UNIVERSIDAD
NACIONAL
DE COLOMBIA

Modelo predictivo para la ocurrencia de leishmaniasis cutánea en Colombia, a partir de variables ambientales y socioeconómicas

José Daniel Salazar Mora

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de ingeniería de Sistemas e industrial,
Maestría en ingeniería de Sistemas y Computación
Bogotá, Colombia
2021

Modelo predictivo para la ocurrencia de LC en Colombia, a partir de variables ambientales y socioeconómicas

José Daniel Salazar Mora

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

Magister en Ingeniería de Sistemas y Computación

Director (a):

Luis Fernando Niño Vásquez, Ph.D., profesor titular Universidad Nacional de Colombia,

Codirector (a):

Juan David Gutiérrez, Ph.D., profesor titular Universidad de Santander

Línea de Investigación:

Ciencia de datos e inteligencia artificial

Grupo de Investigación:

Laboratorio de investigación en sistemas inteligentes (LISI)

Universidad Nacional de Colombia

Facultad de ingeniería, Departamento de ingeniería de sistemas e industrial

Bogotá, Colombia

2021

(Dedicatoria o lema)

Este trabajo es un avance en mi vida profesional y personal y está dedicado a mis padres, mi pareja, mi director de tesis y codirector, a la Universidad Nacional de Colombia y demás personas que me apoyaron en la realización de esta meta.

Agradecimientos

Este trabajo llevó bastante tiempo y muchos intentos para llegar a buenos resultados, hay mucho que agradecer y a muchas personas, seres e instituciones. En primer lugar, quiero extender mi más grande agradecimiento a Dios, Él siempre ha estado a mi lado y en la realización de esta tesis siempre me apoyó y me brindó su compañía aun cuando las cosas parecían no tener solución o me quería rendir.

En segundo lugar, a la majestuosa Universidad Nacional de Colombia, quien me ha dado mucho, no hay nada que no haya obtenido sin ella. De igual manera, también a las personas que he conocido a lo largo de mi vida académica y profesional y que, de alguna manera, ya sea directa o indirectamente han aportado a mi crecimiento personal y profesional, además de apoyarme en la realización de este trabajo. A ellos mis más sinceros agradecimientos.

Además, agradezco al profesor titular del departamento de ingeniería de sistemas e industrial de la Universidad Nacional de Colombia, mi director de tesis, Luis Fernando Niño Vásquez, PhD, una gran persona, quien siempre estuvo apoyándome, acompañándome en este proceso y enseñándome. Otra persona muy importante presente en la realización de este proyecto es el profesor Juan David Gutiérrez, PhD. mi codirector en esta tesis, el cual es un gran ser humano, un apoyo incondicional y siempre con su buena energía y disposición me brindó su ayuda en todos los momentos que la necesité, ya fueran académicos o personales. Al grupo de investigación LISI, el cual me brindó bastante ayuda y orientación en la realización de esta tesis. Finalmente, quiero agradecer a mi pareja, mi compañera de vida, quien ha estado conmigo en grandes momentos apoyándome para alcanzarlos.

Resumen

Modelo predictivo para la ocurrencia de leishmaniasis cutánea en Colombia, a partir de variables ambientales y socioeconómicas

Se crearon varios modelos predictivos para la ocurrencia de leishmaniasis cutánea en Colombia a partir de un conjunto de variables socioeconómicas y ambientales. Con este conjunto de datos (*dataset*) se hizo un trabajo de ciencia de datos utilizando el proceso de *KDD* (*Knowledge Discovery in Databases*), pasando por cada una de sus etapas. Particularmente, se recolectó y organizó el conjunto de datos, se elaboró una descripción y revisión de este y se hizo un análisis estadístico descriptivo. Después, se realizó el preprocesamiento de los datos, se hicieron transformaciones de estos y se implementaron técnicas de reducción de dimensionalidad.

Posteriormente, se procedió a utilizar diferentes técnicas de aprendizaje de máquina, tanto para clasificación como regresión. Para clasificación se implementaron varios métodos: naive bayes, redes neuronales (perceptrón multicapa), árboles de decisión y redes bayesianas, los cuales permitieron generar un modelo predictivo de clasificación, obteniendo los mejores resultados con el algoritmo *XGBoost* sobre un *set* de datos municipal con datos reportados mensualmente. De la misma forma, se realizó un modelo de regresión a través de redes neuronales y *XGBoost*, obteniendo los mejores resultados con el algoritmo *XGBoost*, pero esta vez con un conjunto de datos departamentales con periodicidad mensual. Finalmente, se realizó un análisis de series de tiempo con algoritmos de regresión con redes neuronales y *XGBoost* obteniendo las mejores métricas con *XGBoost* para un modelo departamental con resolución temporal semanal.

Con cada uno de los modelos se identificaron las variables más importantes para la predicción; todos los modelos tuvieron en cuenta al menos las siguientes: el total de la población, precipitación, temperatura, índice de vegetación mejorado (EVI por sus siglas en inglés) y mes. Además, para poder utilizar el modelo de regresión para series de tiempo, se creó una página web que recibe como entrada las variables independientes junto con sus retrasos y genera la predicción de la cantidad de casos futuros a 1, 2 y 4 semanas.

Palabras clave: Modelo predictivo, leishmaniasis cutánea, ciencia de datos, aprendizaje de máquina, *XGBoost*, redes neuronales, series de tiempo.

Abstract

Predictive model for the occurrence of cutaneous leishmaniasis in Colombia, based on environmental and socioeconomic variables

Several predictive models were created for the occurrence of cutaneous leishmaniasis in Colombia from a set of socioeconomic and environmental variables. With this dataset, a data science work was done using the KDD process (Knowledge Discovery in Databases), going through each of its stages. In particular, the data set was collected and organized, a description and review of it was prepared, and a descriptive statistical analysis was carried out. Afterwards, the data was preprocessed, transformations were made of these and dimensionality reduction techniques were implemented.

Subsequently, different machine learning techniques were used, both for classification and regression. For classification, several methods were implemented: naive bayes, neural networks (multilayer perceptron), decision trees and Bayesian networks, which allowed to generate a predictive classification model, obtaining the best results with the *XGBoost* algorithm on a municipal data set with data reported monthly. In the same way, a regression model was carried out through neural networks and *XGBoost*, obtaining the best results with the *XGBoost* algorithm, but this time with a departmental data set on a monthly basis. Finally, a time series analysis was performed with regression algorithms with neural networks and *XGBoost*, obtaining the best metrics with *XGBoost* for a departmental model with weekly temporal resolution.

With each of the models, the most important variables for prediction were identified; all the models took into account at least the following variables: the total population, precipitation, temperature, improved vegetation index (EVI) and month. In addition, to be able to use the regression model for time series, a web page was created that receives as input the independent variables together with their delays and generates the prediction of the number of future cases at 1, 2 and 4 weeks.

