Estimación de la producción de café cereza basada en Series de Tiempo

Jhonn Pablo Rodríguez Muñoz

Tesis de Maestría en Ingeniería Telemática

Director:

PhD. David Camilo Corrales Muñoz

Co-Director:

PhD. Juan Carlos Corrales Muñoz

Universidad Del Cauca

Facultad de Ingeniería Electrónica y telecomunicaciones

Departamento de Telemática

e- @mbiente

Popayán, diciembre de 2021.

Jhonn Pablo Rodríguez Muñoz

Estimación de la producción de café cereza basada en Series de Tiempo

Tesis presentada a la
Facultad de Ingeniería Electrónica y Telecomunicaciones
de la Universidad del Cauca, Colombia
para otorgar el grado académico de

Magíster en:

Ingeniería Telemática

Director:

Dr. David Camilo Corrales Muñoz, PhD

Co-directores:

Dr. Juan Carlos Corrales Muñoz, PhD

Popayán

2021

Agradecimientos

Dedico este trabajo de grado a mis padres: Rosalba Muñoz y Carlos Humberto Rodríguez, gracias por ser los promotores de mis sueños y por cada día confiar en mí. Gracias por cada consejo y por cada palabra que me guiaron durante mi vida y que me han llevado a conseguir tantos logros. Agradezco a mi hermano Carlos Andrés Rodríguez Muñoz, que siempre ha estado junto a mí, brindándome su apoyo y ser el ejemplo a seguir.

A mi novia Isis Victoria, por ser mi compañera de vida, quien siempre ha estado conmigo en los momentos felices y difíciles de la vida. Te agradezco muchísimo amor.

A mis compañeros de estudio: Ana Isabel Montoya, Carlos David Rodríguez y Carlos Andrés Gonzales Amarillo, quienes sin esperar nada a cambio compartieron sus conocimientos y experiencias, en compañía de un buen café. A todos mis demás compañeros, mil gracias.

Agradezco también a mis asesores de tesis, los ingenieros David Camilo Corrales y Juan Carlos Corrales por haberme brindado la oportunidad de recurrir a su capacidad y conocimiento científico, así como también por haberme tenido toda la paciencia para guiarme durante mi vida profesional.

Agradezco al Grupo de Ingeniería Telemática (GIT) y al Departamento de Telemática de la Universidad del Cauca, por haberme abierto las puertas de su seno científico para poder estudiar, y a los diferentes docentes que brindaron sus conocimientos y su apoyo para desarrollar este trabajo de grado. Así como también al proyecto IoT-Agro por su apoyo económico para realizar mis estudios de posgrado.

Resumen estructurado

Antecedentes: El cultivo del café está situado en más de 60 países, donde Colombia siempre ha ocupado un lugar importante como productor, con 22 departamentos y 590 municipios cafeteros, que suman 853.698 hectáreas sembradas. El sector cafetero Colombiano, está basado en la productividad y reducir los costos. La estimación de la producción de café en el sector ha sido de mayor importancia para permitir a los caficultores tener información relevante para la toma de decisiones adecuadas, en la planeación de actividades. Los caficultores colombianos estiman la producción de café cereza basados en un modelo destructivo, lo cual genera pérdidas en sus producciones. Varios trabajos de investigación tienen primeras aproximaciones a la estimación de la producción de café, basados en técnicas de inteligencia artificial, específicamente con conjuntos de imágenes. Los esfuerzos de investigación en los últimos años en otros cultivos se han centrado en la creación de modelos de mayor complejidad con información multi-variable.

Objetivos: Establecer un modelo estadístico para estimar la producción de café cereza para mejorar la planeación de actividades previas a la cosecha.

Métodos: Se propone usar modelos estadísticos para la estimación de la producción de café cereza basada en series de tiempo, a partir de información de manejo de cultivo e información climática para una finca cafetera del Departamento del Cauca.

Resultados: El presente trabajo entrega como resultados un conjunto de series de tiempo que representan la producción de café cereza en una finca cafetera ubicada en el corregimiento “La Venta”, del municipio Cajibio (Cauca). Además un prototipo que estima la producción de café cereza a partir de un modelo estadístico.

Conclusiones: La construcción de un modelo estadístico para la estimación de la producción de café cereza basada en información multi-variable, permite a los caficultores realizar la planeación de actividades previas a la cosecha disminuyendo tiempos y costos. Además les permite la no utilización de modelos destructivos que generan pérdidas en sus producciones.

Palabras clave: café cereza, modelos estadísticos, series de tiempo, información multi-variable.

ABSTRACT

Background: Coffee crop is located in more than 60 countries, where Colombia has always occupied an important place as a producer, with 22 departments and 590 coffee municipalities, which add up to 853,698 hectares planted. The Colombian coffee sector is based on productivity and reducing costs. Estimating coffee production in the sector has been of greater importance in enabling coffee growers to have information relevant to appropriate decision-making in planning activities. Colombian coffee growers estimate cherry coffee production based on a destructive model, resulting in losses in their productions. Several research papers have early approximations of coffee production estimation, based on artificial intelligence techniques, specifically with imaging sets. Research efforts in recent years in other crops have focused on creating more complex models with multi-variable information.

Objectives: Establish a statistical model to estimate the production of cherry coffee to improve the planning of pre-harvest activities.

Methods: It is proposed to use statistical models to estimate cherry coffee production based on time series, based on information on crop management and climate information for a coffee farm in the Department of Cauca.

Results: The present work provides, as a result, a set of time series that represent the production of cherry coffee in a coffee farm located in the village of "La Venta", Cajibio municipality (Cauca). In addition, a prototype that estimates the production of cherry coffee from a statistical model.

Conclusions: The construction of a statistical model for estimating cherry coffee production based on multi-variable information allows coffee growers to plan activities prior to harvest, reducing times and costs. It also allows them not to use destructive models that generate losses in their productions.

Keywords: cherry coffee, statistical models, time series, multi-variable information.

Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

En este capítulo se describe el planteamiento del problema, el escenario de motivación, el objetivo general y los objetivos específicos que guían el presente trabajo de maestría, y finalmente se presentan las contribuciones.

Introducción

1.1 Planteamiento del problema

El café es la segunda bebida más consumida a nivel mundial después del agua, según *The International Coffee Organization* (ICO) [1]. El gran consumo de café ofrece importantes oportunidades de crecimiento para los países exportadores del grano; siendo Brasil, Vietnam y Colombia los mayores exportadores de café en el mundo [2]. En Colombia, el café cumple un papel prioritario en la generación de empleo rural, en esta actividad se ocupa más de 785 mil personas de manera directa, siendo el 26% de la totalidad de empleos en el sector agrícola en Colombia [3].

Debido a la importancia del cultivo de café en Colombia, los productores estiman la producción de café cereza (grano en estado de madurez que ha sido extraído de las plantas de café en la etapa de “Recolección” de la cadena productiva [4]), con la finalidad de soportar la toma de decisiones en la planeación de actividades, número de trabajadores requeridos, infraestructura necesaria, negociaciones anticipadas y pérdidas de producción de café en un determinado territorio [5].

Actualmente, los caficultores colombianos estiman la producción de café cereza basados en mediciones directas en campo [6]. El proceso de recolección de muestras consiste en seleccionar 60 árboles de café por hectárea, de los cuales se extraen los granos de café y posteriormente se pesan [6]. Debido a la complejidad y costos en el proceso de recolección de muestras, se cuentan con pocos datos, lo cual limita a los caficultores a utilizar este tipo de medidas que no tienen un intervalo de confianza suficiente para estimar correctamente la producción de café cereza. Adicionalmente, los granos extraídos de los cafetales para este proceso se descartan de la cadena productiva de café, lo cual genera pérdidas a los caficultores (Modelo destructivo).

Teniendo en cuenta la necesidad de estas herramientas para el uso adecuado de las labores de campo, a nivel de Colombia, el sector cafetero adolece de modelos no destructivos, de bajo costo y tiempo para estimar la producción de café cereza a partir de información multi-variable como, clima (datos de sensores climáticos, anuarios meteorológicos), practicas agronómicas (históricos de manejo de cultivo, estudios de suelo, reportes técnicos, etc.), conocimiento experto, etc.

Basado en lo anterior, se plantea la siguiente pregunta de investigación:

¿Cómo estimar la producción de café cereza a partir de información multi-variable para mejorar la planeación de actividades previas a la cosecha?

1.2 Escenario de motivación

El cultivo del café está situado en más de 60 países, se estima que son más de 125 millones las personas cuya subsistencia depende del café de las cuales países como Colombia siempre han ocupado un lugar importante como productor; con más de 540.000 familias productoras, siendo el patrimonio social estratégico más importante de la nación; contando con 22 departamentos y 590 municipios cafeteros, que cultivan 853.698 hectáreas [3].

El modelo actual del sector cafetero Colombiano, está basado en la productividad, y el desarrollo de variedades de café altamente productivas, donde se pretende incrementar el volumen de cosechas, y reducir los costos unitarios en cosecha y procesamiento. La estimación de la producción de café en el sector ha sido de mayor importancia para permitir a los caficultores tener información relevante para la toma de decisiones adecuadas, en la planeación de actividades, número de trabajadores requeridos, infraestructura necesaria, negociaciones anticipadas y pérdidas de producción de café [5].

1.3 Objetivos

1.3.1 Objetivo general

Establecer un modelo estadístico para estimar la producción de café cereza para mejorar la planeación de actividades previas a la cosecha.

1.3.2 Objetivos específicos

- Construir una serie de tiempo para estimar la producción de café cereza.
- Definir un modelo estadístico que estime la producción de café cereza a partir de la serie de tiempo.
- Desarrollar un prototipo que valide el modelo estadístico propuesto.

1.4 Contribuciones

Las principales contribuciones de este trabajo son:

- Un conjunto de datos en escala semanal y mensual que representa la producción de café cereza, que involucra información de manejo de cultivo e información climática.
- Un prototipo que implementa el modelo estadístico para la estimación de la producción de café cereza.
- Un artículo publicado en “Pattern Recognition Letter”, titulado: “*A computer vision system for automatic cherry beans detection on coffee trees*”. Computer Vision and Pattern Recognition. Elsevier. Publindex A1.
- Un artículo publicado en “Computers and Electronics in Agriculture”, titulado: “*IoT-Agro: A smart farming system to Colombian coffee farms*”. Publindex A1.
- Un artículo publicado en “Computers, Materials & Continua”, titulado: “*A Non-Desctructive Time Series Model for the Estimation of Cherry Coffee Production*”. Publindex A1.

1.5 Contenido de la monografía

La monografía se encuentra organizada a partir de las fases de la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) [7], la cual ofrece seis fases (comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implantación). Cada una de estas fases se presenta en los siguientes capítulos del presente documento. La monografía se encuentra organizada en siete capítulos, los cuales son resumidos a continuación:

- **Capítulo 2. Estado actual del conocimiento / comprensión del negocio**

  Presenta una visión general de los conceptos y los trabajos relacionados que giran alrededor del problema de investigación declarado.

- **Capítulo 3. Comprensión y preparación de los datos**

  Presenta los conjuntos de datos usados y los métodos utilizados para la preparación de los datos.

- **Capítulo 4. Modelado**

  Explica las series de tiempo propuestas para el presente trabajo de investigación y los modelos estadísticos para la estimación de la producción de café cereza.

- **Capítulo 5. Experimentación y evaluación**

  Presenta el proceso de evaluación y las pruebas ejecutadas con el fin de analizar la calidad de los resultados y su rendimiento.

- **Capítulo 6. Prototipo**

  Presenta el proceso de desarrollo de software llevado a cabo para construir el prototipo que contiene los modelos estadísticos para la estimación de la producción de café cereza.

- **Capítulo 7. Conclusiones y trabajos futuros**

  En este capítulo se analizan los resultados del trabajo realizado, se detallan las principales contribuciones obtenidas en la ejecución del trabajo de maestría y se expone un conjunto de recomendaciones importantes para el desarrollo de trabajos futuros.

Capítulo 2: Estado actual del conocimiento / comprensión del negocio

En este capítulo, se presentan los precedentes teóricos que permiten comprender el contexto del presente trabajo de maestría, el cual consiste en la estimación de la producción de café cereza basada en series de tiempo. Luego, se presentan los trabajos de investigación relacionados al planteamiento del problema de la tesis de maestría. Finalmente, se realiza un resumen que describe los principales aportes del capítulo.

2.1 Conceptos generales

En esta sección se explican los conceptos: producción de café cereza, series de tiempo y modelos estadísticos con el objetivo de contextualizar el presente trabajo de investigación.

2.1.1 Producción de café cereza

La cadena de valor del café [4] es una serie de procesos, en los cuales los granos de café sufren una transformación, hasta la constitución de un producto final y su colocación en el mercado. La Figura 1 presenta la cadena productiva del café, compuesta por 10 procesos (rectángulos color azul) y las flechas representan el flujo de secuencia y la salida que tiene al finalizar cada proceso. Las circunferencias verde y roja, representan el inicio y fin de la cadena productiva.

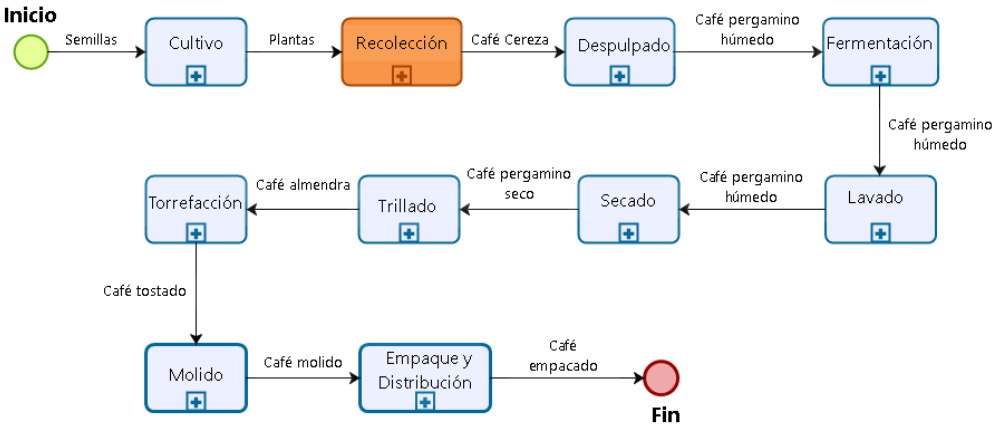

Figura 1. Cadena de valor del café. Fuente [8]

La producción de café cereza está involucrada al finalizar el proceso de "*Recolección*" (rectángulo color naranja), donde se recogen uno a uno los granos de café maduros (café cereza) de las plantas de café, y es ahí donde se pesan y se obtiene la producción de café cereza en kilogramos.

2.1.2 Series de tiempo

Las series de tiempo o series temporales [9] son datos estadísticos que se recopilan, observan o registran en intervalos de tiempo regulares (diario, semanal, mensual, semestral, anual, entre otros). El término serie de tiempo se aplica a datos registrados en forma periódica. Matemáticamente una serie de tiempo se define como:

$$S = [(t_1, c_1), (t_2, c_2), \ldots, (t_i, c_i), \ldots, (t_n, c_n)] \text{ con } (t_1 < t_2 < \ldots < t_i < \ldots < t_n)$$

Donde c_i es un punto en el espacio que pertenece a los datos, y t_i el instante de tiempo al cual corresponde el dato c_i [10].

La existencia de un orden temporal entre los puntos de datos que forman una serie hace que su análisis sea diferente al de otros problemas que no presentan esta característica.

El cambio de la media a través del tiempo en las series temporales, se conoce como tendencia [10]. Existen dos tipos de tendencia: tendencia lineal, la cual refleja un crecimiento o decrecimiento lineal de las variables de la serie temporal. La tendencia no lineal, no se ve reflejada como línea recta, puede ser exponencial, logarítmica, entre otras.

2.1.3 Modelos estadísticos

Los modelos estadísticos hacen uso de funciones multivariables que transforman la información extraída desde los datos [11]. Este tipo de modelos, estiman el valor de la variable dependiente a partir de un conjunto de variables independientes o una serie de tiempo [12]. En este orden de ideas, en la presente propuesta de investigación, la variable dependiente será la producción de café cereza, y las variables independientes involucrarán inicialmente información climática e información de manejo del cultivo. Dentro de los modelos estadísticos más representativos se encuentran: Linear Regression (LR), Decision Trees (DT), Bayesian Networks (BN), Artificial Neural Network (ANN), XgBoost y Support Vector Regression (SVR) [13]. Por otra parte, la construcción de la serie de tiempo juega un papel muy importante, ya que el pre-

procesamiento de los datos determina la estrategia de estimación de la variable dependiente.

2.2 Trabajos relacionados

En esta sección se explican las aproximaciones existentes relacionadas con el problema de investigación declarado. Las propuestas de estos trabajos están orientadas a la estimación de producción de café y a la estimación de la producción en diferentes cultivos.

2.2.1 Estimación de la producción de café

En cultivos de café son pocos los trabajos que se enfocan a la estimación de la producción, las investigaciones que se han realizado hasta el momento están enfocadas en la detección de enfermedades como la roya. Sin embargo, existen trabajos con primeras aproximaciones a la estimación de la producción, los cuales se describen a continuación.

