquidación final cosecha 2019-2020 [28]
	Precio de liquidación final cosecha 2018-2019 [29]
	Precio de liquidación final cosecha 2017-2018 [30]
	Precio de liquidación final cosecha 2016-2017 [31]
Crecimiento económico mundial	Crecimiento del PIB (% anual) [32]
Producción nacional	Informe sobre la actividad cafetalera de Costa Rica 2023 [33]
	Informe sobre la actividad cafetalera de Costa Rica 2021 [34]
Oferta Mundial	Informe sobre la actividad cafetalera de Costa Rica 2023 [33]
	Informe sobre la actividad cafetalera de Costa Rica 2021 [34]
Demanda mundial	Informe sobre la actividad cafetalera de Costa Rica 2023 [33]
	Informe sobre la actividad cafetalera de Costa Rica 2021 [34]
Precio según la Bolsa de Nueva York	Precios históricos del café en Nueva York – 2024 [35]
	Precios históricos del café en Nueva York – 2023 [36]
	Precios históricos del café en Nueva York – 2022 [37]
	Precios históricos del café en Nueva York – 2021 [38]
	Precios históricos del café en Nueva York – 2020 [39]

Instituto Tecnológico de Costa Rica, noviembre 1,
2024, Cartago, Costa Rica

Factor	**Fuente de información del factor relevante**
	Precios históricos del café en Nueva York – 2019 [40]
	Precios históricos del café en Nueva York – 2018 [41]
	Precios históricos del café en Nueva York – 2017 [42]
Precio de liquidación	Precio de liquidación final cosecha 2022-2023 [25]
	Precio de liquidación final cosecha 2021-2022 [26]
	Precio de liquidación final cosecha 2020-2021 [27]
	Precio de liquidación final cosecha 2019-2020 [28]
	Precio de liquidación final cosecha 2018-2019 [29]
	Precio de liquidación final cosecha 2017-2018 [30]
	Precio de liquidación final cosecha 2016-2017 [31]