Keywords: Predictive model, cutaneous leishmaniasis, data science, machine learning, XGBoost, neural networks, time series.

Esta Tesis de Maestría fue calificada en septiembre de 2021 por los siguientes evaluadores:

Diego F. Cuadros, Ph.D.
Associate Professor, University of Cincinnati
Daniel Restrepo-Montoya, Ph.D.
Postdoctoral Researcher - Bioinformatics, Department of Crop Science, North Carolina State University

Contenido

Capítulo 1: Introducción

Las condiciones climáticas, geográficas y ecológicas hacen que la población colombiana esté expuesta a sufrir de enfermedades transmitidas por vectores, una de las cuales es la leishmaniasis cutánea (LC). Esta es una enfermedad infecciosa producida por parásitos del género *Leishmania*. La infección es transmitida a los humanos por la picadura de insectos (vectores) del género *Lutzomyia*, conocidos popularmente en Colombia como arenillas. Esta enfermedad afecta la piel produciendo laceraciones o grandes úlceras en cualquier parte del cuerpo, lo cual puede causar incapacidad o incluso pérdida de la extremidad afectada [1]. El ciclo de transmisión de la infección involucra un reservorio animal (exclusivamente mamíferos) que porta el parásito, un vector que se alimenta de dicho reservorio y un humano que es afectado por la enfermedad [1].

Esta enfermedad es considerada un problema de salud pública en Colombia, no sólo por la reducción de la calidad de vida de las personas que la padecen, sino también por los cambios en el comportamiento del evento, la aparición de nuevos focos, la domiciliación del vector dada por la frecuencia de casos en estudiantes y amas de casa que favorece un mayor número de casos en diferentes grupos de edad [1]. También, es importante anotar que existen diferentes factores socioeconómicos asociados a la enfermedad. Estos factores en general son: la pobreza, malnutrición y falta de saneamiento básico [1].

La LC afecta en promedio al 2% de la población colombiana y se estima que en el país existen más de 11 millones de personas que están en riesgo de contraer la enfermedad [2]. Según el Instituto Nacional de Salud, en 2017 se presentaron 7,696 casos de LC en Colombia, lo que equivale a una incidencia de 71.6 casos por 100,000 habitantes. Sin

embargo, en departamentos como el Guaviare la incidencia alcanza los 499 casos por 100.000 habitantes.

En este trabajo se realizaron varios modelos predictivos para la ocurrencia de LC implementando minería de datos y aprendizaje de máquina. Los modelos predictivos con mejores resultados obtenidos fueron: un modelo de clasificación con dos clases, los municipios que presentarán uno o más casos y los que presentarán cero casos a nivel mensual. El segundo modelo fue un modelo de regresión con un set de datos departamental a nivel mensual y un modelo de regresión con series de tiempo, donde se predice la cantidad de casos a nivel departamental ya sea una semana a futuro, dos semanas o cuatro. A partir de este último modelo se realizó una aplicación web que permite utilizar el modelo predictivo de manera real y práctica.

El presente documento se encuentra estructurado en 6 capítulos, donde se empieza con esta introducción, en seguida, se encuentra un marco conceptual el cual consta de dos secciones que son la parte teórica de leishmaniasis y la parte conceptual de aprendizaje de máquina y minería de datos utilizada en esta investigación. Después de ello se dedica un capítulo a los estudios previos con relación a esta investigación y otro capítulo para explicar la metodología seguida. Una vez se tiene detallado lo anteriormente mencionado se procede a presentar los resultados y discusión de la investigación, los cuales se enfocaron principalmente en las técnicas de minería de datos y aprendizaje de máquina utilizadas. En el capítulo 6 se expone el funcionamiento y arquitectura de la aplicación web desarrollada para la predicción. Finalmente, se exponen las conclusiones y el trabajo futuro, terminando con la bibliografía.

Capítulo 2: Marco conceptual

Este capítulo presenta las dos ramas conceptuales más importantes en esta investigación, a saber, los conceptos fundamentales sobre la leishmaniasis y los aspectos teóricos de la ciencia de datos utilizados para el desarrollo de este trabajo.

2.1. Leishmaniasis

Como se mencionó anteriormente en la introducción, la leishmaniasis es una enfermedad infecciosa producida por parásitos del género *Leishmania* y transmitida por la picadura de vectores del género *Lutzomyia.* Existen tres tipos de leishmaniasis que se describen a continuación.

2.1.1. Tipos de leishmaniasis

Leishmaniasis cutánea: Tiene diferentes formas clínicas y varía desde lesiones cerradas como pápulas, nódulos y placas con aspecto verrugoso, hasta las formas en úlceras. La enfermedad tiende a ser dolorosa cuando hay sobreinfección bacteriana. La LC puede tornarse crónica luego de doce o más semanas sin cierre de la úlcera o transformaciones de la misma en placas con costras [3].

Leishmaniasis mucocutánea (mucuosa): Ocurre como resultado de la diseminación linfohematógena del parásito. Afecta las mucosas de las vías aéreas superiores, nariz, faringe, laringe, boca y tráquea. Entre el 3% y 5% de los pacientes con LC, suelen presentar también la mucocutánea. Se producen lesiones en las zonas afectadas,