En [14], realizan un sistema de *Computer Vision* para el conteo de granos de café en las ramas de los árboles. Los autores utilizaron un conjunto de datos de 1018 imágenes de ramas de los árboles de café en diferentes etapas de maduración. El sistema fue validado en cuatro parcelas de café variedad Castillo, en diferentes etapas de desarrollo y con diferentes densidades; con una correlación superior al 90% en las primeras etapas del desarrollo del cultivo.

En [15], proponen un enfoque clásico de *Computer Vision* para detectar granos de café cereza en los cultivos, a partir de imágenes capturadas por un celular de gama media en ambiente no controlado. El conjunto de imágenes utilizado en este estudio contenía imágenes de la planta de café completa de 3 variedades de café (caturra, bourbon y castillo). El sistema logró los mejores resultados para los cafetales bourbon con 0.59 de precisión; 67% del total de granos de cereza relevantes correctamente clasificados.

En [16], evaluaron la capacidad de un modelo de *Extreme Learning Machine* (ELM) para analizar las propiedades de la fertilidad del suelo (materia orgánica del suelo, potasio, boro, azufre, zinc, fósforo, nitrógeno, calcio intercambiable, magnesio y pH) y generar una estimación precisa del rendimiento del café Robusta. En comparación con los modelos Multiple Linear Regression y Random Forest, la adopción del modelo ELM

contribuye a la selección adecuada de las propiedades más óptimas del suelo que se pueden utilizar en el rendimiento del café. El modelo ELM construido con características de materia orgánica, potasio y azufre como variables predictoras generó la estimación del rendimiento de café más precisa.

Los autores de [17], proponen un modelo estadístico para pronosticar los rendimientos del café usando datos recopilados durante un período de 10 años de cuatro provincias productoras de café en Vietnam, los modelos involucran métodos bayesianos basados en variables agroclimáticas, la evapotranspiración real derivada de satélites y la información sobre el manejo de cultivos. El modelo logro una mediana de pronóstico MAPE y RMSE que varió entre 13% y 16% y entre 420 kg ha - 1 y 456 kg ha- 1.

En [18], combinan técnicas de vehículos aéreos no tripulados (UAV) con modelos estadísticos para estimar el rendimiento del cultivo de café. Los autores utilizaron los UAV equipados con una cámara RGB (Rojo, Verde, Azul) y algoritmos de visión por computadora para estimar la altura y el diámetro de la copa del cafeto en cultivos de Brasil, esto con el fin de recolectar datos. Luego implementaron modelos estadísticos para predecir el rendimiento, de los cuales el que mejor resultado obtuvieron fue el Neuroevolución de topologías de aumento (NEAT), con un MAPE del 37% para un conjunto de datos que solo contenía: índice de área foliar – LAI y diámetro de la copa.

Los trabajos de investigación descritos anteriormente, se enfocan a la estimación de la producción de café en el sector cafetero. Los principales enfoques usados para la estimación de la producción en cultivos de café son *Computer Vision* y *Modelos Estadísticos*.

2.2.2 Estimación de la producción en diferentes cultivos

Teniendo en cuenta que, hasta el momento, existen pocos trabajos de investigación orientados a la estimación de la producción de café, a continuación, se mencionan los trabajos más relevantes para la estimación de producción en otros cultivos.

El modelo propuesto en [19] utiliza dos conjuntos de datos de los años 1983 al 2013 con las características climáticas (temperatura, lluvia y humedad) y la producción del cultivo de arroz de tres distritos costeros pertenecientes a “Odisha” ubicados en la India. Utilizan el método AdaBoost para predecir la producción de arroz y el cual combina los resultados obtenidos por los modelos estadísticos: lineal, LASSO, cresta y una máquina de vector de soporte para regresión (SVR). Los resultados de los

errores calculados (MAE, MSE, MAD y R-Squared) para el método AdaBoost, representan calidad de predicción del método de conjunto propuesto en este trabajo.

En [20] proponen un modelo de regresión lineal múltiple (MLR) para estimar la producción de caña de azúcar en el estado de São Paulo en Brasil, con base en series temporales de datos meteorológicos y agroclimáticos. El modelo propuesto en este trabajo, utiliza variables de área plantada, índice de vegetación de diferencia normalizada (NDVI) e índice de satisfacción de requerimiento de agua (WRSI), los cuales presentaron coeficientes de correlación alrededor del 0.9. Además, los modelos mostraron una relación directamente proporcional entre la producción de caña de azúcar y el NDVI; e inversamente proporcional al WRSI.

En [21] los autores combinaron datos de imágenes satelitales de MODIS (Espectro radiómetro de imágenes de resolución moderada) con información de producción del cultivo, para desarrollar un modelo de estadístico para pronosticar la producción de trigo de invierno y cebada en Irak. A partir de la información de las imágenes satelitales, los autores hallaron los índices de vegetación: NDVI y EVI (índice de vegetación mejorado); donde el mejor resultado de la investigación lo obtuvieron con el índice de vegetación NDVI.

Un enfoque para el pronóstico de cosechas de aceitunas con base en los recuentos de polen de oliva en el aire y observaciones meteorológicas y prácticas agronómicas es desarrollado en [22], este estudio fue realizado en la Campiña Alta, Córdoba, (suroeste de España). Los datos de las diferentes fuentes de información son combinados y se obtuvieron cuatro ecuaciones matemáticas para pronosticar el cultivo con 6 meses de antelación, con diversos grados de confiabilidad. La regresión lineal $Y = -1.90 * 10^4 + 2.35\,X + 53.94\,Z$, obtuvo los mejores resultados; donde Y es la producción de aceitunas, X el recuento de polen del olivo y Z la precipitación antes de la floración.

En [23], utilizan los datos históricos del rendimiento de trigo de la Universidad de Agricultura y Tecnología Odisha de la India, para pronosticar la producción de trigo, a partir de ecuaciones matemáticas de grado 1 a 4. En este trabajo clasifican la variable de producción de trigo en 7 intervalos difusos (rendimiento muy deficiente, rendimiento deficiente, no tan buen rendimiento, producción media, buen rendimiento, muy buen rendimiento, excelente rendimiento). Los autores obtuvieron el MSE (Error cuadrático medio) más bajo en la ecuación cúbica de tercer grado (MSE = 180.98).

En [24], proponen una arquitectura del modelo de predicción del rendimiento del cultivo que incluye un módulo de entrada, que es responsable de tomar la opinión del agricultor. El módulo de entrada considera el nombre del cultivo, área de la tierra, tipo de suelo, pH del suelo, detalles de plagas, clima, nivel de agua, tipo de semilla. El módulo de selección es responsable de la selección de subconjuntos de un atributo. Luego, el modelo de predicción del rendimiento de los cultivos es utilizado para predecir el crecimiento de las plantas y enfermedades de las plantas. Después de la selección de entidades, los datos van a la regla de clasificación para agrupar contenidos similares. Los autores utilizan un modelo de regresión para la predicción de producción, a partir de variables como la variedad, área de cultivo, tipo de suelo, pH del suelo, control de plagas, nivel del agua, tipo de semilla. Por último, realizan la selección de variables, agrupando los datos con variables climáticas y parámetros de cultivos utilizados para predecir la producción del cultivo.

En [25], implementan dos técnicas para la predicción de cultivos (regresión lineal múltiple MLR y clustering basado en la densidad DBC), estos modelos fueron experimentados en el distrito de East Godavari de Andhra Pradesh en la India. Los autores consideraron las siguientes variables de entrada para los modelos: año (fecha en la que fue capturado el dato), pluviosidad, área de siembra, rendimiento, fertilizantes (nitrógeno, fósforo, potasio) y la producción del cultivo. Los resultados obtenidos en este trabajo fueron alrededor del 2% de diferencia entre el valor real y los valores pronosticados.

En [26], presentan un sistema basado en aprendizaje máquina que utiliza datos de múltiples fuentes para realizar pronósticos de rendimiento de soja y maíz. El sistema está compuesto por una red neuronal recurrente (RNN) entrenada con variables de precipitación, temperatura, suelo y producción histórica de soja o maíz, en regiones de Brasil y Estados Unidos. Para evaluar los resultados obtenidos con diferentes configuraciones de RNN, los autores midieron el coeficiente de determinación obteniendo valores entre 0.55 - 0.75.

A través de regresiones lineales [27], en China predicen la producción del trigo a partir de datos de cobertura terrestre, datos de producción (rendimiento del trigo, maíz, soja y arroz), NDVI, información auxiliar (límites administrativos, calendario de cultivos y distribución geográfica). Primero los autores extraen indicadores analíticos de los NDVI de la serie temporal, y luego eliminan la tendencia de la serie de tiempo utilizando una función lineal de tendencia al alza. Luego utilizan regresiones lineales con las demás

características mencionadas. El porcentaje de error de predicción obtenido en este trabajo es inferior al 8% y también obtienen un coeficiente de determinación del 86.6%.

En [28], es presentada una herramienta llamada CST (Herramienta de Estadísticas de Cultivos), la cual permite predecir el rendimiento de cultivos, específicamente en el maíz, utilizando datos históricos de manejo de cultivos y datos meteorológicos o de detección remota, a partir de análisis de regresión múltiple o análisis de escenarios (busca los años más similares al año actual). En este estudio encontraron que los rendimientos de maíz en Etiopía están altamente correlacionados con la precipitación de la vegetación de la región. Particularmente, en Oromiya (Etiopia), la herramienta obtuvo para 4 zonas un R^2 cerca de 0.9.

Modelos lineales simples son implementados en [29], utilizando datos de detección remota para pronosticar el rendimiento de cultivos de arroz en la provincia de Hubei. Compararon el rendimiento real del cultivo a partir de datos estadísticos con los resultados del modelo. Los resultados indican que el error varía entre -14.38% y 11.31% en comparación con los datos reales, y el coeficiente de correlación es de 0.87.

En [30], construyen un modelo de regresión no paramétrico para el pronóstico de rendimiento agrícola, implementado en cultivos de manzanas en Corea, basado en datos climáticos mensuales de 33 años (temperatura máxima, temperatura mínima, temperatura promedio, pluviosidad y horas de sol) en tiempo real y datos de producción del cultivo. El modelo obtuvo para el mes de diciembre un Error Porcentual Absoluto Medio (MAPE) de 5.713087e-12 y un coeficiente de determinación igual a 1.

Un modelo gris (GM) y un modelo auto-regresivo integrado de media móvil (ARIMA), son implementados en [31], para predecir el rendimiento de los cultivos de granos, a partir de datos de rendimiento de los años 1998 al 2008. Las predicciones fueron realizadas entre los años 2009 al 2013, y obtuvieron un error promedio de 7.88% (GM) y 12.32% (ARIMA) y una precisión promedio de 92.12% y 87.68% respectivamente. Siendo los resultados del modelo GM los mejores para la predicción del rendimiento de cultivos de granos.

En [32] son utilizados varios modelos de regresiones lineales y no lineales mediante la validación cruzada, para predecir la producción de trigo en distintos condados de Estados Unidos. El conjunto de datos utilizado en este trabajo de investigación involucró 2 años de datos de variables de geo localización de condados, datos climáticos, índices de vegetación y la variable objetivo la producción de trigo. El modelo

Random Forest obtuvo el mejor rendimiento, con un valor de R^2 de 0.83, un RMSE de 5.3 y un error porcentual absoluto medio del 5%.

En [33] predicen el rendimiento del cultivo de cacao, a través de un conjunto de variables fotosintéticas, morfológicas, climáticas, químicas y físicas (años 2015, 2016 y 2017), a partir de dos modelos (Modelo Lineal Generalizado - GLM y las Máquinas de Vector de Soporte - SVR). La construcción y comparación de los resultados de los dos modelos, fue útil para comprobar que las variables explicativas: diámetro del tronco, fósforo, magnesio, radiación, temperatura, humedad y lluvias acumuladas son las variables que explican en mayor medida el rendimiento del cultivo de cacao. El modelo con mayor precisión corresponde al Lineal Generalizado, el cual obtuvo un bajo coeficiente de correlación (R^2 = 0.13) y errores altos (RMSE = 1708.29; MAE = 1028.78).

En [34], realizan la predicción de producción del cultivo de macadamia en seis regiones de Australia, a partir de variables del cultivo (Edad de los árboles, Variedad, Región y Espaciado de árboles) y de variables climáticas (Temperatura máxima y mínima, Evaporación, Radiación solar, Lluvia, Eficiencia de transpiración modelada, Estrés hídrico y el Índice de suelo-agua modelado). Con la validación cruzada los autores demostraron que los modelos de regresión LASSO con errores de MAE <10%, superaron a los modelos lineales generales y a la regresión parcial de mínimos cuadrados.

El trabajo presentado en [35], desarrollan un sistema automatizado con redes convolucionales (CNN) para la predicción del rendimiento del algodón a partir de imágenes en color adquiridas por un dispositivo móvil simple. Los modelos son entrenados con imágenes adquiridas en diferentes momentos durante el día y evaluando tres escenarios diferentes (recursos computacionales de baja, media y alta demanda). El sistema propuesto obtiene un error del 17,86% al predecir el rendimiento utilizando 205 imágenes del conjunto de datos de prueba.

En [36], presentan un modelo de simulación en cultivos de arroz en Tamil Nadu para predecir la producción del cultivo a través del algoritmo de clasificación ZeroR. Este modelo está basado en datos de un periodo de 10 años los cuales incluyen: datos de suelo, clima, fertilizantes y riego; la evaluación del modelo obtuvo un error absoluto medio de 0.0364 en sus resultados.

En [37], utilizan imágenes de cultivos de: yuca, cítricos, coco, maíz, algodón, café, anacardo, vides, soja, caña de azúcar y trigo; donde abarcan muchas características

del cultivo, pero con la desventaja de que el conjunto de imágenes es pequeño. La principal limitación de este trabajo es que las bases de datos construidas no son lo suficientemente grandes para la implementación de redes neuronales convolucionales (CNN). El autor concluye que cuando una CNN es entrenada sólo con imágenes capturadas en ambiente no controlado, la precisión disminuye del 99% al 68%; y cuando las imágenes se invierten, la precisión cae al 33%.

En la Figura 2, se presentan el número de trabajos de investigación encontrados para la estimación de la producción en diferentes cultivos. Se puede observar que el mayor número de esfuerzos en investigación ha sido para el cultivo del trigo, debido a que es el cultivo más dominante en el comercio mundial [38].

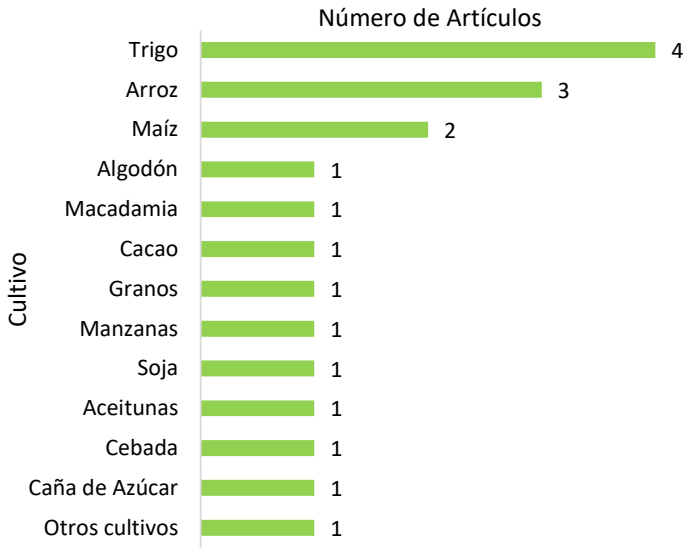

Figura 2. Trabajos encontrados por tipo de cultivo.

Los trabajos de investigación encontrados en esta sección utilizan 2 enfoques de la inteligencia artificial (Deep Learning y Modelos Estadísticos), como se puede observar en la Figura 3.

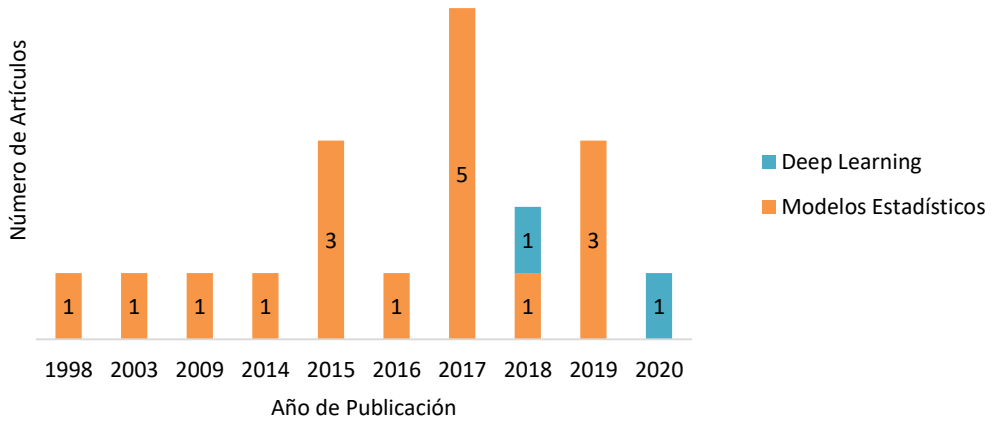

Figura 3. Enfoques utilizados para la estimación de producción en cultivos.