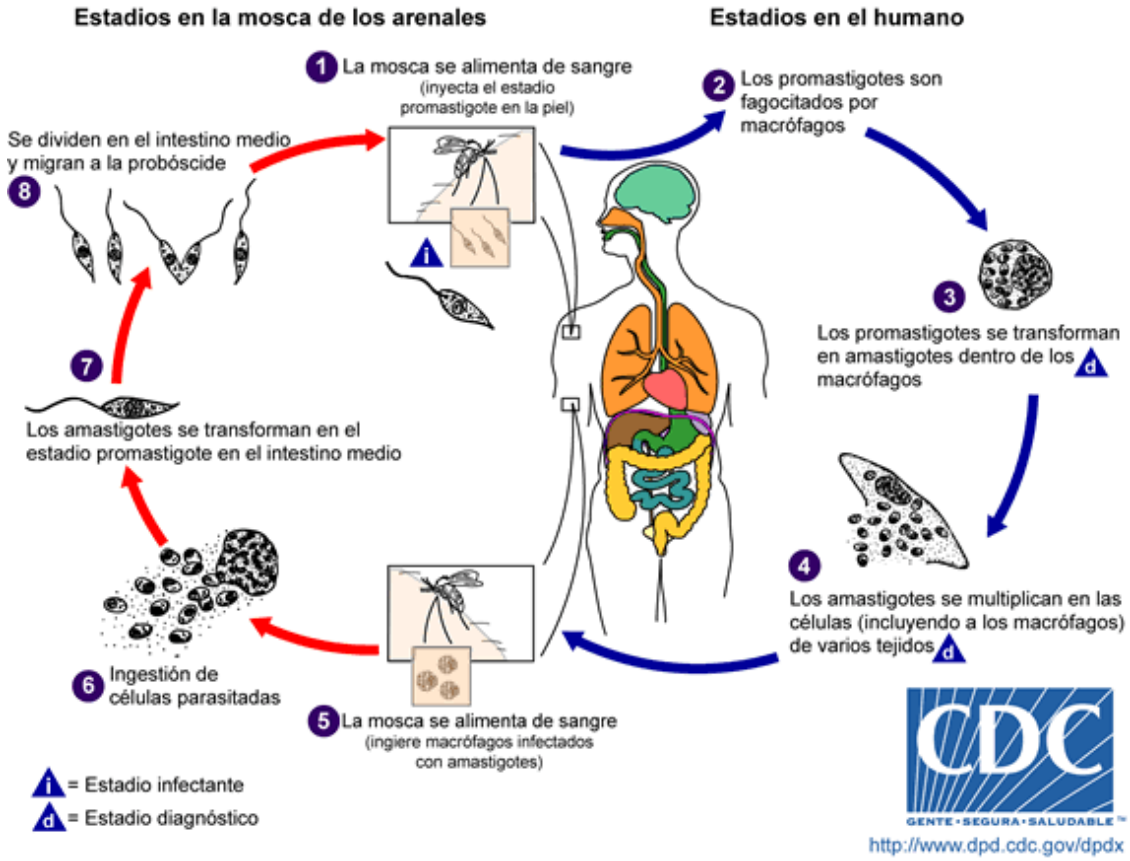
congestión y obstrucción. Además, de producir graves malformaciones por falta de atención oportuna [3].

Leishmaniasis visceral: Es una enfermedad del sistema retículo endotelial. Se caracteriza por fiebre, anemia, leucopenia, trombocitopenia y debilidad progresiva. También se pueden presentar diarreas e infecciones respiratorias [3].

2.1.2. Ciclo del parásito

El ciclo biológico de transmisión de la leishmaniasis se puede dividir en 8 etapas.

Figura 1: Ciclo biológico de la leishmaniasis



CDC. *Parásitos y salud, leishmaniasis*. [Figura]. 2004. Tomado de: [4]

En la Figura 1 se puede observar el ciclo de transmisión de la leishmaniasis, es importante saber que éste ciclo es digénico, es decir, que una parte del ciclo se desarrolla en el hospedador, por ejemplo, perros o conejos, y otra parte en insectos del género

Phlebotomus. Los parásitos de *Leishmania* están dentro del tubo digestivo del insecto, el cual se infecta al picar a un mamífero infectado, donde luego se transforman dentro del sistema digestivo del insecto y al picar a otro animal o ser humano le transmite la enfermedad, este ciclo puede durar de 4 a 20 días [5].

Tomando como referencia la Figura 1, el mosquito hembra se alimenta de un humano e inyecta el promastigote (fase 1), los promastigotes en el punto de la infección por el insecto son fagocitados por los macrófagos (fase 2), los promastigotes se transforman en amastigotes (fase 3), estos amastigotes se multiplican y afectan diferentes tejidos causando manifestaciones clínicas de la leishmaniasis (fase 4), los mosquitos pican animales infectados y adquieren el parásito (fase 5 y fase 6), en el aparato digestivo del mosquito, el parásito se desarrolla en promastigotes (fase 7), para completar el ciclo, los promastigotes se multiplican se desarrollan y se trasladan a las partes bucales del vector, donde al picar a otra persona se contagia la enfermedad (fase 8). [4].

2.1.3. Relación con las variables ambientales y socioeconómicas

Existen diferentes estudios que han demostrado la relación de la ocurrencia de casos de LC con factores ambientales y socioeconómicos. Por ejemplo, Hernández y colaboradores (2019) mencionan que la ocurrencia de LC se asocia principalmente con factores ambientales como la elevación, temperatura, precipitación y cobertura de tierra [6]. Colombia experimenta factores adicionales de índole socioeconómico, como conflicto armado, el cual causa desplazamiento de la población y conduce a condiciones de mayor presencia de personas en áreas de alto riesgo de la enfermedad. La mayoría de las estrategias que se han hecho para el control y la prevención de la leishmaniasis presuponen estabilidad social, política y contextos económicos para la presentación de servicios, puntos críticos en muchas regiones del país [6].

2.2. Minería de datos y aprendizaje de máquina

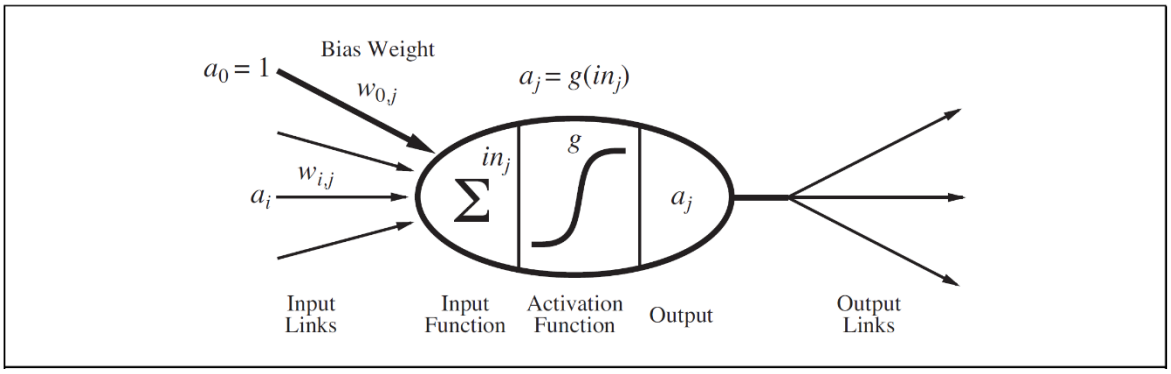
Esta sección del presente capítulo 2, está dedicada a presentar los conceptos y terminología más importante que se utilizó para el desarrollo de la investigación en términos de la inteligencia artificial y ciencia de datos implementada.

Inteligencia artificial (IA): Es el estudio del diseño de agentes inteligentes que son capaces de tomar la decisión de qué acciones tomar y cuándo tomar estas acciones [7].

Aprendizaje de máquina: Es una rama de la inteligencia artificial y tiene como fin desarrollar sistemas computacionales con la capacidad de aprender, adaptarse a nuevas circunstancias, detectar y extrapolar patrones [7].

Redes neuronales: Primero es importante mencionar que "una neurona es una célula del cerebro cuya función principal es la recolección, procesamiento y emisión de señales eléctricas" [7]. La idea es que las redes neuronales artificiales simulen el comportamiento de las neuronas biológicas, es así como las redes neuronales artificiales, están compuestas por unidades a través de conexiones dirigidas. Una conexión de una unidad con otra sirve para propagar la activación, además, cada conexión tiene un peso numérico asociado que determina su fuerza y el signo de conexión. Cada unidad primero calcula una suma ponderada de sus entradas. Luego se aplica una función de activación a la suma para producir la salida [7], la Figura 2 muestra esto en detalle.

Figura 2: Modelo matemático de una neurona



Russell & Norving. *Artificial Intelligence A Modern Approach.* [Figura], pp. 728

Clustering: Es una técnica de aprendizaje no supervisado [7], que consiste en tomar un conjunto de datos y generar diferentes grupos a partir de características que compartan los elementos de dicho conjunto.

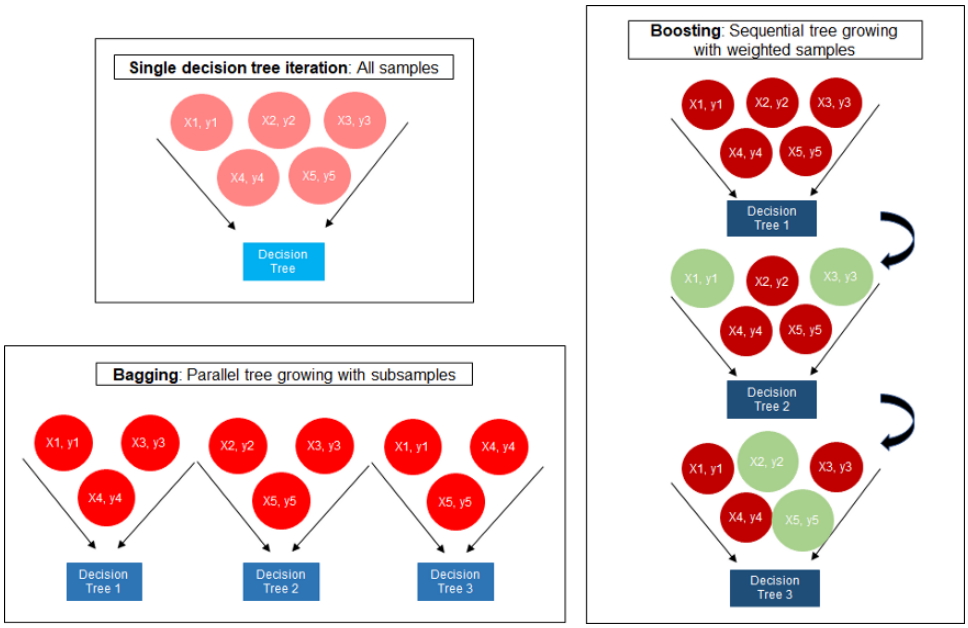
Series de tiempo: Es una secuencia de n observaciones ordenadas cronológicamente sobre varias características de una unidad observable en diferentes momentos (diario, semanal, mensual, etc.) [8].

Tuning: Conocido también como afinamiento u optimización de hiperparámetros, consiste en buscar objetivamente diferentes valores para los hiperparámetros de un modelo y elegir un subconjunto que dé como resultado un modelo con el mejor rendimiento para el conjunto de datos suministrado [9]. Para esta investigación se utilizó el método *RandomizedSearchCV* de la biblioteca *Sklearn*, el cual consiste en realizar una búsqueda de los mejores parámetros para el modelo de forma aleatoria a través del universo de hiperparámetros dado. De igual forma, existe otro método llamado *GridSearchCV*, también de *Sklean*, el cual, a diferencia del anterior, hace una búsqueda exhaustiva y prueba con cada uno de los hiperparámetros definidos en el universo de éstos, produciendo un mejor modelo a partir de cada una de las combinaciones y selección de parámetros, sin embargo, tiene un costo computacional mayor.

***XGBoost*:** Es una técnica basada en algoritmos de gradientes de árboles reforzados (*gradient boosting*). Estos algoritmos son basados en árboles de decisión [10] y son de aprendizaje supervisado, los cuales intentan predecir una variable objetivo combinando estimaciones de modelos más sencillos [11]. Adicional a ello, *XGBoost* funciona similar a los algoritmos de bosques aleatorios (*random forest)* en los cuales se crean muchos árboles y éstos son independientes entre sí, es decir no se necesita saber el resultado de uno para entrenar el otro. En el método *XGBoost* el entrenamiento se hace de manera secuencial, es decir, se entrena un árbol y se presta atención a los casos que peor se clasificó o peor resultado se obtuvo en regresión y ese margen de error se le pasa al árbol que es presentado a continuación para que intente mejorar ese resultado, ya sea de clasificación o de regresión. El resultado promedio de esos dos árboles se incorpora como referencia a un tercer árbol y así sucesivamente hasta encontrar el punto máximo de la cantidad de árboles (*n estimators*) buscando evitar el sobreajuste.

La Figura 3 permite entender el funcionamiento del método *XGBoost*, comparado con los modelos de árbol de decisión y el bosque aleatorio. La ilustración en la parte superior izquierda hace referencia al árbol de decisión, debajo de éste se encuentra representado gráficamente el funcionamiento del método de *random forest* y al lado derecho se observa el algoritmo basado en *gradient boosting*.

Figura 3: Árboles de decisión vs *random forest* vs *XGBoost*



Towards data science. *The Ultimate Guide to AdaBoost, random forests and XGBoost*. [Figura], 2020. Tomado de: [12]

Capítulo 3: Objetivos y pregunta de investigación

OBJETIVO GENERAL

Desarrollar un modelo predictivo para la ocurrencia de leishmaniasis cutánea a partir de datos ambientales y socioeconómicos mediante técnicas de inteligencia artificial y minería de datos.

OBJETIVOS ESPECIFICOS

- Construir el conjunto de datos a partir variables ambientales, socioeconómicas y epidemiológicas.
- Seleccionar e Implementar técnicas de minería de datos a las variables descritas anteriormente para asociarla y obtener patrones entre ellas
- Seleccionar e implementar técnicas de aprendizaje de máquina para la predicción de la ocurrencia de casos de leishmaniasis a nivel municipal. Teniendo en cuenta métricas de desempeño y su comportamiento en el conjunto de datos en cuestión.
- Proponer el modelo computacional que permita predecir la ocurrencia de casos de leishmaniasis cutánea.

PREGUNTA DE INVESTIGACIÓN

¿Cómo implementar un modelo predictivo con técnicas de inteligencia artificial para la ocurrencia de casos de LC en Colombia utilizando variables ambientales y socioeconómicas?