Se puede percibir que la implementación de modelos estadísticos para la estimación de la producción se ha mantenido frecuentemente en el tiempo, y en los últimos años el Deep Learning ha aparecido como una nueva alternativa que hace uso de bases datos de imágenes. En la Tabla 1, se observan las principales brechas encontradas de los trabajos anteriormente mencionados.

Cultivo	Brecha	Referencia
Café	No realizan estimación de producción, además utilizan conjunto de imágenes para detectar el fruto.	[14]
	Los autores no realizan una estimación de producción, ni tienen en cuenta información de diferentes fuentes.	[15]
	En este trabajo utilizan información de propiedades del suelo para obtener el mejor rendimiento del cultivo, sin embargo no estiman la producción del cultivo.	[16]
	Realizan predicción del rendimiento mas no de la producción de café, además no tienen en cuenta las series temporales ni variables climáticas relevantes al cultivo como la humedad relativa y radiación solar y tampoco actividades de manejo de cultivo como control y limpieza.	[17]
	Los autores utilizan información recolectada por UAV, y no tienen en cuenta información climática, ni de manejo de cultivo.	[18]
Arroz	El conjunto de datos utilizado solo cuenta con variables climáticas.	[19]
	Los autores utilizan un conjunto de imágenes para entrenar los modelos y comparar con los datos reales, pero no involucran información relevante al cultivo ni climática.	[29]
	Este estudio combina diferentes variables como clima, propiedades del suelo y otras, sin embargo no utilizan información de manejo de cultivo.	[36]

Caña de azúcar	Utilizan un conjunto de imágenes, de las cuales solo extraen información climática y no tienen en cuenta información de manejo de cultivo.	[20]
Trigo y cebada	Involucran información de índices de vegetación obtenidos de imágenes satelitales, pero no de manejo de cultivo.	[21]
Aceituna	Combinan diferentes fuentes de información de clima y manejo de cultivo, pero consiguen que solo dos variables sean necesarias para la estimación.	[22]
Trigo	Solo tienen en cuenta la variable de producción del cultivo por año y no incluyen otro tipo de fuente de información.	[23]
	Estiman la producción a partir de información histórica de producción e índices de vegetación, pero no involucran información climática.	[27]
	El conjunto de datos contiene información climática y de producción, sin embargo no utilizan información de manejo de cultivo.	[32]
Soja y maíz	En este estudio utilizan información de clima para entrenar redes neuronales, pero no tienen en cuenta información de manejo de cultivo u otro tipo de fuente de información.	[26]
Maíz	Realizan un análisis de índices de vegetación y algunas variables climáticas pero no incluyen información de manejo de cultivo.	[28]
Manzanas	El conjunto de datos consta de 33 años de variables climáticas y de producción del cultivo, sin embargo no tienen en cuenta información de manejo de cultivo.	[30]
Granos	Las series de tiempo utilizadas en este estudio solo tienen la variable de producción del cultivo, sin tener en cuenta otras fuentes de información relevantes al cultivo.	[31]
Cacao	Los modelos fueron entrenados con información fotosintética, morfológica, climática, química y física. Sin embargo no combinan información de manejo de cultivo.	[33]
Macadamia	El conjunto de datos combina información climática como información de cultivo, específicamente datos tomados en campo, sin embargo no tienen en cuenta información de actividades de control o fertilización del cultivo.	[34]
Algodón	La CNN utilizada en este trabajo solo es entrenada con información de imágenes tomadas de un dispositivo, y no involucran información climática ni tampoco de manejo de cultivo.	[35]
Otros	En este trabajo la información utilizada para entrenar los modelos es tanto climática como propiedades del suelo y tipos de semilla del cultivo.	[24]
	Los autores utilizan información climática y características del cultivo en el conjunto de datos, pero no utilizan información de manejo de cultivo.	[25]
	Utilizan un conjunto de imágenes muy pequeño y que han sido capturadas en ambientes no controlados. Por consiguiente no tienen en cuenta información climática ni de manejo de cultivo para la estimación.	[37]

Tabla 1. Brechas en trabajos encontrados.

La mayoría de los trabajos de investigación encontrados incluyen conjuntos de datos con información relevante al cultivo, sin embargo no combinan diferentes fuentes de información para crear modelos explicativos a partir de diferentes variables (meteorológicas, manejo del cultivo, etc). Aunque existen trabajos que se enfocan en el uso de imágenes (haciendo uso de índices de vegetación, propiedades del suelo, clima, etc) con técnicas provenientes de computer visión, el presente trabajo parte de la premisa de la construcción de series de tiempo multivariadas a partir de datos recolectados que no involucran imágenes.

2.3 Resumen

Con el fin de comprender la temática del presente trabajo de maestría, este capítulo presentó los conceptos teóricos relacionados con la producción de café cereza, series de tiempo, y modelos estadísticos. Finalmente, fueron expuestos los trabajos relacionados respecto al problema declarado mediante una revisión literaria divida en estimación de la producción de café y la estimación de producción en diferentes cultivos. Como conclusión, los trabajos encontrados enfocados en la estimación de la producción de café son pocos y además utilizan imágenes para la detección de frutos lo que conlleva la implementación de enfoques de computer visión, sin embargo, los trabajos más recientes implementan modelos estadísticos para realizar las estimaciones sin tener en cuenta diferentes fuentes de información. En cuanto a los trabajos encontrados para la estimación de producción en diferentes cultivos, podemos afirmar que hay distintos enfoques dependiendo del conjunto de datos utilizado en cada uno de ellos. Por ejemplo para conjunto de imágenes los enfoques están basado en deep learning y para conjunto de datos los enfoques implementados son basados en modelos estadísticos. Algunos trabajos no tienen en cuenta la información de diferentes fuentes ni datos relevantes a los cultivos, solo incluyen información que está disponible. Además los algoritmos implementados en los trabajos encontrados pertenecientes a los modelos estadísticos, son pocos, por consiguiente falta explorar con otros algoritmos de este enfoque.

Capítulo 3: Comprensión y preparación de los datos

En este capítulo se describe el proceso de comprensión de datos, que abarca la producción de café cereza y la descripción del conjunto de datos utilizado. En la segunda parte de este capítulo se presenta el proceso de preparación de datos (pre-procesamiento), con el objetivo de adaptar el conjunto de datos a los *Modelos Estadísticos*.

3.1 Comprensión de los datos

A continuación, se explica en detalle las dos fuentes de información (manejo de cultivo y clima) utilizadas en el presente trabajo de maestría.

La información de manejo de cultivo es proveniente de la finca “Los Naranjos” perteneciente a la empresa Supracafé, ubicada en el municipio de Cajibio, Cauca (21°35’08”N, 76°32’53”W). La finca está compuesta por 38 lotes (Figura 4); cada uno se distingue por la variedad de café sembrado, y cuenta con la siguiente información: número del lote, variedad de café, control, fertilización, limpieza y la producción de café cereza para los años 2012, 2013, 2014, 2016, 2017 y 2018 (Tabla 2). El año 2015 no se tuvo en cuenta debido a que la información de manejo de cultivo no estaba completa y presentaba muchos errores.

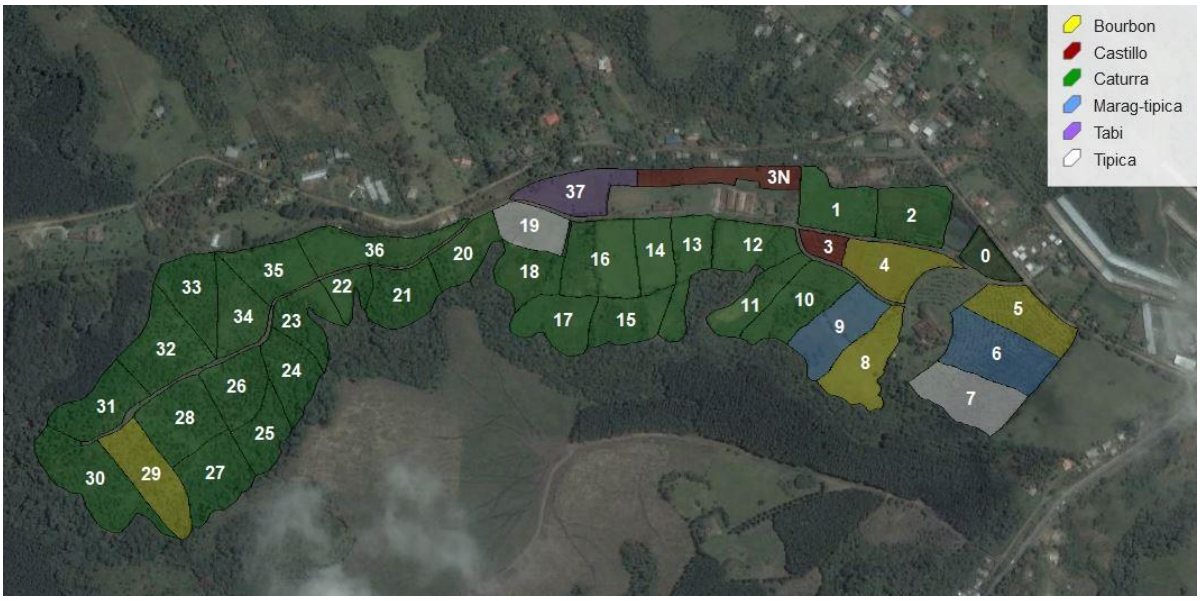

Figura 4. Finca cafetera "Los Naranjos". [Fuente propia]

ATRIBUTO	ABREVIACION	TIPO	DESCRIPCION
Año	Año	Numérico	Año en el que se captura la muestra de producción de café cereza.
Mes	Mes	Nominal	Mes en el que se captura la muestra de producción de café cereza.
Lote	Lote	Numérico	Número del lote.
Edad del cultivo	EdadCultivo	Numérico	Edad del cultivo en meses, calculado a partir de la fecha de siembra.
Densidad	Densidad	Numérico	Densidad de siembra del lote.
Variedad de Café	VariedadCafe	Nominal	Variedad de café.
Control en el mes	Control	Nominal	Actividad de Control en el lote.
Fertilización en el mes	Fertilizacion	Nominal	Actividad de Fertilización en el lote.
Limpieza en el mes	Limpieza	Nominal	Actividad de Limpieza en el lote.
Producción de café cereza	Produccion	Numérico	Producción de café cereza en kilogramos.

Tabla 2. Descripción de los atributos de la información de manejo de cultivo.

La información de manejo de cultivo se diferencia por número de lote, según la distribución de la finca. Cada lote tiene una variedad de café sembrada; en toda la finca existen 6 variedades de café (Figura 4): Bourbon, Castillo, Caturra, Marag-típica, Tabi y Típica. Según la variedad de café sembrada la proporción de la cantidad de producción del lote es diferente, como puede ser que una variedad de café produzca más cantidad también puede ser más susceptible a enfermedades que otras variedades; como también depende el manejo de los caficultores a los cultivos en cuanto a fertilizaciones y controles que deban realizar.

La densidad de siembra indica el número de plantas sembradas por área de terreno. Este atributo tiene un marcado efecto sobre la producción del cultivo y se considera como un insumo, además, la densidad de siembra produce en la planta la competencia de otras plantas de la misma o de otra especie, y también, con una mayor o menor eficiencia de captación de la radiación solar [39].

La edad del cultivo influye de manera muy importante en la producción de café [40], debido a que el cultivo del café tiene un crecimiento en la producción desde el primer año hasta aproximadamente los 5 años, dependiendo del manejo del cultivo por parte del caficultor. Posteriormente a los 5 años tiende a disminuir la producción de café hasta alrededor de los 8 años. De tal manera, que los caficultores toman medidas de

renovación de cultivos a partir de este tiempo para estabilizar la producción de café en sus fincas.

Ahora bien, los cultivos de café necesitan actividades realizadas por humanos con las cuales contrarresten factores de riesgo a los cultivos, como por ejemplo, plagas y enfermedades que ataquen a las plantas de café [40]. Las actividades se pueden clasificar en actividades de fertilización, control y limpieza. Las actividades de fertilización se refieren a la administración de nutrientes o abonos a las plantas, que sirven para fortalecer ya sea las plantas o el suelo donde están sembradas. Las actividades de control, corresponden a la aplicación de productos químicos u orgánicos a las plantas o suelo del cultivo para erradicar plagas o enfermedades presentes en el terreno; en esta práctica influye la variedad de café que está sembrada, debido a que algunas variedades de café son más susceptibles o resistentes a las enfermedades. Finalmente, las actividades de limpieza, hacen referencia a las prácticas que realizan los caficultores para mantener limpio los terrenos de otros cultivos que puedan ser competencia a la hora de absorber nutrientes del suelo. También en estas actividades se incluyen las prácticas de renovación de cultivo, con el objetivo de aumentar la productividad (sujeto a la edad del cultivo, el cual indica si es necesario realizar esta práctica).

Por otra parte, la información climática utilizada en este trabajo se obtuvo entre los años 2012 a 2018; sin embargo como el año 2015 no se tiene información de manejo de cultivo, no se tiene en cuenta los datos climáticos del año 2016, debido a que los datos climáticos influyen aproximadamente un año atrás a la cosecha principal de café. Los datos climáticos se obtienen para toda la región cafetera del departamento del Cauca (Colombia) y son suministrados por Meteoblue (Servicio meteorológico creado en la Universidad de Basilea, Suiza, en cooperación con la Administración Nacional Oceánica y Atmosférica de los Estados Unidos y los Centros Nacionales de Predicción Ambiental). La información climática consta de 4 variables con escala de tiempo horaria (temperatura, humedad relativa, precipitación y radiación). Además, a partir de las variables obtenidas se calcularon las variables: temperatura promedio, temperatura máxima y temperatura mínima (Tabla 3).

ATRIBUTO	ABREVIACION	TIPO	DESCRIPCION
Temperatura promedio.	TempPromedio	Numérico	Temperatura promedio medida desde las 6 hasta las 18 horas del día.
Temperatura Mínima.	TempMinima	Numérico	Temperatura mínima medida desde las 6 hasta las 18 horas del día.
Temperatura Máxima.	TempMaxima	Numérico	Temperatura máxima medida desde las 6 hasta las 18 horas del día.
Precipitación.	Precipitacion	Numérico	Precipitación acumulada del día.
Radiación.	Radiacion	Numérico	Radiación acumulada del día.
Humedad Relativa.	HumedadRelativa	Numérico	Porcentaje de humedad relativa del día.

Tabla 3. Descripción de los atributos de la información climática.

Las variables de temperaturas, precipitación, radiación y humedad relativa son consideradas las variables climáticas más importantes para los sistemas de producción de café [41], esto se debe a que medidas altas o bajas en estas variables afectan el ciclo de vida de la planta y del fruto, teniendo en cuenta la etapa en la que se encuentre el cultivo. Por ejemplo, el déficit hídrico, la alta radiación, las fuertes lluvias (aguacero) esporádicas, el exceso de lluvias, el exceso de nubosidad, las temperaturas medias por encima del óptimo y los cambios bruscos de temperatura, durante el periodo de floración hacen que se interrumpa el crecimiento de las flores de la planta y ocasionando una alta probabilidad de que la planta no germine frutos.

3.2 Preparación de los datos

En esta sección se presenta el procesamiento de los datos, que involucra principalmente el tratamiento de valores faltantes.

Las dos fuentes de información fueron verificadas. La información de manejo de cultivo (ya que son observaciones recolectadas de manera manual), que presentaron diferentes problemas de calidad de datos: valores faltantes, datos duplicados, errores de formato (fechas inconsistentes y ortográficos, *p. e., Maragtipica = marag-típica = Marag Típica*), valores fuera del rango del dominio de la variable.

Los datos climáticos no presentaron errores sintácticos ya que los datos suministrados por Meteoblue son previamente validados por modelos de simulación y estaciones meteorológicas en campo.

3.2.1 Tratamiento de valores faltantes

Teniendo en cuenta que los valores faltantes se presentaron en los datos de manejo de cultivo (Tabla 4), los valores perdidos fueron rellenados a partir de un panel de expertos de la finca “Los Naranjos”, para corroborar la información, a partir de la información registrada por ellos.

FUENTE	ATRIBUTO	PORCENTAJE DE VALORES FALTANTES
Manejo de cultivo	Año	0 %
	Mes	0 %
	Lote	0 %
	EdadCultivo	1 %
	Densidad	0 %
	VariedadCafe	4 %
	Control	2 %
	Fertilizacion	3 %
	Limpieza	2 %
	Produccion	0 %
Clima	TempPromedio	0 %
	TempMinima	0 %
	TempMaxima	0 %
	Precipitacion	0 %
	Radiacion	0 %
	HumedadRelativa	0 %

Tabla 4. Valores faltantes en el conjunto de datos.

Los pasos a seguir para completar la información fueron reuniones con el panel de expertos, donde se realizan las respectivas dudas de los datos faltantes en el conjunto de datos y también validaciones de algunos datos. Sin embargo, en alguno de los casos el panel de experto no contaba con la información en el momento y quedaban de averiguar o corroborar los datos con otras personas que llevaran ese registro.

Particularmente para la información de manejo de cultivo del año 2015, el panel de expertos trato de recopilar esta información, debido a que presentaba mucha inconsistencia y valores faltantes, pero no fue posible porque no se encontraron los registros y además el personal de ese entonces ya no labora actualmente en la finca, A partir de ello, se decide descartar este año para el entrenamiento de los modelos.