Capítulo 4: Estado del arte

Existen varios estudios previos relacionados con la LC, estos estudios se pueden dividir en dos grandes grupos. Aquellos que se centran en la descripción de la enfermedad y su tratamiento y los estudios que se enfocan en generar métodos predictivos para prevenir la enfermedad. Estos últimos son los que más interesan en el presente trabajo.

King y colaboradores (2004) desarrollaron un modelo predictivo de la ocurrencia de LC dividiendo el territorio nacional de Colombia en 5 zonas biogeográficas para el análisis epidemiológico. Dichos autores crearon un modelo de regresión logística a partir de la asociación entre la altura del municipio, la cobertura del suelo (25 variables categóricas) y la probabilidad de que se haya reportado al menos un caso de la enfermedad [13].

La relación entre el fenómeno climático de El Niño y la incidencia de la enfermedad en Costa Rica fue estudiada mediante el análisis de ondulas (*wavelets*). En este caso, la incidencia de la enfermedad fue predicha hasta con 12 meses de anterioridad [14].

Los modelos dinámicos lineales de series de tiempo han sido también implementados para evaluar el retraso entre el momento de infección y la aparición de los síntomas clínicos de la LC americana. En dicho estudio se determinó que, en general, hay un retraso de 5 meses entre la infección y el inicio de los síntomas [15]. Lo cual contrasta con lo conocido sobre el periodo de incubación de la enfermedad.

El análisis de las series de tiempo de variables meteorológicas mediante modelos autorregresivos de media móvil (ARIMA) y modelos estacionales autorregresivos de media móvil (SARIMA) ha sido implementado tanto en Brasil [16] como en Irán [17]. En ambos

casos los resultados sugieren que existe una asociación entre algunas variables climáticas como la temperatura y la precipitación y la ocurrencia de casos de LC.

Valderrama y colaboradores (2010) trabajaron en un análisis de los factores ambientales de riesgo para la LC. Este análisis se centró en el brote de 2003 a 2007 en Colombia. En dicha investigación utilizaron un modelo condicional autorregresivo de Poisson para la correlación espacial, las variables predictoras fueron, el uso de la tierra, la elevación y variables climáticas como la temperatura media y la precipitación. La variable respuesta fue el total de casos de LC en los 5 años de estudio, con una función de vinculo logarítmica y el logaritmo de la población como compensación. Los resultados de este estudio mostraron que la mayor incidencia de LC se presentó en temperaturas promedio de 26°C. Este estudió mediante sus hallazgos confirmó el papel del clima y el uso de la tierra en la transmisión de LC [18].

Chaves y colaboradores (2014) estudiaron la asociación del fenómeno de El Niño con los vectores del género *Lutzomyia*. En dicho estudio se utilizaron series de tiempo mensuales con los casos de LC, además de la información de precipitación y temperatura. Se implementaron modelos lineales generalizados con efecto mixto para estudiar los patrones de abundancia de los vectores a través de las fases del ciclo de El Niño y se evidenció una estrecha relación entre la abundancia de vectores y los cambios bruscos de temperatura y la precipitación, lo cual generaría que la cantidad de vectores aumente el potencial de brotes epidémicos [19].

Pérez y colaboradores (2016) identificaron los factores de riesgo ambientales para la LC en Colombia (región andina 715 municipios rurales y urbanos) utilizando 10 años de vigilancia (2000 - 2009) mediante análisis espaciotemporales en modelos de efectos aleatorios autorregresivos condicional de Poisson, para modelar la dependencia de la incidencia de LC en el uso de la tierra, el clima, la elevación y la densidad de población. Los resultados de esta investigación identificaron que las selvas tropicales, los bosques, la vegetación secundaria, la temperatura y la precipitación anual están asociadas positivamente con la incidencia de LC. Además, que en general el clima y el uso de la tierra se pueden utilizar para identificar áreas de alto riesgo de LC [20].

Gutiérrez y colaboradores (2018) estudiaron la incidencia de LC en los departamentos colombianos de Santander y Norte de Santander. Para dicho estudio, estos autores trabajaron con los casos de LC a nivel municipal, durante el periodo de 2007 a 2016; utilizaron una regresión binomial negativa para obtener la tasa de incidencia ajustada para variables ambientales y socioeconómicas. Los resultados de dicha investigación confirmaron la importancia de los determinantes ambientales, como la altura sobre el nivel del mar, la cobertura de bosques, cultivos permanentes y zonas agrícolas heterogéneas para la ocurrencia de LC [21].

Como puede evidenciarse de los trabajos anteriormente descritos, la mayoría de las investigaciones han implementado técnicas estadísticas y análisis de series de tiempo, con el uso de covariables principalmente meteorológicas y ambientales.

El uso de técnicas de inteligencia artificial (IA) para la estimación de la ocurrencia de casos de LC ha sido muy escasa. Sin embargo, se han implementado técnicas de aprendizaje profundo (*deep learning*) para predecir perfiles epidemiológicos a partir de series de tiempo, haciendo uso de redes neuronales recurrentes (RNN) y redes neuronales convolucionales (CNN). La comparación de los resultados de dichas técnicas con respecto a modelos autorregresivos de series de tiempo mostró un mejor desempeño [22]. De acuerdo con la revisión del estado del arte realizada, esta indica que existe un único trabajo en Colombia con datos de dengue en el que se implementaron técnicas de inteligencia artificial (bosques aleatorios, *random forest*, y redes neuronales artificiales) para generar una predicción semanal de la ocurrencia de casos [23].

Lo anterior sugiere la necesidad de implementar técnicas de IA para mejorar la predicción de la ocurrencia de casos de LC, con el objetivo de brindar a las autoridades de salud una herramienta robusta capaz de integrar la información climática, ambiental y socioeconómica asociada con la enfermedad, de tal manera que se pueda contribuir de forma eficiente y oportuna en la prevención de nuevos casos.

Las consecuencias de no disponer de una herramienta predictiva para la ocurrencia de nuevos casos, no solo de LC sino de cualquier otra enfermedad, están asociados a la dificultad de las autoridades de salud de prepararse de forma oportuna a la ocurrencia de

brotes epidémicos, durante los cuales el número de personas que requieren atención es mucho mayor que el esperado, por lo que probablemente no hay suficientes facilidades clínicas ni suficientes medicinas disponibles. Disponer de un modelo de predicción de casos futuros puede mejorar la preparación de los sistemas de salud para la admisión de un número masivo de nuevos pacientes.

Capítulo 5: Metodología

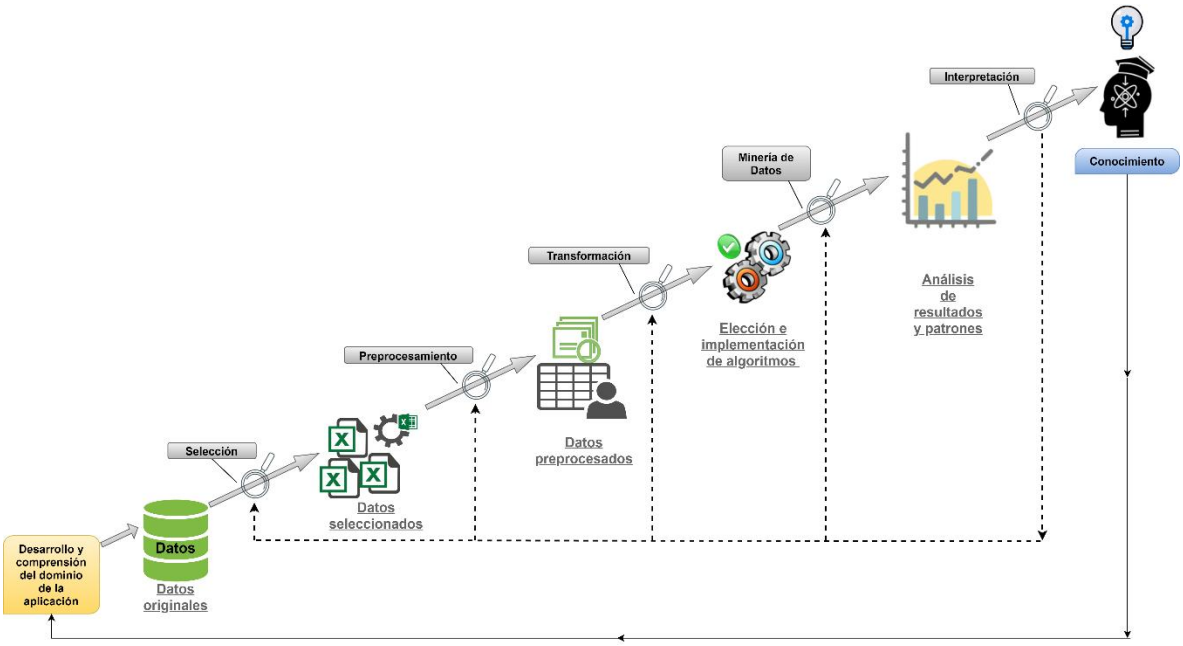
En este capítulo se presenta de manera conceptual el proceso KDD, el cual fue el que se siguió para la metodología de esta investigación; se explica el paso a paso de cada una de las fases realizadas en este estudio y cómo se resolvió cada etapa con la ayuda del proceso KDD.

5.1. Proceso KDD

Para el desarrollo de esta investigación se utilizó la metodología basada en el proceso KDD (*knowledge discovery in databases,* por su sigla en inglés), el cual se refiere a descubrir o encontrar conocimiento útil a partir de datos y este proceso involucra la minería de datos, la cual es un paso en particular en *KDD*. Allí la minería de datos (*data mining*) es utilizada para aplicar algoritmos y extraer patrones de interés de los datos [24].

En la Figura 4 se muestra la metodología seguida en esta investigación, basada en el proceso KDD.

Figura 4: Metodología *KDD* seguida en esta investigación



El proceso de KDD es de naturaleza iterativa e interactiva. Hay varios pasos o etapas diferentes de este modelo y se detallan a continuación [25].

Desarrollo y comprensión del dominio de la aplicación: Es el primer paso de esta metodología y consistió en analizar y definir los objetivos que se querían con la investigación y ciencias de datos aplicada, se definió crear modelos predictivos para la ocurrencia de casos de LC en Colombia a partir un *dataset* con variables ambientales y socioeconómicas.

Selección: Etapa en la cual se procedió a recolectar, crear, organizar y escoger los datos necesarios para la investigación.

Preprocesamiento: En esta fase se tomó el conjunto de datos y se hizo el análisis respectivo para poder eliminar registros con información faltante, datos erróneos y valores inconsistentes.

Transformación: En este paso se hizo la preparación final de conjunto de datos, para lo cual se redujo la dimensionalidad, utilizando la matriz de correlación de Pearson, con un umbral de 0.6 de correlación; se pasó de 65 variables a 45 variables predictoras. Y,

finalmente, se hizo la estandarización de los datos mediante z-score y la organización de las matrices de corrimientos para el caso de los modelos con series de tiempo.

Minería de datos: Esta etapa consistió en la definición de la tarea de minería de datos, para este trabajo fue la clasificación binaria para el conjunto de datos municipal a nivel mensual, regresión con el *set* de datos departamental a nivel mensual y regresión con series de tiempo para el *dataset* semanal a nivel departamental. Para esto se hizo una selección de los algoritmos de minería de datos a aplicar, particularmente, *XGBoost* (basados en árboles de decisión), redes neuronales (perceptrones multicapa), máquinas de soporte vectorial, *naive bayes*, árboles de decisión y bosques aleatorios.

Interpretación: Fase que consistió en realizar la evaluación del desempeño de los modelos implementados previamente, se interpretaron y analizaron los resultados; de igual forma, acá se hizo validación cruzada y análisis de métricas de los algoritmos para corroborar que sí se cumple con el objetivo propuesto.

Conocimiento: Es la última fase de este proceso y consistió en el uso de los resultados y los modelos obtenidos en esta investigación. Es así como se procedió a realizar una aplicación web que permitió hacer uso de los modelos predictivos construidos, esto con el fin de poder ser usados por organismos de la salud en Colombia para tomar medidas preventivas en municipios o departamentos del país.

Es importante aclarar que, al ser un proceso iterativo y cíclico [26], en cualquier etapa de este si no se obtienen los resultados esperados es posible devolverse a etapas previas en cualquier momento y volver a realizar el proceso en estas fases.

Capítulo 6: Resultados y discusión

El presente capítulo muestra los resultados obtenidos a lo largo de la investigación, partiendo desde la recolección y construcción del conjunto de datos hasta llegar a la definición de los modelos predictivos y sus métricas obtenidas. También, se menciona las diferentes técnicas e intentos por obtener mejores resultados, los cuales se pueden ver en más detalle en la sección de anexos correspondiente. Paralelamente también se presenta la discusión de los resultados.

6.1. Generación del conjunto de datos y abstracción de la información

En esta sección se explica el desarrollo de la primera fase del trabajo propuesto, la cual se centró en la construcción del conjunto de datos con todas las variables a utilizar y a partir de ello se realizó el tratamiento de la información.

Antes de construir el conjunto de datos original que se utilizó para crear los modelos predictivos, se inició esta investigación con un *dataset* de pocas variables ambientales y socioeconómicas, con datos registrados a mano en un archivo de Excel obtenidos de diferentes fuentes. Sin embargo, al empezar a utilizar dicho conjunto y hacer el respectivo análisis estadístico se encontraron muchas inconsistencias, datos faltantes, erróneos, etc. Esto llevó a que se tuviera que construir un nuevo *dataset* con información descargada directamente de las fuentes originales y confiables, que se almacenaron automáticamente en un archivo el cual fue utilizado para el desarrollo de la investigación.