3.3 Resumen

En este capítulo se describieron los procesos llevados a cabo para la comprensión y preparación de los conjuntos de datos. En primera instancia se presenta la descripción de las dos fuentes de información para la creación del conjunto de datos que representa la producción de café cereza. De forma seguida, se mencionan los atributos con sus respectivas abreviaciones, tipo y una descripción de cada uno de ellos. Además de la importancia o relevancia que tiene cada uno de los atributos con la producción de café. Finalmente, son mencionadas las actividades llevadas a cabo para procesar los conjuntos de datos antes de ser usados en el proceso de modelado. En conclusión, las dos fuentes de información juegan un papel muy importante en la estimación de la producción de café, debido a que cada uno de los atributos que componen el conjunto de datos impacta la productividad del cultivo, tal como es mencionado anteriormente en la descripción de los atributos utilizados para este trabajo.

Capítulo 4: Modelado

En este capítulo se presentan los modelos estadísticos para la estimación de la producción del café. En primer lugar, se plantean dos series de tiempo que representan la producción de café en diferente escala de tiempo. En segundo lugar, se presentan los modelos estadísticos implementados para la estimación de la producción de café basado en series de tiempo.

4.1 Series de tiempo de producción de café cereza

A partir de las dos fuentes de información (manejo de cultivo y clima), el panel de expertos (administrativos, técnicos e ingenieros agrónomos), e informes técnicos [40], se logró modelar los datos en tres series de tiempo que representan la producción de café en la cosecha principal.

4.1.1 Serie de Tiempo Mensual (STM)

Esta serie de tiempo está construida en escala de tiempo mensual, donde los atributos de manejo de cultivo: Control, Fertilización y Limpieza, se modelan respecto al conocimiento experto, y los meses que más influyen estas actividades en la cosecha principal, son: noviembre y diciembre (año pasado a la cosecha principal) para las actividades de Control (color rojo) y Limpieza (color amarillo); los meses de noviembre (año pasado a la cosecha principal) y febrero (año presente a la cosecha principal) para las actividades de Fertilización (color verde), como se muestra en la Figura 5.

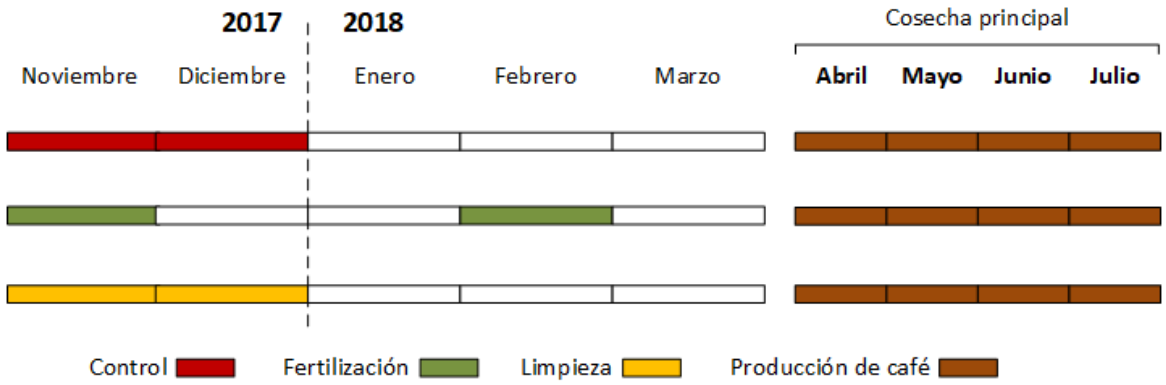

Figura 5. Modelado de los datos de manejo de cultivo para STM. [Fuente propia]

Para la información de clima el modelado es distinto, ya que el clima influye en la producción de café de cada mes de la cosecha principal, en cambio para la información de manejo de cultivo influye de manera directa a los cuatro meses de cosecha. A partir

de las 4 variables de clima se calcularon las temperaturas: mínima, máxima y promedio, como se muestra a continuación:

$$TempPromedio = \frac{\sum_{i=hInicio}^{hFinal}(t_i)}{(hFinal - hInicio) + 1}$$

Ecuación 1. Temperatura promedio.

$$TempMinima = Min(t_{hInicio}, t_{hInicio+1}, .., t_i, .., t_{hFinal})$$

Ecuación 2. Temperatura mínima.

$$TempMaxima = Max(t_{hInicio}, t_{hInicio+1}, .., t_i, .., t_{hFinal})$$

Ecuación 3. Temperatura máxima.

Donde,

$hInicio$: Hora de inicio del día.

$hFinal$: Hora final del día.

t: Temperatura en una hora del día.

i: Hora del día.

Los parámetros $hInicio$ y $hFinal$, fueron ajustados acorde al panel de expertos; donde la hora inicial del día se considera a las 6 y la hora final del día las 18. Esto debido a que las temperaturas mínimas, máximas y promedio no se consideran en las franjas de la madrugada, ni en las noches, debido a que no representarían las temperaturas ideales del día.

Las variables climáticas fueron procesadas para siete meses antes de una ventana de tiempo de 4 meses, anterior a la cosecha. Por ejemplo, para la producción de café (color café) en el mes de Abril del año 2018, el clima (color azul) influyente es el de los meses de Mayo (2017) hasta el mes de noviembre (2017). Así que esta franja de tiempo de 7 meses se va trasladando para cada mes de la cosecha principal, conservando la ventana de tiempo de 4 meses, como se puede observar en la Figura 6.

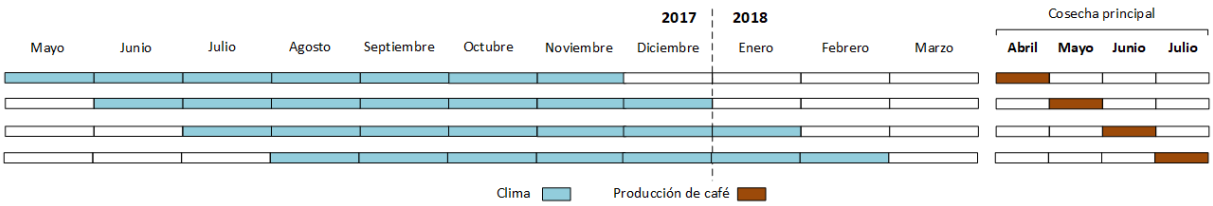

Figura 6. Modelado de los datos de clima mensual. [Fuente propia]

De esta manera la serie de tiempo STM, está conformada por 54 atributos (12 de manejo de cultivo y 42 de clima), los cuales se describen en la Tabla 5.

ATRIBUTO	TIPO	DESCRIPCION
Año	Numérico	Año en el que se captura la muestra de producción de café cereza.
Mes	Nominal	Mes en el que se captura la muestra de producción de café cereza.
Lote	Numérico	Número del lote.
EdadCultivo	Numérico	Edad del cultivo en meses, calculado a partir de la fecha de siembra.
Densidad	Numérico	Densidad de siembra del lote.
VariedadCafe	Nominal	Densidad de siembra del lote.
Control(noviembre)	Nominal	Actividad de Control en el lote en el mes de noviembre del año pasado a la cosecha principal.
Control(diciembre)	Nominal	Actividad de Control en el lote en el mes de diciembre del año pasado a la cosecha principal.
Fertilizacion(noviembre)	Nominal	Actividad de Fertilización en el lote en el mes de noviembre del año pasado a la cosecha principal.
Fertilizacion(febrero)	Nominal	Actividad de Fertilización en el lote en el mes de febrero del año presente a la cosecha principal.
Limpieza(noviembre)	Nominal	Actividad de Limpieza en el lote en el mes de noviembre del año pasado a la cosecha principal.
Limpieza(diciembre)	Nominal	Actividad de Limpieza en el lote en el mes de diciembre del año pasado a la cosecha principal.
TempPromedio(meses)	Numérico	Temperatura promedio de los meses 11, 10, 9, 8, 7, 6 y 5 con una ventana de tiempo de 4 meses antes de la cosecha. Es decir 7 atributos.
TempMinima(meses)	Numérico	Temperatura mínima de los meses 11, 10, 9, 8, 7, 6 y 5 con una ventana de tiempo de 4 meses antes de la cosecha. Es decir 7 atributos.
TempMaxima(meses)	Numérico	Temperatura máxima de los meses 11, 10, 9, 8, 7, 6 y 5 con una ventana de tiempo de 4 meses antes de la cosecha. Es decir 7 atributos.
Precipitacion(meses)	Numérico	Precipitación de los meses 11, 10, 9, 8, 7, 6 y 5 con una ventana de tiempo de 4 meses antes de la cosecha. Es decir 7 atributos.
Radiacion(meses)	Numérico	Radiación de los meses 11, 10, 9, 8, 7, 6 y 5 con una ventana de tiempo de 4 meses antes de la cosecha. Es decir 7 atributos.
HumedadRelativa(meses)	Numérico	Humedad relativa de los meses 11, 10, 9, 8, 7, 6 y 5 con una ventana de tiempo de 4 meses antes de la cosecha. Es decir 7 atributos.

Tabla 5. Atributos de la serie de tiempo STM.

En la Tabla 5, se presentan los 6 atributos básicos de clima. Por cada atributo básico, se calculan 7 atributos de clima (para un total de 42 atributos climáticos), como se explica en la columna descripción de la Tabla 5. Para mayor comprensión se puede ver el Anexo 4, donde se encuentra toda la información de los atributos de la serie de tiempo STM.

4.1.2 Serie de Tiempo Mensual por Variedad de Café (STMVC)

Esta serie de tiempo sigue el mismo proceso como fue construida STM, con la diferencia que se dividió toda la serie en 6 sub-series, dependiendo de la variedad de café. Es decir, que son utilizados los mismos atributos de clima y de manejo de cultivo, como se observa en la Tabla 6.

SUB-SERIE	NUMERO DE ATRIBUTOS	NUMERO DE INSTANCIAS
Caturra	54	536
Castillo		20
Típica		36
Marag-típica		40
Bourbon		80
Tabi		20

Tabla 6. Descripción de la serie de tiempo STMVC.

Por otra parte, la sub-serie de Caturra contiene más muestras, debido a que la mayor parte de los lotes de la finca están sembrados de esta variedad de café.

4.1.3 Serie de Tiempo Semanal (STS)

La serie de tiempo STS, está conformada por los mismos atributos de la serie de tiempo STM, pero con escala semanal. Es decir, se separaron las actividades de manejo de cultivo en semanas y además pasaron de ser atributos nominales a atributos numéricos, donde se cuenta el número de actividades realizadas en la semana.

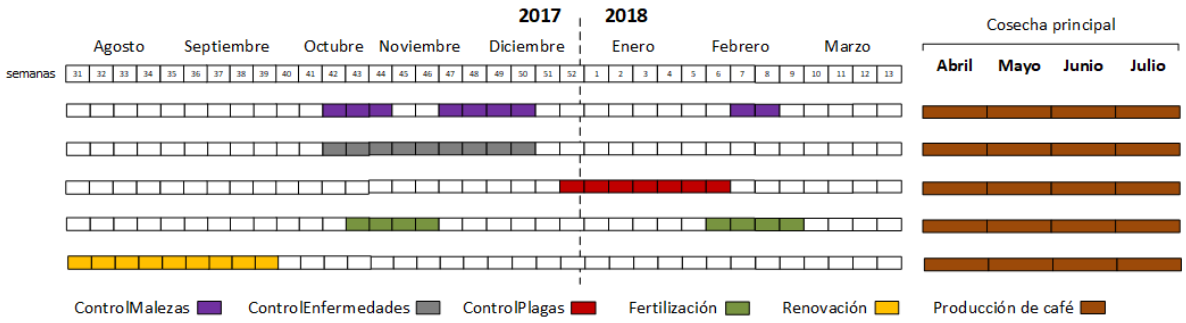

Figura 7. Modelado de los datos de manejo de cultivo para STS. [Fuente propia]

En la Figura 7, se puede observar el modelado de la información de manejo de cultivo respecto a los meses de la cosecha principal de café. También cabe aclarar que el atributo *Control* que se tenía en la serie de tiempo STM, es dividido en los atributos *ControlMalezas*, *ControlEnfermedades* y *ControlPlagas*. De igual forma el atributo *Limpieza* de STM, corresponde al atributo *Renovación* en la serie de tiempo STS. Para estos nuevos atributos se evaluó nuevamente la influencia que tienen respecto a la cosecha principal con el conocimiento experto. En la Figura 7 se puede observar que para esta serie de tiempo se tiene en cuenta aproximadamente un año atrás, es decir, a 51 semanas antes, con una ventana de tiempo de una semana antes de la cosecha principal.

El modelado de la información climática para la serie de tiempo STS, fue de igual forma a la serie de tiempo STM, pero con escala de tiempo semanal. Utilizando las mismas ecuaciones (Ecuación 1**,** Ecuación 2 y Ecuación 3), pero con la diferencia que se calcularon para cada semana del año.

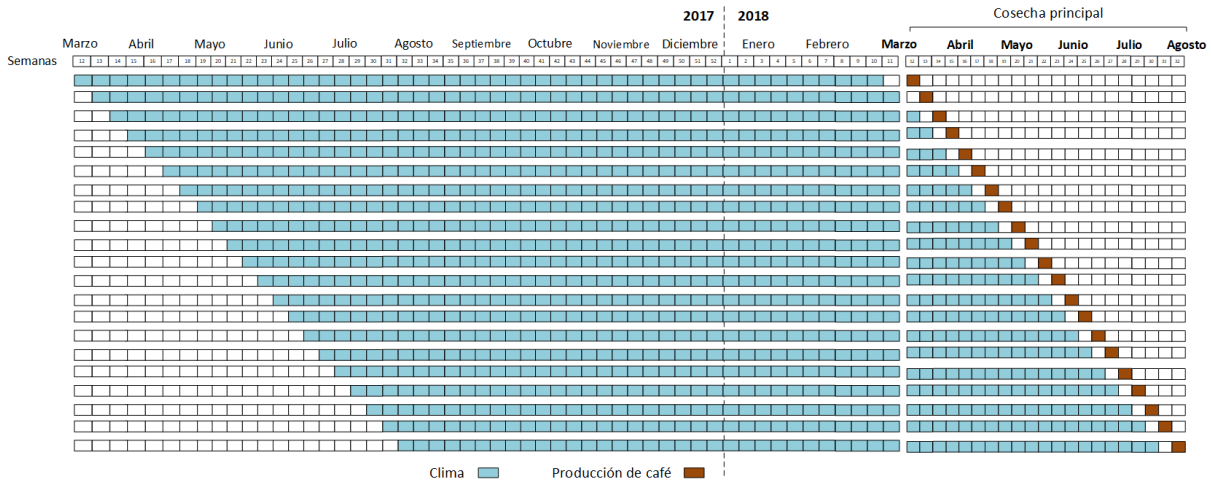

Figura 8. Modelado de los datos de clima semanal.

En la Figura 8, se puede observar el modelamiento de la información climática respecto a la información de la producción de café en la cosecha principal. Esta serie de tiempo STS, permite aumentar el número de instancias a 107 respecto a la serie de tiempo STM. De esta manera la serie de tiempo STS, está conformada por 318 atributos (12 de manejo de cultivo y 306 de clima), los cuales se describen en la Tabla 7.

ATRIBUTO	TIPO	DESCRIPCION
Año	Numérico	Año en el que se captura la muestra de producción de café cereza.
Mes	Nominal	Mes en el que se captura la muestra de producción de café cereza.
Semana	Numérico	Semana en la que se captura la muestra de producción de café cereza.
ControlMalezas(42-44)_a	Numérico	Número de actividades de control de malezas en las semanas 42 a la 44 del año pasado a la cosecha principal.
ControlMalezas(47-50)_a	Numérico	Número de actividades de control de malezas en las semanas 47 a la 50, del año pasado a la cosecha principal.
ControlMalezas(7-8)_p	Numérico	Número de actividades de control de malezas en las semanas 7 y 8, del año presente a la cosecha principal.
ControlEnfermedades(42-50)_a	Numérico	Número de actividades de control de enfermedades en las semanas 42 a la 50, del año pasado a la cosecha principal.
ControlPlagas(52)_a	Numérico	Número de actividades de control de plagas en la semana 52, del año pasado a la cosecha principal.
ControlPlagas(1-6)_p	Numérico	Número de actividades de control de plagas en las semanas 1 a la 6, del año presente a la cosecha principal.
Fertilizacion(43-46)_a	Numérico	Número de actividades de fertilización en las semanas 43 a la 46, del año pasado a la cosecha principal.
Fertilizacion(6-9)_p	Numérico	Número de actividades de fertilización en las semanas 6 a la 9, del año presente a la cosecha principal.
Renovacion(31-39)_a	Numérico	Número de actividades de renovación en las semanas 31 a la 39, del año pasado a la cosecha principal.
TempPromedio(semanas)	Numérico	Temperatura promedio de la semana 51 hasta la semana 2, a la semana de cosecha. Es decir 51 atributos.
TempMinima(semanas)	Numérico	Temperatura mínima de la semana 51 hasta la semana 2, a la semana de cosecha. Es decir 51 atributos.
TempMaxima(semanas)	Numérico	Temperatura máxima de la semana 51 hasta la semana 2, a la semana de cosecha. Es decir 51 atributos.
Precipitacion(semanas)	Numérico	Precipitación de la semana 51 hasta la semana 2, a la semana de cosecha. Es decir 51 atributos.
Radiacion(semanas)	Numérico	Radiación de la semana 51 hasta la semana 2, a la semana de cosecha. Es decir 51 atributos.
HumedadRelativa(semanas)	Numérico	Humedad relativa de la semana 51 hasta la semana 2, a la semana de cosecha. Es decir 51 atributos.