6.1.1. Identificación de las fuentes de información

Para la obtención de la información, se investigaron las diferentes fuentes que pudieran contener datos de interés y se procedió a la descarga de estos. Las fuentes identificadas fueron: el Departamento Nacional de Planeación (DNP), el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), el Sistema Nacional de Vigilancia en la Salud Pública (SIVIGILA), de bases de públicas de la Administración Nacional de la Aeronáutica y del Espacio (NASA) y de bases de datos atmosféricos de la Universidad de Arizona.

El Anexo 1 muestra la descripción y caracterización de las variables incorporadas, e información detallada de cada variable.

6.1.2. Descarga de la información de las bases de datos

La descarga de información se inició con los datos de las fuentes locales, de la página de SIVIGILA se descargó la información de los casos de LC de 2007 a 2016 con periodicidad mensual y diaria que luego fue agrupada en semanas, para cada municipio. Esta página genera un archivo en formato de Excel con los datos de casos junto con la población de cada municipio. De la misma forma del IDEAM y del DANE se obtuvieron variables ambientales y socioeconómicas respectivamente (Anexo 1).

Las variables con resolución temporal mensual y temporal diaria, correspondieron al índice mejorado de vegetación (EVI) que fue descargado desde la Universidad de Arizona (producto: VIP01P4), la precipitación y temperatura, que se obtuvieron de la NASA (productos: M2SDNXSLV y GLDAS_NOAH025_M) (Anexo 1).

La descarga automatizada de las variables temporales se realizó mediante *scripts* que fueron construidos en el lenguaje de programación R, los cuales descargaban la información de tipo *raster* de los correspondientes servidores, a partir del rango de coordenadas seleccionadas y que correspondía a todo el territorio nacional (Spatial Coverage: -82.222,-4.638,-66.313,16.808).

Debido a que la resolución espacial de los archivos en formato ráster obtenidos correspondió a 25 km, los datos en formato ráster fueron acoplados con un mapa vectorial a nivel municipal. Así, cada municipio tenía asociado uno o más píxeles con los datos de cada una de las variables de temperatura, precipitación e índice de vegetación. Para poder obtener el valor de la variable por municipio, cuando este coincidía con más de 1 pixel en su ubicación, se realizó un promedio ponderado, donde el pixel que tenía más cobertura del municipio tuvo más peso en el promedio y los que tenían menos cobertura del municipio, tuvieron menos peso. Los pesos se obtuvieron gracias a la función *extract* de R, la cual suministra el peso a los pixeles, de acuerdo con su cobertura en el polígono manipulado.

Las unidades de temperatura y precipitación fueron transformadas, para el caso de la temperatura, los valores originales en grados Kelvin se convirtieron a grados Celsius (centígrados), para el caso de la precipitación, se convirtió de $\frac{kg}{m^2 \times s}$ a $\frac{mm}{dia}$ (milímetros por día).

Por último, las variables temporales que fueron descargadas originalmente a escala diaria se agruparon para obtener valores promedio para cada semana epidemiológica y cada mes.

El conjunto de datos consistió de dos *datasets*, uno con datos mensuales en el periodo de estudio (2007 - 2016) y otro con datos semanales (semanas epidemiológicas) en el mismo periodo, en ambos casos, la variable objetivo fue el número de casos de LC. Ambos *datasets* fueron objeto de análisis en las fases posteriores.

Las variables predictoras fueron discriminadas en variables de contexto y variables del nivel de observación. Las variables de contexto corresponden a variables que tienen el mismo valor para un mismo municipio durante todo el periodo de estudio y describen el contexto ambiental y socioeconómico de los municipios. Las variables del nivel de observación corresponden a variables con mediciones semanales o mensuales (precipitación, temperatura e índice EVI) que coinciden en su resolución temporal con la variable repuesta cuando esta es semanal o mensual.

6.2. Análisis descriptivo de los datos

Cuando ya se logró obtener el conjunto de datos, se procedió a realizar un análisis descriptivo detallado del mismo, esto con el fin de conocer cada una de las variables, identificar valores erróneos, inconsistentes, faltantes y saber cómo es el comportamiento de cada variable y cómo son en relación con la variable respuesta. Esto se explica a continuación.

6.2.1. Descripción detallada del conjunto de datos.

Después de haber obtenido el conjunto de datos tanto mensual como semanal, se procedió a analizarlo detalladamente en Python, usando la librería de pandas y otras para facilitar su visualización y descripción.

Se partió de dos *datasets,* el mensual con **65 columnas** y **131,280 filas** y el *dataset* semanal con **65 columnas** y **571,068 filas.** Se analizó cada una de las variables usando algunos estadísticos básicos como la media, la desviación estándar, su valor máximo y mínimo. A continuación, se muestran estadísticas para las variables de observación y la variable respuesta tanto para el conjunto de datos semanal como para el mensual (ver Tabla 1).

En la Tabla 1 se puede observar como para la variable *casos* la media es muy cercana a cero (*dataset* mensual y semanal), lo que indica que se tienen desde el principio muchos registros con valor cero, siendo esta una variable discreta que toma valores de 0 a 228 (*dataset* semanal) y valores de 0 a 509 (*dataset* mensual).

El valor máximo del número de casos por semana correspondió a la semana 26 (mes de julio) del año 2009 para el municipio de La Macarena en el departamento del Meta. Por otro lado, por ejemplo, para la variable temperatura se puede detallar que su valor promedio fue de 20.595 (set semanal), lo que indica que es la temperatura promedio en general de Colombia durante los 10 años de estudio. El valor mínimo de temperatura correspondió a la primera semana de 2015 en varios municipios de los departamentos de Boyacá y Santander. El valor máximo de temperatura fue registrado en el municipio de San Fernando en Bolívar en la semana 12 del 2016.

Tabla 1: Estadísticos básicos de las variables de observación y variable respuesta para los *datasets* semanal y mensual.

Conjunto de datos semanal				
Variable	**Media**	**Desviación estándar**	**Valor mínimo**	**Valor máximo**
Casos	0.158	1.151	0	228
Temperatura	20.595	4.883	9.638	35.276
Precipitación	61.251	52.691	45.932	541.775
Índice de vegetación EVI	0.422	0.077	-0.296	0.671
Conjunto de datos mensual				
Variable	**Media**	**Desviación estándar**	**Valor mínimo**	**Valor máximo**
Casos	0.688	4.406	0	509
Temperatura	19.429	5.850	3.349	32.755
Precipitación	211.352	124.486	0	2039.617
Índice de vegetación EVI	0.451	0.077	0.021	0.694

La Tabla 2 muestra los 10 municipios con más casos reportados en todo el periodo de estudio. El municipio con más casos fue La Macarena en el departamento del Meta. De la misma forma, los departamentos de Tolima, Antioquia y del Meta son los que tienen más municipios entre los primeros en dicho ranking.