Tabla 7. Atributos de la serie de tiempo STS.

En la Tabla 7, se presentan los 6 atributos básicos de clima. Por cada atributo básico, son calculados 51 atributos (para un total de 306 atributos climáticos), como se puede explicar en la columna descripción de la Tabla 7. Para mayor comprensión se puede ver el Anexo 4, donde se encuentra toda la información de los atributos de la serie de tiempo STS.

4.2 Modelos estadísticos para la estimación de la producción de café cereza basada en Series de Tiempo

En esta sección se describen los modelos estadísticos utilizados para la estimación de la producción de café cereza basada en las series de tiempo descritas en la sección anterior. En el siguiente capítulo se explican los resultados obtenidos por cada modelo.

4.2.1 Tree Regressor (TreeRegresor)

Es un modelo lineal basado en árboles [42]. Algunas de las características importantes de este modelo son las siguientes: los requisitos computacionales crecen de una manera que puede abordar problemas con alta dimensionalidad, y los árboles generados son más pequeños que otros métodos como CART y MARS.

4.2.2 Linear Regression (LR)

La regresión lineal es utilizada para describir el comportamiento de una variable numérica dependiente, en función de un conjunto de variables independientes [43]. La LR es explicada a través de coeficientes agregados a cada variable independiente mediante el principio de correlación lineal.

4.2.3 Artificial Network Neural (ANN)

Las redes neuronales se componen de nodos (llamados neuronas) organizados en capas: entrada, ocultas y de salida [44]. En este modelo cada neurona dentro de una capa está conectada a cada neurona de la siguiente capa. Adicionalmente cada conexión entre neurona tiene asignado un peso, que se establecen a lo largo de un proceso de aprendizaje (algoritmo de aprendizaje de retro propagación).

4.2.4 Extreme Gradient Boosting (XGBoost)

Es uno de los algoritmos más dominantes en el aprendizaje automático [45], está basado en arboles de decisión que a la vez implementa un marco potenciado de gradientes. Cuando se van a tratar datos estructurados de tamaño pequeño o mediano, este algoritmo se considera que obtiene mejores resultados que otros.

También una de las ventajas de este algoritmo es el bajo consumo de memoria y tiempo de cómputo en su proceso de entrenamiento.

4.2.5 Random Forest (RForest)

Es un modelo de árboles de decisión combinado con el método de *Bagging* [46]. *Bagging* es un método que consiste en que distintos árboles ven diferentes porciones de los datos; ningún árbol entrena con todos los datos. Esto hace que cada árbol se entrene con diferentes muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, los errores se compensan con otros y se tiene una estimación que generaliza mejor.

4.2.6 Support Vector Machine para Regresión (SVR)

Las máquinas de vector de soporte de regresión parte de los mismos principios que la máquina de vector de soporte para clasificación (SVM): trata de encontrar una función que aproxime los datos de entrenamiento con el menor error [47]. SVR pretende encontrar la distancia adecuada entre los vectores de soporte y el hiperplano óptimo de decisión, que reúna el mayor número de elementos

4.3 Resumen

En este capítulo se describieron las tres series de tiempo que representan la producción de café a partir de información de manejo de cultivo e información climática. Primero, la serie de tiempo que estaba basada en información mensual. Segundo, la serie de tiempo basada en información mensual, pero discriminada por variedad de café. Y la tercera serie de tiempo basada en información semanal, aumentando el número de instancias respecto a las otras dos series de tiempo. Luego, se explican los modelos estadísticos utilizados para la estimación de producción de café cereza basada en las series de tiempo anteriormente mencionadas.

Capítulo 5: Experimentación y evaluación

En este capítulo se presenta el proceso de experimentación y las pruebas ejecutadas sobre la estimación de la producción de café cereza basada en series de tiempo. Primero se describen las métricas de evaluación utilizadas para evaluar los resultados obtenidos por los modelos estadísticos. Segundo, el plan de pruebas de los experimentos realizados para la estimación de producción de café cereza. Finalmente, los resultados obtenidos.

5.1 Métricas de evaluación

En esta sección se presentan las métricas de evaluación utilizadas para evaluar los modelos estadísticos en cada uno de los experimentos planteados en este trabajo de investigación.

5.1.1 Error absoluto medio (MAE)

Esta métrica de evaluación la media entre los valores estimados y reales del modelo [48]. Esta medida está dada por la Ecuación 4.

$$MAE = \frac{1}{n} * \sum_{1}^{n} (valor\_estimado - valor\_real)$$

Ecuación 4. Error absoluto medio.

Donde,

n: número de muestras.

$valor\_estimado$: valor estimado de la producción de café cereza por el modelo.

$valor\_real$: valor real de la producción de café cereza.

5.1.2 Error cuadrático medio (RMSE)

El error cuadrático medio, mide el promedio de los errores al cuadrado entre los valores estimados y reales del modelo [48]. Esta medida está dada por la Ecuación 5.

$$RMSE = \sqrt{\frac{\sum_{1}^{n}(valor\_estimado - valor\_real)^2}{n}}$$

Ecuación 5. Error cuadrático medio.

Donde,

n: número de muestras.

$valor\_estimado$: valor estimado de la producción de café cereza por el modelo.

$valor\_real$: valor real de la producción de café cereza.

5.1.3 Error cuadrático relativo (RSE)

El error cuadrático relativo [48], mide el promedio de los errores. Esta medida toma el error cuadrático total y lo normaliza dividiéndolo por el error cuadrático total. Esta medida está representada por la Ecuación 6.

$$RSE = \frac{\sum_1^n (valor\_estimado - valor\_real)^2}{\sum_1^n (valor\_estimado - media\_valor\_real)^2}$$

Ecuación 6. Error cuadrático relativo.

Donde,

n: número de muestras.

$valor\_estimado$: valor estimado de la producción de café cereza por el modelo.

$valor\_real$: valor real de la producción de café cereza.

$media\_valor\_real$: media aritmética del valor real de la producción de café cereza.

5.2 Plan de pruebas

En esta sección se presentan los experimentos realizados a partir de las tres series de tiempo planteadas (sección 4.1), los experimentos están organizados como se muestran en la Tabla 8. En total son 6 experimentos, en los cuales se fue manipulando las series de tiempo e implementando métodos de selección de características.

Se utilizaron dos métodos de selección de características para encontrar las variables independientes más representativas: un método de filtro de correlación de Pearson (PC) y eliminación de características recursivas (RFE) [49].

- El método de filtro de PC utiliza la correlación de pearson para seleccionar el mejor subconjunto de variables del valor absoluto más alto de correlación entre una variable independiente y una dependiente. Se descartaron las variables con un coeficiente de correlación de pearson $<= 0.2$ (ver Anexo 3).
- RFE selecciona un subconjunto de variables comenzando con todas las características en el conjunto de datos de entrenamiento y eliminando características de manera recursiva hasta obtener la mejor calidad de predicción, utilizando validación cruzada con 10 folds.

EXPERIMENTO	SERIE DE TIEMPO	MODELO ESTADISTICO	SELECCIÓN DE ATRIBUTOS	NUMERO DE ATRIBUTOS
E-01	STM	TreeRegresor	Todos los atributos	54
		LR		
		ANN		
		SVR		
E-02	STMVC	TreeRegresor	Todos los atributos	54
		LR		
		ANN		
		SVR		
E-03	STS	TreeRegresor	Todos los atributos	318
		ANN		
		XGBoost		
		SVR		
		RForest		
E-04	STS	TreeRegresor	PC	112
		ANN		
		XGBoost		
		SVR		
		RForest		
E-05	STS	TreeRegresor	PC + Clase objetivo normalizada	112
		ANN		
		XGBoost		
		SVR		
		RForest		
E-06	STS	TreeRegresor	RFE + Clase objetivo normalizada	51
		ANN		
		XGBoost		
		SVR		
		RForest		

Tabla 8. Plan de pruebas.

El experimento E-01 considera todas las características de STM con una escala de tiempo mensual. Luego, para aumentar el número de instancias y características, se

construyó la serie de tiempo STMVC para realizar el experimento E-02 considerando todas las características. Para el experimento E-03, se implementó la serie de tiempo STS que organiza la información en escala de tiempo semanal. En el experimento E-03 se aumentan el número de instancias (107) y atributos (318) a comparación de los experimentos anteriores. Debido a los resultados desfavorables obtenidos en los experimentos E-01 y E-02 con el algoritmo LR (ver Anexo 1 y 2), se decidió descartar dicho algoritmo en los siguientes experimentos, y se agregaron los algoritmos XGBOOST y RF.

En el experimento E-04, se aplicó un método de selección de características con el objetivo de disminuir la alta dimensionalidad de la serie de tiempo. Se utilizó un criterio de PC para seleccionar las características relevantes en función de la correlación más alta entre la variable dependiente y las variables independientes. A partir del experimento E-04, para reducir los errores de evaluación, se normalizó la variable objetivo en el experimento E-05 (es decir, la producción de café cereza se transforma de tal manera que los valores están entre 0 y 1). Finalmente, el experimento E-06 con la variable dependiente normalizada y la técnica RFE, 51 variables independientes fueron seleccionadas.

5.3 Resultados

5.3.1 Experimento E-01

Los mejores resultados se obtuvieron mediante el modelo TreeRegresor para las series de tiempo de los lotes 23, 28 y 29 como se puede observar en la Tabla 9. Según estos resultados se puede concluir que los árboles de regresión funcionan mejor que el resto de modelos en este tipo de experimentos con series de tiempo que contienen pocas instancias (alrededor de 20 instancias).

LOTE	MODELO ESTADISTICO	MAE	RMSE	RSE	PRODUCCION DE CAFÉ CEREZA (Kg)	
					REAL	ESTIMACION
23	SVR	557.35	646.46	72.1189	5877	6056.81
28	ANN	553.77	615.86	68.70	7459	7425.33
29	TreeRegresor	439.55	490.63	99.36	6421	6412.48

Tabla 9. Métricas de evaluación para E-01.

En la Figura 9 se puede observar el mejor resultado obtenido por el modelo TreeRegresor para el lote número 29. La diferencia de la producción total real de la

cosecha principal del año 2017 con respecto a la estimada es de 8.5 kg. Sin embargo, los valores de la producción estimada para cada mes presentan una mayor diferencia.

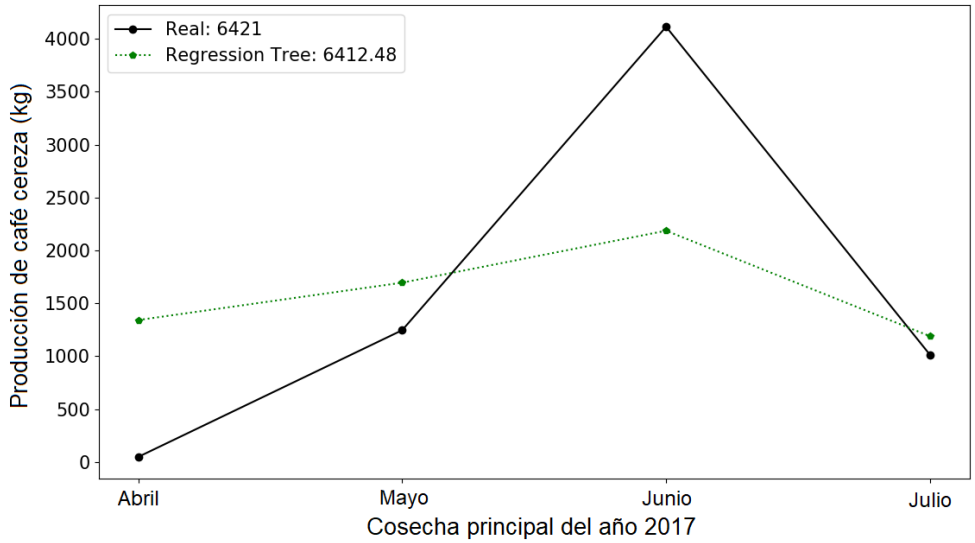

Figura 9. Resultados experimento E-01.

5.3.2 Experimento E-02

Para este experimento, los modelos TreeRegresor y ANN proporcionaron los mejores resultados en comparación con los modelos SVR y LR (Tabla 10). La serie de tiempo creada para la variedad de café bourbon, incremento el número de instancias, lo que hizo que el modelo TreeRegresor obtuviera resultados satisfactorios para estas métricas.

VARIEDAD DE CAFE	LOTE	MODELO ESTADITICO	MAE	RMSE	RSE	PRODUCCION DE CAFÉ CEREZA (Kg)	
						REAL	ESTIMACION
Castillo	3	TreeRegresor	3136.74	3443.45	370.76	5595	18141.96
Tipica	7	ANN	491.10	627.93	186.53	2098	2093.51
Marag-tipica	9	ANN	1360.98	1658.43	172.00	5624	6239.32
Caturra	28	ANN	531.53	536.32	35.90	7459	7425.33
Bourbon	29	TreeRegresor	961.41	1185.21	78.19	6421	6412.47
Tabi	37	ANN	921.15	1115.83	104.86	6246	6524.41

Tabla 10. Métricas de evaluación para E-02.

El uso de modelos ANN también proporcionó resultados satisfactorios incluso para las series de tiempo con pocas instancias (Tabi, Marag-tipica). Como muestra la Figura 10, el mejor resultado para este experimento se obtuvo con el modelo ANN con la serie

temporal de la variedad de café de caturra para el lote número 28. La diferencia entre la estimación y la producción real de cada mes disminuyó con respecto al experimento E-01. Sin embargo, al calcular la cantidad total de la cosecha, la diferencia fue mayor (33,60 kg).

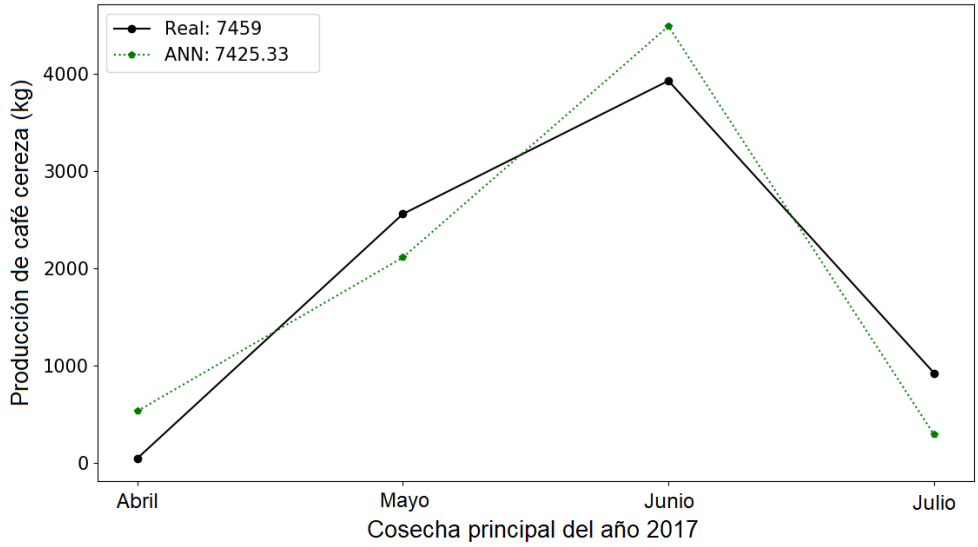

Figura 10. Resultados experimento E-02.

Cabe aclarar que la implementación de los modelos XGBoost y RForest no mejoraron los resultados descritos para los experimentos E-01 y E-02.

5.3.3 Experimento E-03

Los modelos XGBoost y RForest se implementaron en el experimento E-03 dado su uso generalizado para tareas similares. El modelo LR fue descartado debido los resultados desfavorables obtenidos en los experimentos E-01 y E-02 (Ver Anexo 1 y 2). La Tabla 11, muestra que los modelos XGBoost y RForest proporcionan la diferencia más pequeña de la producción estimada con la real.

MODELO ESTADISTICO	MAE	RMSE	RSE	PRODUCCION DE CAFÉ CEREZA (Kg)	
				REAL	ESTIMACION
TreeRegresor	3068.5	94165500.33	0.05	180672	92190
ANN	4760.39	52139096.81	0.474	180672	374868
XGBoost	2752.00	46149548.01	0.535	180672	193920
SVR	2815.06	128383997.75	-0.295	180672	339849
RForest	3043.87	39196021.92	0.605	180672	222719

Tabla 11. Métricas de evaluación para E-03.

En la Figura 11, se pueden observar las estimaciones proporcionadas por los modelos XGBoost y RForest, que son las más similares a la curva de la producción real de la cosecha principal del año 2018. Los modelos ANN y SVR, por otro lado, presentan resultados superiores a los valores esperados de producción de café. Finalmente, el modelo TreeRegresor proporciona resultados por debajo de la curva.

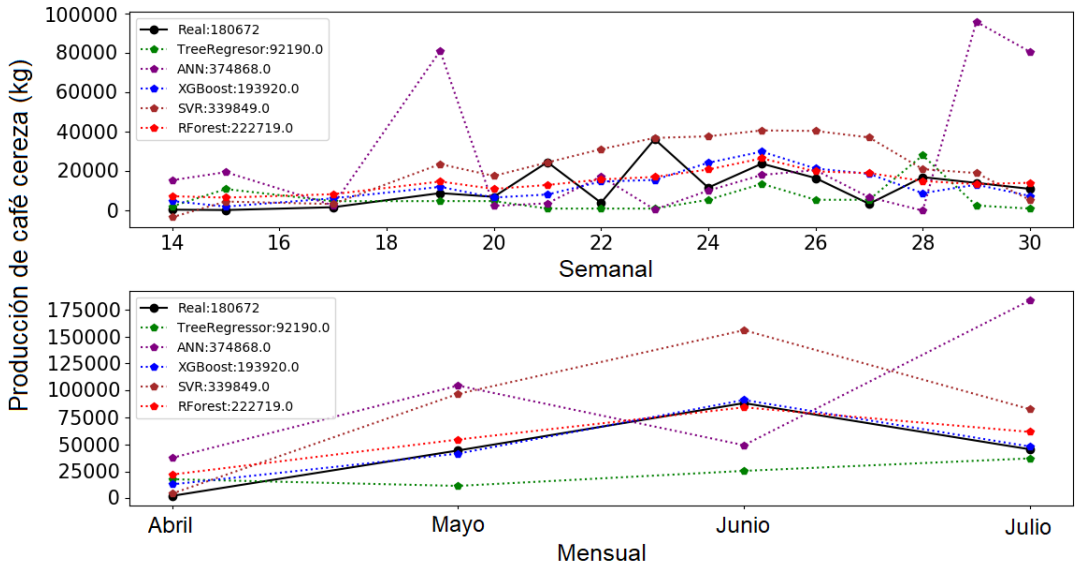

Figura 11. Resultados experimento E-03.

5.3.4 Experimento E-04

En este experimento, el conjunto de datos se redujo a 112 atributos, debido al uso del método de filtro para la selección de características. Los resultados de las métricas de evaluación para los modelos XGBoost y RForest, son mejores que los resultados proporcionados por el resto de modelos. Sin embargo, los errores MAE, RMSE y RSE siguen siendo altos (Tabla 12). Adicionalmente, el modelo TreeRegresor mejoró sus resultados utilizando menos atributos que en el experimento anterior.

MODELOS ESTADISTICOS	MAE	RMSE	RSE	PRODUCCION DE CAFÉ CEREZA (Kg)	
				REAL	ESTIMACION
TreeRegresor	6147.00	178105397.43	-0.796	180672	311203
ANN	4760.39	52139096.81	0.474	180672	304696
XGBoost	2845.28	45467185.29	0.541	180672	192599
SVR	5566.33	104603881.97	-0.055	180672	322665
RForest	3257.40	39889453.89	0.598	180672	227854

Tabla 12. Métricas de evaluación para E-04.

Las estimaciones de producción de los modelos mejoran con respecto al experimento E-03, como se puede observar en la Figura 12. Sin embargo, las estimaciones generadas por XGBoost y RForest aún brindan los mejores resultados para la producción de café cereza en la cosecha principal del año 2018.

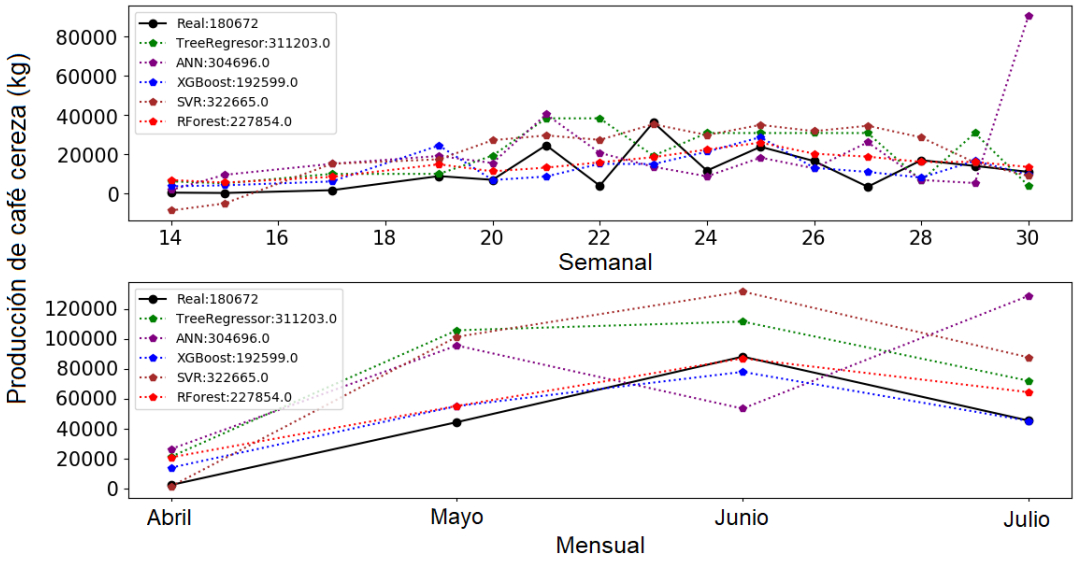

Figura 12. Resultados experimento E-04.

5.3.5 Experimento E-05

Los resultados obtenidos por las métricas de evaluación disminuyeron considerablemente en los experimentos E-01, E-02, E-03 y E-04. Adicionalmente, los modelos XGBoost y RForest también obtienen mejores resultados comparados con el resto de modelos estadísticos en el experimento E-05. Las estimaciones de los 5 modelos estadísticos (excepto SVR) mejoran con respecto a los experimentos anteriores.

MODELOS ESTADISTICOS	MAE	RMSE	RSE	PRODUCCION DE CAFÉ CEREZA (Kg)	
				REAL	ESTIMACION
TreeRegresor	0.065	0.014	0.262	180672	133141
ANN	0.065	0.010	0.474	180672	214668
XGBoost	0.032	0.008	0.585	180672	179637
SVR	0.091	0.026	-0.429	180672	357519
RForest	0.044	0.007	0.597	180672	235800

Tabla 13. Métricas de evaluación para E-05.

Los resultados obtenidos por el modelo XGBoost están muy cerca de la producción real de café cereza de la cosecha principal del año 2018 (Figura 14) y es de destacar, que los errores disminuyen respecto a los experimentos anteriores (Tabla 13).

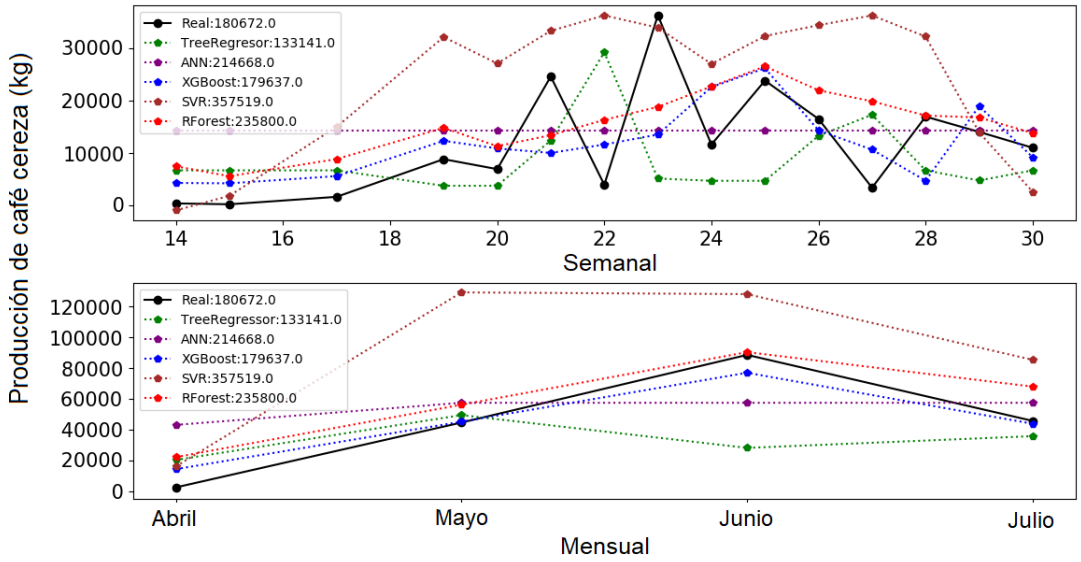

Figura 13. Resultados experimento E-05.

5.3.6 Experimento E-06

La Tabla 14 muestra que los resultados de las métricas de evaluación para el experimento E-06 son muy similares a los resultados descritos para los experimentos anteriores. Los modelos XGBoost y RForest proporcionan la estimación más cercana a la producción real de café cereza del año 2018.

MODELOS ESTADISTICOS	MAE	RMSE	RSE	PRODUCCION DE CAFÉ CEREZA (Kg)	
				REAL	ESTIMACION
TreeRegresor	0.068	0.033	-0.799	180672	234594
ANN	0.087	0.031	-0.66	180672	214668
XGBOOST	0.046	0.01	0.456	180672	186110
SVR	0.059	0.008	0.546	180672	264720
RForest	0.043	0.008	0.56	180672	213317

Tabla 14. Métricas de evaluación para E-06.

Hay una diferencia de aproximadamente 1 mil kilogramos para el modelo XGBoost y 55 mil kilogramos para el modelo RForest, como se muestra en la Figura 14. La

diferencia entre el valor estimado y el real es bastante grande, debido a que los modelos no consideran variables de pérdida de granos de café cereza en la etapa de “Recolección” de la cadena de valor del café (Figura 1).

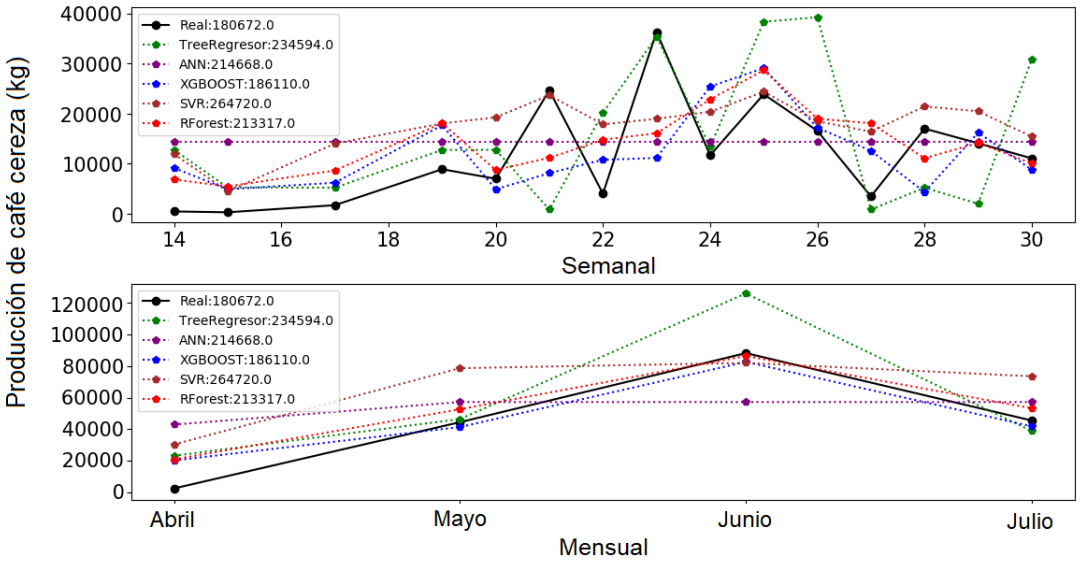

Figura 14. Resultados experimento E-06.

Según un estudio publicado por CENICAFE [35], la pérdida media en la cosecha tradicional es de 4.46 granos de café por planta. Sin embargo, un cultivo de café tiene aproximadamente de 5 mil a 10 mil plantas de café por hectárea. Además, 1 kilogramo de café maduro (cereza) tiene un promedio de 555 granos de café.

Con respecto a las estimaciones que brindan los diferentes modelos estadísticos para la producción de café en la cosecha 2018, se encuentran bastante cerca del valor real. Sin embargo, la diferencia del mejor resultado obtenido por el algoritmo XGBoost es mayor que el valor correspondiente obtenido para el experimento E-05. La Figura 14 muestra que los algoritmos XGBoost y RForest proporcionan resultados satisfactorios también considerando que el número de atributos ha disminuido considerablemente.

Como se describió anteriormente, los mejores resultados se han obtenido para los experimentos E-05 y E-06 dado que la clase objetivo se normalizó para reducir los errores observados en los experimentos E-01, E-02, E-03 y E-04. También, luego de analizar las métricas de evaluación en los experimentos E-05 y E-06. Los modelos TreeRegresor, ANN, XGBoost y RForest proporcionan los valores más bajos para el error RMSE en el experimento E-05 (Figura 15a). Sin embargo, en ambos experimentos, los modelos XGBoost y RForest proporcionan errores bajos

independientemente del método de selección de características utilizado en el proceso de evaluación.

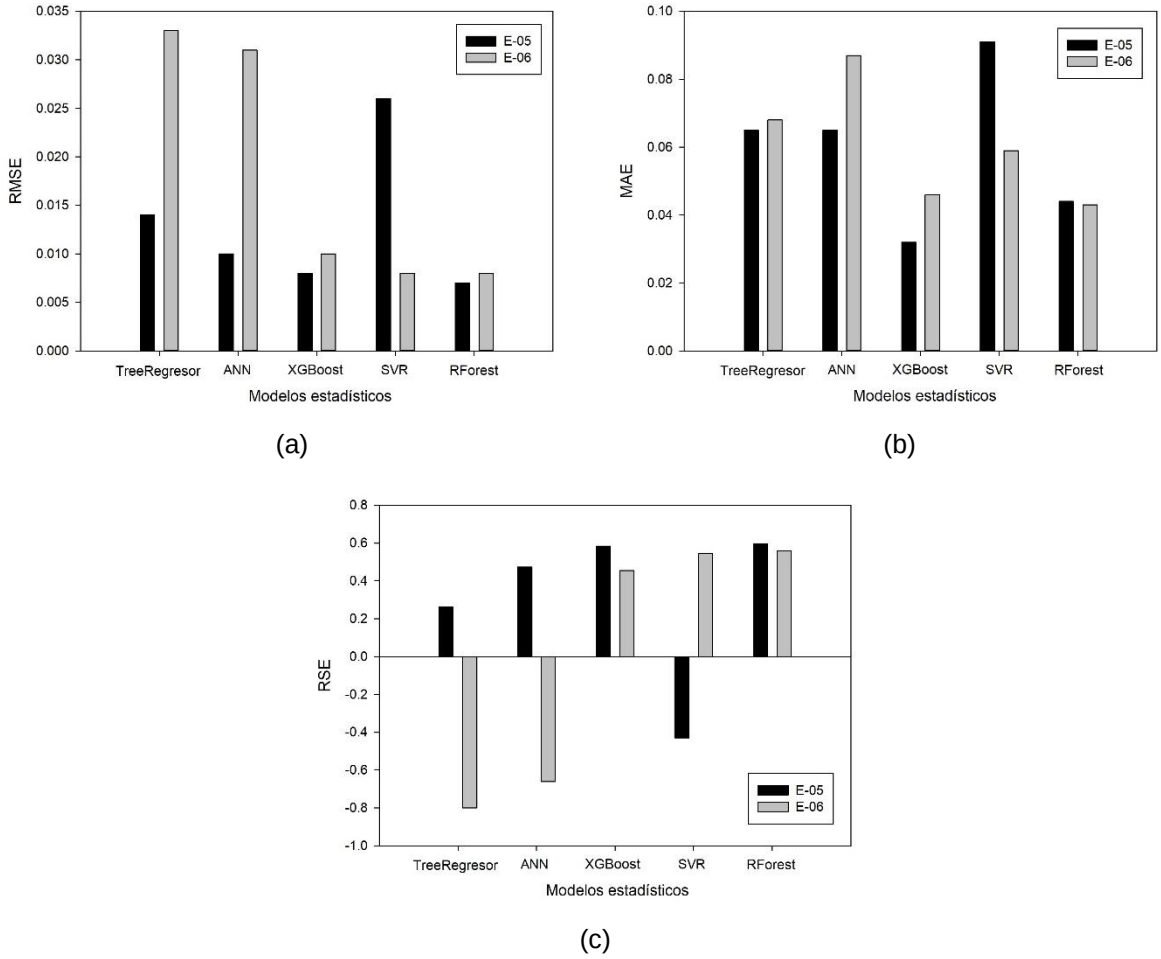

Figura 15. Métricas de evaluación para los experimento E-05 y E-06.

5.4 Resumen

En este capítulo se realizaron seis experimentos para evaluar tanto las series de tiempo, como los modelos estadísticos propuestos en este trabajo de maestría. El experimento E-01, consiste en implementar los modelos estadísticos con la serie de tiempo mensual, que es relevante a los datos de clima y manejo de cultivo. El experimento E-02, se diferencia en que la serie de tiempo es discriminada por variedad de café. El experimento E-03, se aumenta el volumen de la serie de tiempo, pasando

de escala mensual a semanal. El experimento E-04, utiliza la serie de tiempo semanal pero se implementa un método de filtro basado en la correlación de Pearson (PC) para la selección de atributos. El experimento E-05, es tal cual el experimento E-04 pero adicionalmente se normaliza la variable objetivo (producción de café cereza). Y finalmente, el experimento E-06, al igual que el E-05 se normaliza la variable objetivo, pero se utiliza eliminación de atributos recursiva (RFE) para la selección de atributos.

Las métricas utilizadas en los distintos experimentos fueron: error absoluto medio (MAE), error medio cuadrático (RMSE) y error cuadrático relativo (RSE).

Los resultados de los experimentos E-01 y E-02 son prometedores por la aproximación de la producción estimada a la real, pero el análisis de los valores de las métricas de evaluación no son muy buenos; estos dos experimentos están atados a la poca información debido a la escala mensual de las series de tiempo. En cambio, los experimentos E-03, E-04 y E-05 obtuvieron mejores valores en las métricas de evaluación y por consiguiente la diferencia de la producción real de café es menor a la estimada como se pueden observar en la figuras donde se encuentran las curvas de la producción real de café cereza y la estimada en cada experimento por los modelos estadísticos para la cosecha principal del año 2018. Finalmente, el experimento E-06 es el más destacado por sus buenos resultados y obtener la curva más similar a la real en cuanto a la producción de café cereza, eso también lo muestra en los resultados logrados por las métricas de evaluación.

Capítulo 6: Prototipo

En este capítulo se presenta un prototipo que implementa la estimación de café cereza basada en series de tiempo. Para la construcción del prototipo se utilizó el marco de trabajo ágil de desarrollo de software SCRUM [50].

6.1 Marco de trabajo ágil SCRUM

De acuerdo al marco de trabajo ágil SCRUM, se han definido las siguientes historias de usuario que para el desarrollo del prototipo (Tabla 16). Cabe mencionar que el prototipo esta implementado dentro de la plataforma del proyecto IoT-Agro, el cual tiene otros módulos diferentes a la estimación de la producción de café cereza.

#	HISTORIA DE USUARIO (HU)	PRIORIDAD	DESCRIPCION
1	Definición de arquitectura	Alta	Realizar la respectiva investigación de las tecnologías necesarias para el procesamiento y almacenamiento de la información, también los recursos software para el desarrollo de los componentes y módulos requeridos de la plataforma.
2	Almacenamiento de la información	Alta	Crear la estructura adecuada para el almacenamiento de la información.
3	Diseño de interfaces	Media	Definir las respectivas vistas y diseños de la plataforma para manejar una ambigüedad en la navegación de la plataforma.
4	Procesamiento de la información	Media	Modelar la información de las dos fuentes de información, con el fin de ser el insumo de los modelos estadísticos.
5	Modulo estimación de la producción de café cereza	Media	Construcción del módulo de la estimación de la producción de café cereza con los respectivos parámetros de entrada para los modelos y graficar los resultados.

Tabla 15. Historias de usuario.

A partir de las HU definidas anteriormente, se especifican una serie de iteraciones (sprint) que satisfacen los objetivos del trabajo (Tabla 16).

SPRINT (ITERACION)	OBJETIVO	HU
1	Recursos software	Definición de arquitectura
2	Datos agrupados e Interfaces intuitivas	Almacenamiento de la información
		Diseño de interfaces
3	Series de tiempo	Procesamiento de la información
4	Estimación de la producción de café cereza	Modulo estimación de la producción de café cereza

Tabla 16. Iteraciones SCRUM.

6.1.1 Sprint 1

Debido a que el prototipo esta implementado en la plataforma agroclimática IoT-Agro, esta plataforma tiene las siguientes características, en cuanto a la infraestructura.

- VPS (Virtual Private Server).
- Intel(R) Xeon(R) CPU E5-2660 v2.
- Windows Server 2016.
- Disco solido de 500gb.
- Memoria ram de 32gb.

Estos recursos computacionales son adquiridos por el proyecto IoT-Agro con el proveedor HostingRed (https://www.hostingred.com/colombia/), los cuales tienen una duración de 3 años, que es el tiempo de duración del proyecto. Estos servicios adquiridos son servidores con recursos dedicados. Diseñados para aplicaciones que requieren hacer uso intenso de la CPU y de almacenamiento, ya sea para análisis de datos, ejecutar consultas de Bases de datos grandes y/o para sitios o aplicaciones que soportan escritura y lectura de usuarios a gran escala.

Para el desarrollo de la plataforma se utilizó el Framework CakePhp 3.3 y además algunas librerías de JavaScript (HighCharts) para las gráficas. Para el almacenamiento de la información se utilizó el motor de bases de datos MySQL Workbench 8.0.

6.1.2 Sprint 2

Las dos fuentes de información son almacenadas en el motor de bases de datos MySql, con el fin de más adelante ser procesada para el entrenamiento de los modelos estadísticos, como se muestra en la Figura 16.

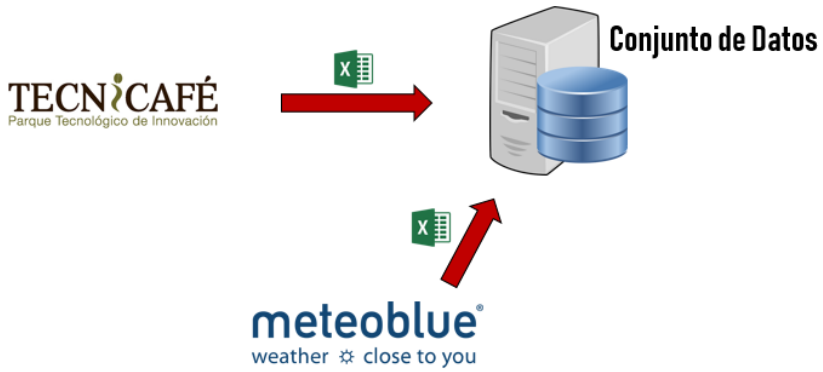

Figura 16. Almacenamiento de información.

El almacenamiento de la información climática proveniente del servicio de Meteoblue, se define en el diagrama entidad-relación (ER), el cual se muestra en la Figura 17.

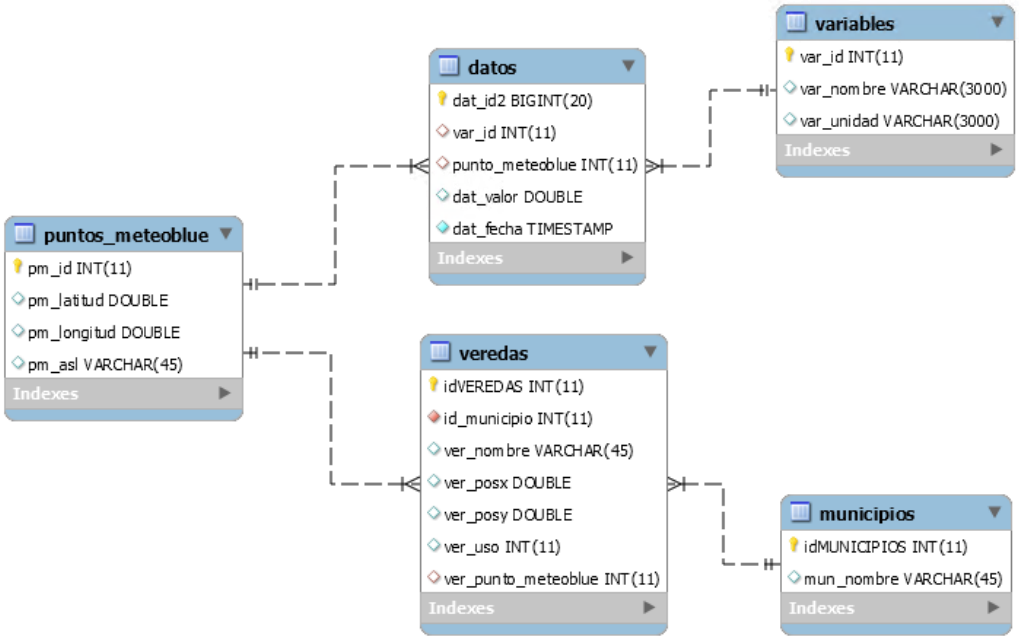

Figura 17. Diagrama ER – información climática.

Debido a que Meteoblue entrega la información climática por puntos de coordenadas, la información se discrimina por vereda y municipio.

Esta información es consumida por un servicio REST-API, que responde con un archivo con formato CSV, el cual es almacenado en la base de datos relacional. La petición al servicio se realiza dos veces al día cada 12 horas para obtener los datos climáticos recientes al día actual.

De igual forma, la información de manejo de cultivo se almacena en la base de datos relacional del servidor en la nube. Esta información viene registrada en un archivo por cada mes con formato XLSX.

El diseño de las interfaces es de acuerdo al dominio de aplicación (cultivo de café), por eso los colores, fondos, y formas de las vistas. También, con el fin de ser amigables y de fácil interacción con los usuarios.

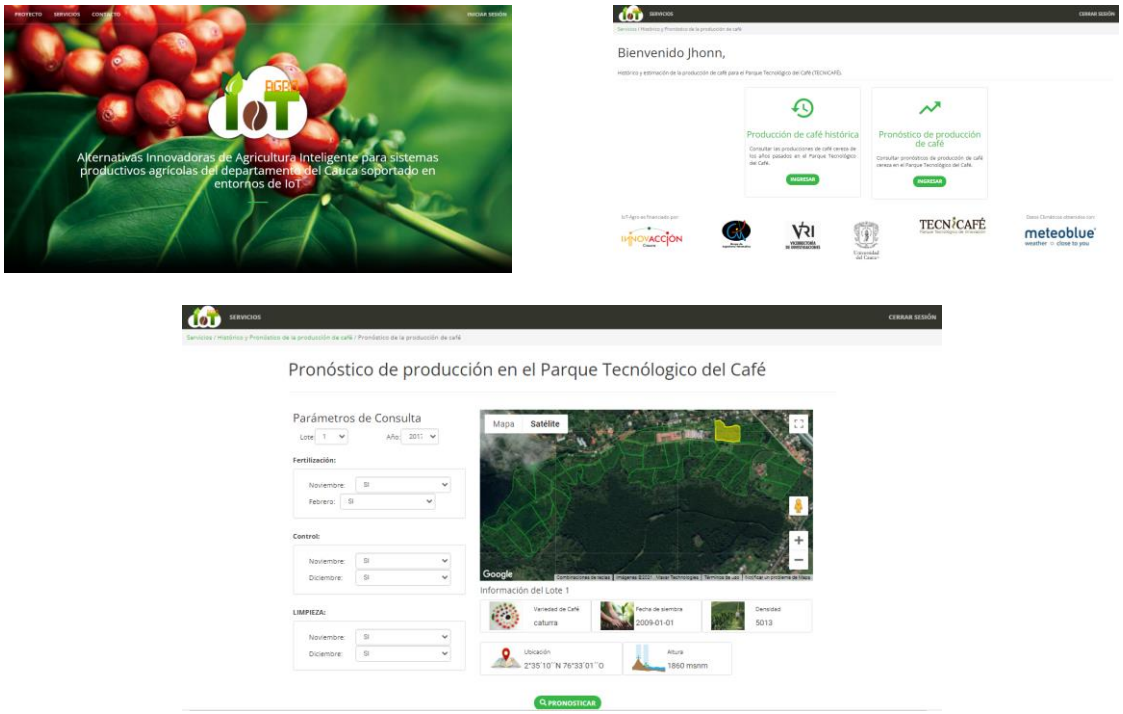

Figura 18. Interfaces.

6.1.3 Sprint 3

Luego de haber almacenado la información tal como viene de las fuentes, se transforma o modela la información con el fin de ser el insumo de los modelos estadísticos. La información climática como la información de manejo de cultivo es modelada a partir de la relevancia a la producción de café.

En el módulo de la estimación de la producción de café se grafica la información climática utilizada para entrenar los modelos, como se muestra en la Figura 19.

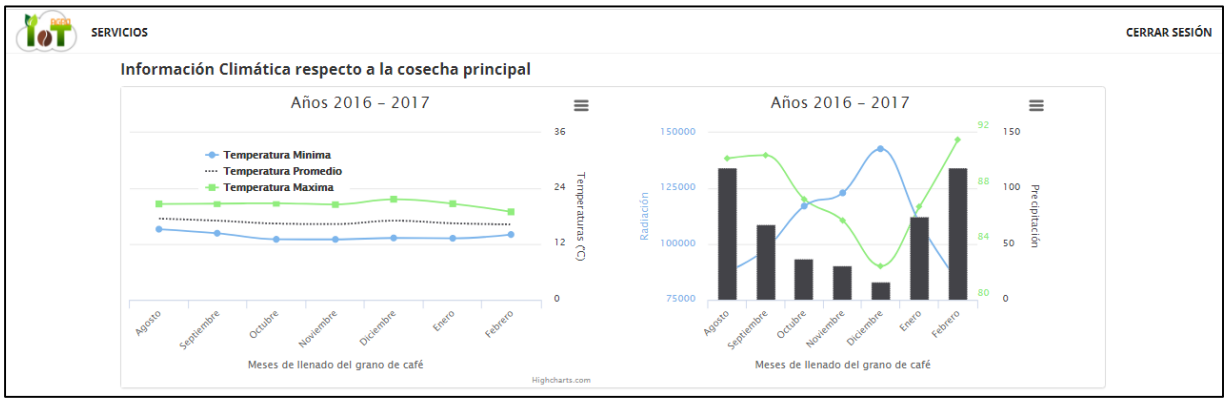

Figura 19. Prototipo - información climática.

La información climática cuando es transformada se almacena en una tabla temporal de la base de datos relacional. Por otro lado, la información de manejo de cultivo se visualiza en la parte superior del prototipo, como se muestra en la Figura 20.

Figura 20. Prototipo - información manejo de cultivo.

Al final las dos fuentes de información son integradas en series de tiempo que representan la producción de café cereza.

6.1.4 Sprint 4

El módulo de estimación de la producción de café cereza se realiza a partir de la información relevante al entrenamiento de los modelos estadísticos.

Este módulo privado, por consiguiente requiere credenciales de usuario para poder ingresar y utilizarlo. En la interfaz se observa los parámetros de entrada y un mapa donde se ubica la finca los Naranjos con la división por lote.

Primero se deben ingresar los datos respectivos y luego oprimir el botón “PRONOSTICAR”, al presionar este botón el prototipo iniciara el proceso de obtener los datos climáticos respectivos y graficarlos, esto puede tomar un tiempo (Figura 21).

Figura 21. Prototipo – cargue de información.

Finalmente, se visualizara la gráfica con los resultados estimados para la producción de café cereza por mes.

En la Figura 22, se observa en la parte derecha la gráfica del resultado, el cual en el eje x son los meses de la cosecha principal del año, y en el eje y la cantidad en kilogramos de la producción de café cereza. En la parte izquierda, se encuentra el valor acumulado de la producción de café cereza real como la de la estimada.

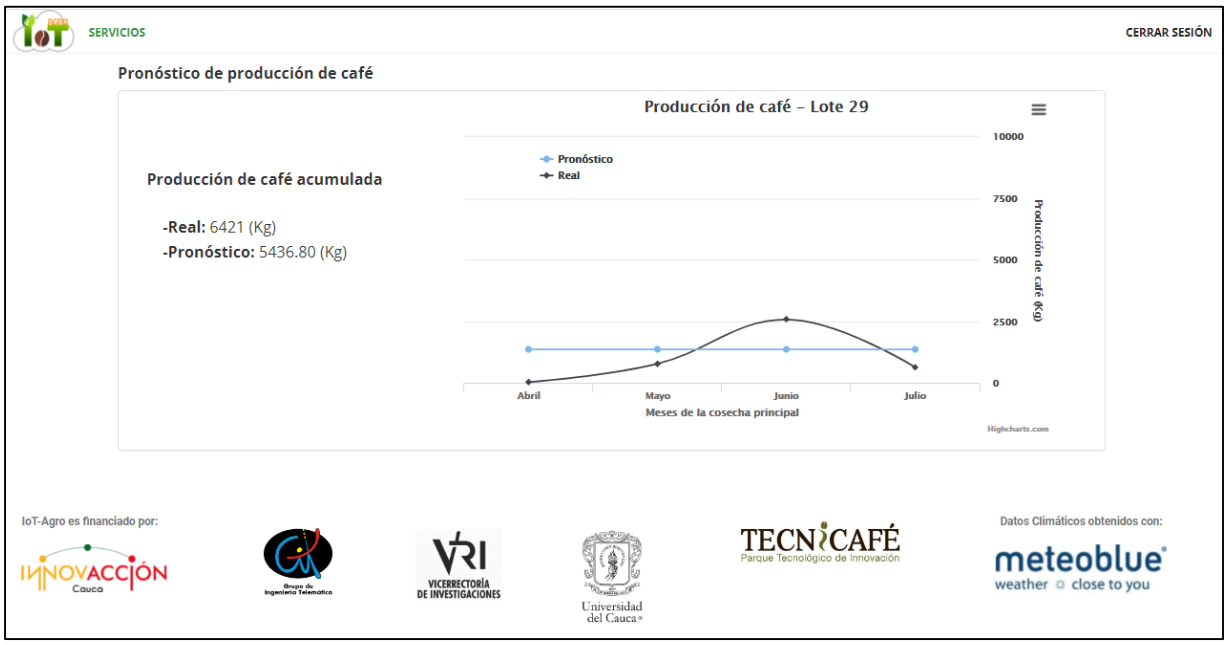

Figura 22. Prototipo – estimación de la producción de café cereza.

Por último, el prototipo funciona en línea (internet) en la plataforma de IoT-Agro: https://www.iot-agro.com, hasta el 20 de noviembre del 2021, debido a que los servicios de recursos computacionales contratados por el proyecto caducan. Posterior a esta fecha la plataforma como el prototipo no estarán disponibles.

6.2 Resumen

En este capítulo se expusieron los elementos requeridos por el marco de trabajo ágil SCRUM para el desarrollo del prototipo que implementa la estimación de la producción de café cereza basada en series de tiempo construido en este trabajo de Maestría. En primer lugar, se muestra las historias de usuario que fueron construidas para la primera versión del prototipo, todo esto apoyado de diagramas de entidad-relacion. Luego, se definen los Sprints necesarios para el desarrollo del producto software. Finalmente se muestran las interfaces principales del prototipo para este proyecto de Maestría.

Capítulo 7: Conclusiones y trabajos futuros

7.1 Conclusiones

La estimación de la producción de café cereza permite a los caficultores planificar las actividades y recursos requeridos, anticipar negociaciones, precios y pérdidas de producción de café, y ofrecer un producto de alta calidad considerando la cantidad de factores que pueden afectar la producción y no son fáciles de predecir. Sin embargo, como se ha descrito en el documento, hay un número reducido de propuestas de investigación que estimen el rendimiento de los cultivos mediante métodos estadísticos; este número es aún menor para la estimación de café cereza.

En este trabajo hemos presentado una propuesta para la estimación de la producción del café basada en la combinación de series de tiempo y un conjunto de modelos estadísticos. La evaluación detallada de estos modelos muestra que los algoritmos basados en árboles de regresión (XGBoost, TreeRegresor y RForest) proporcionan muy buenos resultados para el conjunto de datos recopilados en condiciones reales durante más de un año. Además, el modelo propuesto en este trabajo de investigación permite estimar las posibilidades de negociar y cumplir los plazos de entrega comprometidos con base en un modelo no destructivo, a diferencia de los métodos utilizados actualmente por los productores de café.

Por otra parte, la finca cafetera “Los Naranjos” (Figura 4) tiene un promedio de 4.57 mil plantas por hectárea y un total de 38 lotes por hectárea. Si se calcula la pérdida de granos de café por planta, se obtiene 20.40 mil granos perdidos por hectárea y 775.35 mil para la finca completa. Si esta cantidad se convierte a kilogramos, el resultado es de 1.39 mil kilogramos (aproximadamente 1.40 toneladas) de pérdida para la finca completa. Por último, también sería importante considerar la pérdida de granos de café generada durante el transporte del cultivo al centro de acopio donde se realiza el pesaje de los granos.

Los modelos propuestos en este trabajo obtienen buenos resultados, y son de gran ayuda para los caficultores, debido a que les permite estimaciones de sus producciones antes de comercializarlas y también para la toma de decisiones con respecto a manejo de cultivo que ocasione aumento en la producción. Los modelos aunque experimentalmente tengan buenos resultados, están ligados a que si las

condiciones cambian debido al cambio climático, los modelos no van a tener buenos resultados en un entorno real. Esto debido a que el cambio climático tiene consecuencias inmediatas como el alza del nivel del mar y el derretimiento de los polos, ocasionando que variables climáticas como las temperaturas de la tierra suban y otras variables también sean alteradas, por consiguiente afectando de forma directa a los cultivos de café.

Finalmente, los modelos propuestos en este trabajo carecen de información disponible para ser entrenados. Por tal motivo, sería interesante integrar información como: índices de vegetación, propiedades del suelo,

7.2 Trabajos futuros

A continuación, se presentan los trabajos futuros:

- Extender la evaluación considerando variables adicionales de la producción de café cereza que actualmente se están recolectando y anotando, como los índices de vegetación, que también pueden tener una alta correlación con la producción de cultivos.
- Incrementar el número de muestras de la producción de café cereza con enfoques de visión por computadora, por medio de conteo de granos a partir de imágenes tomadas en campo.
- Implementar otros tipos de enfoques de la inteligencia artificial para validar las series de tiempo propuestas en este trabajo, como por ejemplo enfoques de deep learning, sistemas expertos, y expert elicitation. Si bien es cierto, deep learning ha tenido gran crecimiento en áreas de reconocimiento de imágenes y detección de objetos, estas técnicas pueden ser llevadas a cabo para estimaciones en cultivos como el café.

REFERENCIAS

[1] International Coffee Organization, “Informe del mercado de café enero 2019,” 2019. http://www.ico.org/documents/cy2018-19/cmr-0119-c.pdf

[2] International Coffee Organization, “Total production by all exporting countries,” 2017. http://www.ico.org/prices/po-production.pdf

[3] Federación Nacional de Cafeteros, “Ensayos sobre ECONOMÍA CAFETERA No.30,” 2015.

[4] R. G. G. Cáceres and É. S. O. Escobar, “Caracterización de las cadenas de valor y abastecimiento del sector agroindustrial del café,” *Cuad. Adm.*, vol. 19, no. 31, pp. 197–217, 2006.

[5] J. R. Rendón-Sáenz, J. Arcila-Pulgarín, and E. C. Montoya-Restrepo, “ESTIMACIÓN DE LA PRODUCCIÓN DE CAFÉ CON BASE EN LOS REGISTROS DE FLORACIÓN,” 2008. Accessed: Nov. 14, 2018. [Online]. Available: https://www.researchgate.net/publication/277139994_ESTIMACION_DE_LA_PRODUCCION_DE_CAFE_CON_BASE_EN_LOS_REGISTROS_DE_FLORACION

[6] P. Ramos, F. Prieto, C. Oliveros, N. Aleixos, F. Albert, and J. Blasco, “Medición del porcentaje de madurez en ramas de café mediante dispositivos móviles y visión por computador,” 2015.

[7] R. Wirth, “CRISP-DM: Towards a standard process model for data mining,” in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.

[8] J. P. Rodríguez, A. I. Montoya-Munoz, C. Rodriguez-Pabon, J. Hoyos, and J. C. Corrales, “IoT-Agro: A smart farming system to Colombian coffee farms,” *Comput. Electron. Agric.*, vol. 190, p. 106442, Nov. 2021, doi: 10.1016/j.compag.2021.106442.

[9] D. Peña, *Análisis de series temporales*. Alianza Editorial, 2010.

[10] C. S. V. Mabel González Castellanos, “Minería de datos para Series Temporales,” 2013, doi: 10.13140/RG.2.1.2571.9841.

[11] IBM, “Modelos estadísticos,” Oct. 24, 2014. www.ibm.com/support/knowledgecenter/es/ss3ra7_sub/modeler_mainhelp_client_ddita/clementine/nodes_statisticalmodels.html

[12] L. Yang, S. Liu, S. Tsoka, and L. G. Papageorgiou, “A regression tree approach using mathematical programming,” *Expert Syst. Appl.*, vol. 78, pp. 347–357, Jul. 2017, doi: 10.1016/j.eswa.2017.02.013.

[13] D. C. Corrales, J. C. Corrales, and A. Figueroa-Casas, “Towards Detecting Crop Diseases and Pest by Supervised Learning,” *Ing. Univ.*, vol. 19, no. 1, pp. 207–228, Jun. 2015, doi: 10.11144/Javeriana.iyu19-1.tdcd.

[14] P. J. Ramos, F. A. Prieto, E. C. Montoya, and C. E. Oliveros, “Automatic fruit count on coffee branches using computer vision,” *Comput. Electron. Agric.*, vol. 137, pp. 9–22, May 2017, doi: 10.1016/j.compag.2017.03.010.

[15] J. P. Rodríguez, D. C. Corrales, J.-N. Aubertot, and J. C. Corrales, “A computer vision system for automatic cherry beans detection on coffee trees,” *Pattern Recognit. Lett.*, vol. 136, pp. 142–153, Aug. 2020, doi: 10.1016/j.patrec.2020.05.034.

[16] L. Kouadio, R. C. Deo, V. Byrareddy, J. F. Adamowski, S. Mushtaq, and V. Phuong Nguyen, “Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties,” *Comput. Electron. Agric.*, vol. 155, pp. 324–338, Dec. 2018, doi: 10.1016/j.compag.2018.10.014.

[17] L. Kouadio, V. M. Byrareddy, A. Sawadogo, and N. K. Newlands, “Probabilistic yield forecasting of robusta coffee at the farm scale using agroclimatic and remote sensing derived indices,” *Agric. For. Meteorol.*, vol. 306, p. 108449, Aug. 2021, doi: 10.1016/j.agrformet.2021.108449.

[18] B. D. S. Barbosa, G. A. e S. Ferraz, L. Costa, Y. Ampatzidis, V. Vijayakumar, and L. M. dos Santos, “UAV-based coffee yield prediction utilizing feature selection and deep learning,” *Smart Agric. Technol.*, vol. 1, p. 100010, Dec. 2021, doi: 10.1016/j.atech.2021.100010.

[19] S. Mishra, D. Mishra, and G. H. Santra, “Adaptive boosting of weak regressors for forecasting of crop production considering climatic variability: An empirical assessment,” *J. King Saud Univ. - Comput. Inf. Sci.*, 2017, doi: 10.1016/j.jksuci.2017.12.004.

[20] R. R. do V. Gonçalves, J. Zullo, T. M. Peron, S. R. M. Evangelista, and L. A. S. Romani, “Numerical models to forecast the sugarcane production in regional scale based on time series of NDVI/AVHRR images,” in *2015 8th International Workshop on the Analysis of Multitemporal Remote Sensing Images (Multi-Temp)*, Jul. 2015, pp. 1–4. doi: 10.1109/Multi-Temp.2015.7245806.

[21] S. H. Qader, J. Dash, and P. M. Atkinson, “Forecasting wheat and barley crop production in arid and semi-arid regions using remotely sensed primary productivity and crop phenology: A case study in Iraq,” *Sci. Total Environ.*, vol.

613–614, no. Supplement C, pp. 250–262, 2017, doi: 10.1016/j.scitotenv.2017.09.057.

[22] P. C. Fernandez-Mensaque, F. J. G. Minero, J. Morales, and C. Tomas, “Forecasting olive (Olea europaea) crop production by monitoring airborne pollen,” *Aerobiologia*, vol. 14, no. 2–3, pp. 185–190, Sep. 1998, doi: 10.1007/BF02694204.

[23] B. Garg, S. Aggarwal, and J. Sokhal, “Crop yield forecasting using fuzzy logic and regression model,” *Comput. Electr. Eng.*, Nov. 2017, doi: 10.1016/j.compeleceng.2017.11.015.

[24] R. Sujatha and P. Isakki, “A study on crop yield forecasting using classification techniques,” in *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE’16)*, Jan. 2016, pp. 1–4. doi: 10.1109/ICCTIDE.2016.7725357.

[25] D. Ramesh, “ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUES,” *Int. J. Res. Eng. Technol.*, vol. 04, no. 01, pp. 470–473, 2015.

[26] I. Oliveira, R. L. F. Cunha, B. Silva, and M. A. S. Netto, “A Scalable Machine Learning System for Pre-Season Agriculture Yield Forecast,” *ArXiv180609244 Cs Stat*, Jun. 2018, [Online]. Available: http://arxiv.org/abs/1806.09244

[27] Feng Zhang, Bingfang Wu, and Chenglin Liu, “Using time series of SPOT VGT NDVI for crop yield forecasting,” in *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*, Jul. 2003, vol. 1, pp. 386–388 vol.1. doi: 10.1109/IGARSS.2003.1293784.

[28] H. Kerdiles, F. Rembold, O. Leo, H. Boogaard, and S. Hoek, “CST, a freeware for predicting crop yield from remote sensing or crop model indicators: Illustration with RSA and Ethiopia,” in *2017 6th International Conference on Agro-Geoinformatics*, Aug. 2017, pp. 1–6. doi: 10.1109/Agro-Geoinformatics.2017.8047071.

[29] J. Sun, J. Huang, J. Chen, and L. Wang, “Grain Yield Estimating for Hubei Province Using Remote Sensing Data Take Semilate Rice as an Example,” in *2009 International Conference on Environmental Science and Information Application Technology*, Jul. 2009, vol. 1, pp. 497–500. doi: 10.1109/ESIAT.2009.69.

[30] H. Lee and A. Moon, “Development of yield prediction system based on real-time agricultural meteorological information,” in *16th International Conference on Advanced Communication Technology*, Feb. 2014, pp. 1292–1295. doi: 10.1109/ICACT.2014.6779168.

[31] Xu Xingmei, Cao Liying, Zhou Jing, and Su Fengyan, “Study and application of grain yield forecasting model,” in *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*, Dec. 2015, vol. 01, pp. 652–656. doi: 10.1109/ICCSNT.2015.7490829.

[32] N. Rale, R. Solanki, D. Bein, J. Andro-Vasko, and W. Bein, “Prediction of Crop Cultivation,” in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, Jan. 2019, pp. 0227–0232. doi: 10.1109/CCWC.2019.8666445.

[33] A. A. Gamboa, P. A. Cáceres, H. Lamos, D. A. Zárate, and D. E. Puentes, “Predictive model for cocoa yield in Santander using Supervised Machine Learning,” in *2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*, Apr. 2019, pp. 1–5. doi: 10.1109/STSIVA.2019.8730258.

[34] D. G. Mayer, K. A. Chandra, and J. R. Burnett, “Improved crop forecasts for the Australian macadamia industry from ensemble models,” *Agric. Syst.*, vol. 173, pp. 519–523, Jul. 2019, doi: 10.1016/j.agsy.2019.03.018.

[35] D. Tedesco-Oliveira, R. Pereira da Silva, W. Maldonado, and C. Zerbato, “Convolutional neural networks in predicting cotton yield from images of commercial fields,” *Comput. Electron. Agric.*, vol. 171, p. 105307, Apr. 2020, doi: 10.1016/j.compag.2020.105307.

[36] A. K. Mariappan and J. A. B. Das, “A paradigm for rice yield prediction in Tamilnadu,” in *2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, Apr. 2017, pp. 18–21. doi: 10.1109/TIAR.2017.8273679.

[37] J. G. A. Barbedo, “Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification,” *Comput. Electron. Agric.*, vol. 153, pp. 46–53, Oct. 2018, doi: 10.1016/j.compag.2018.08.013.

[38] R. Lopez, “WHEAT in the World,” *WHEAT*. https://wheat.org/wheat-in-the-world/ (accessed Nov. 07, 2021).

[39] F. F. Farfan V. and P. M. Sanchez A., “Densidad de siembra del café variedad Castillo en sistemas agroforestales en el departamento de Santander Colombia,” Jul. 2016, Accessed: Dec. 14, 2020. [Online]. Available: https://biblioteca.cenicafe.org/handle/10778/678

[40] J. Arcila, F. Farfan, A. Moreno, L. Salazar, and E. Hincapie, *Sistemas de produccion de cafe en colombia. Cenicafe*. 2007.

[41] J. Arcila, “Crecimiento y desarrollo de la planta de café,” in *Sistemas de producción de café en Colombia*, 2007.

[42] J. R. Quinlan, “Learning With Continuous Classes,” 1992, pp. 343–348.

[43] Kavitha S, Varuna S, and Ramya R, “A comparative analysis on linear regression and support vector regression,” in *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*, Nov. 2016, pp. 1–5. doi: 10.1109/GET.2016.7916627.

[44] D. C. Corrales, G. Gutierrez, J. P. Rodriguez, A. Ledezma, and J. C. Corrales, “Lack of Data: Is It Enough Estimating the Coffee Rust with Meteorological Time Series?,” in *Computational Science and Its Applications – ICCSA 2017*, Jul. 2017, pp. 3–16. doi: 10.1007/978-3-319-62395-5_1.

[45] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[46] M. R. Segal, “Machine Learning Benchmarks and Random Forest Regression,” Apr. 2004, Accessed: Sep. 11, 2020. [Online]. Available: https://escholarship.org/uc/item/35x3v9t4

[47] K. G. Nisha and K. Sreekumar, “A review and analysis of machine learning and statistical approaches for prediction,” in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Mar. 2017, pp. 135–139. doi: 10.1109/ICICCT.2017.7975174.

[48] C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” *Clim. Res.*, vol. 30, no. 1, pp. 79–82, Dec. 2005, doi: 10.3354/cr030079.

[49] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products,” *Chemom. Intell. Lab. Syst.*, vol. 83, no. 2, pp. 83–90, Sep. 2006, doi: 10.1016/j.chemolab.2006.01.007.

[50] A. Mundra, S. Misra, and C. A. Dhawale, “Practical Scrum-Scrum Team: Way to Produce Successful and Quality Software,” in *2013 13th International Conference on Computational Science and Its Applications*, Jun. 2013, pp. 119–123. doi: 10.1109/ICCSA.2013.25.