Tabla 2: 10 municipios con mayor cantidad de casos en todo el periodo de estudio

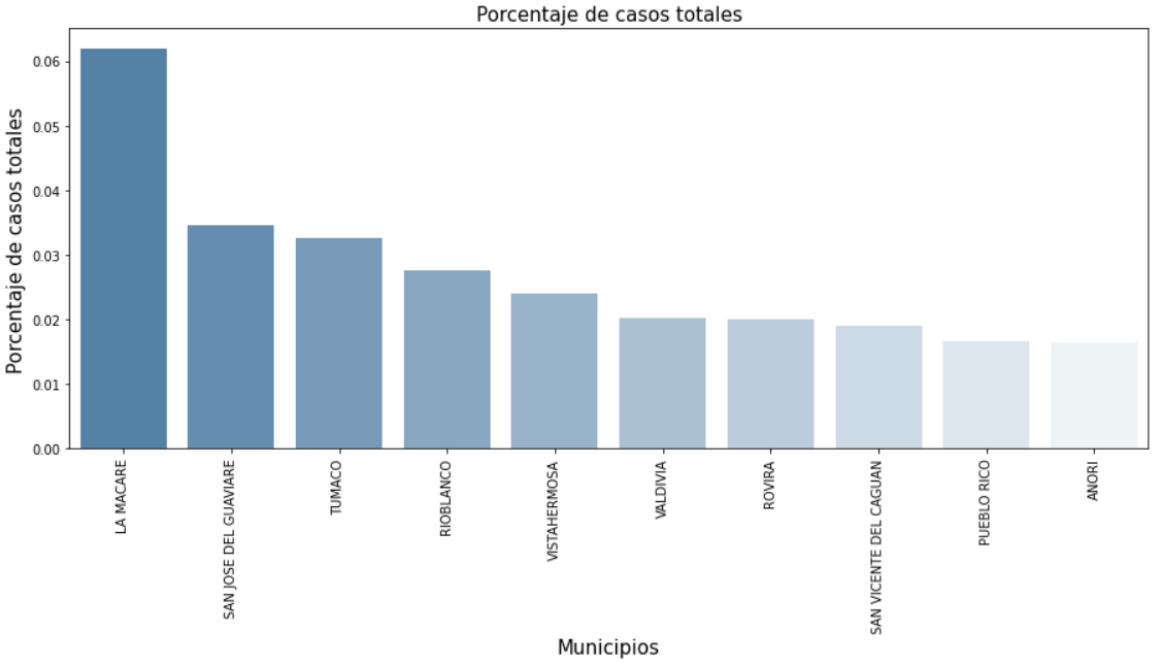
Municipio	Departamento	Total de casos
La Macarena	Meta	5,598
San José del Guaviare	Guaviare	3,123
Tumaco	Nariño	2,938
Rio Blanco	Tolima	2,490
Vista Hermosa	Meta	2,172
Valdivia	Antioquia	1,836
Rovira	Tolima	1,836
San Vicente del Caguán	Caquetá	1,719
Pueblo Rico	Risaralda	1,504
Anorí	Antioquia	1,494

También, se obtuvieron los 5 registros con mayor cantidad de casos en una semana, la mayoría de dichos registros correspondieron al municipio de La Macarena en el Meta, entre los meses de julio y agosto de 2009 (Tabla 3). De los 90,382 casos reportados en Colombia para los 10 años de estudio, más del 6% de dichos casos se reportaron en La Macarena (Figura 5).

Tabla 3: Mayor cantidad de casos en fechas específicas a nivel semanal

Año	Mes	Semana	Total de casos
2009	Julio	26	228
2010	Enero	1	141
2009	Agosto	32	130
2009	Junio	22	119
2009	Julio	27	119

Figura 5: Municipios con mayor porcentaje de casos reportados



Antioquia fue el departamento que más casos reportó (más del 20% del total), seguido por Meta (ver Figura 6). Las figuras 7 y 8 muestran la incidencia por 10,000 habitantes, tanto a nivel municipal para los 10 municipios con más casos, como a nivel departamental. La Macarena presentó la mayor incidencia a escala municipal (Figura 7) y Guaviare fue el departamento con mayor incidencia (ver Figura 8).

Figura 6: Porcentajes de casos por departamento

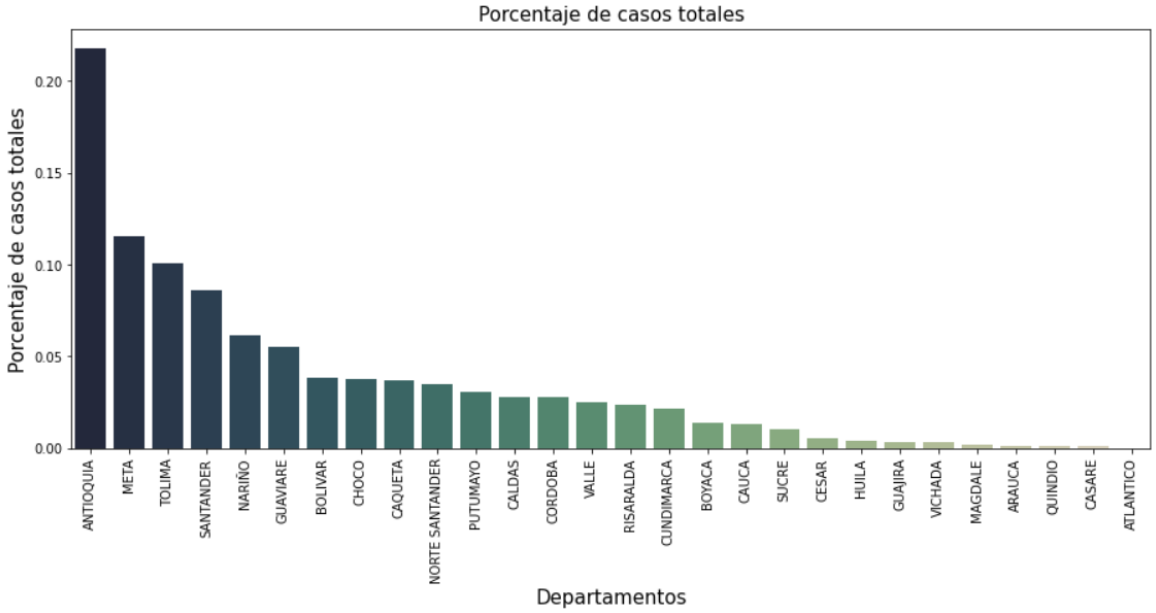


Figura 7: 10 municipios con mayor incidencia

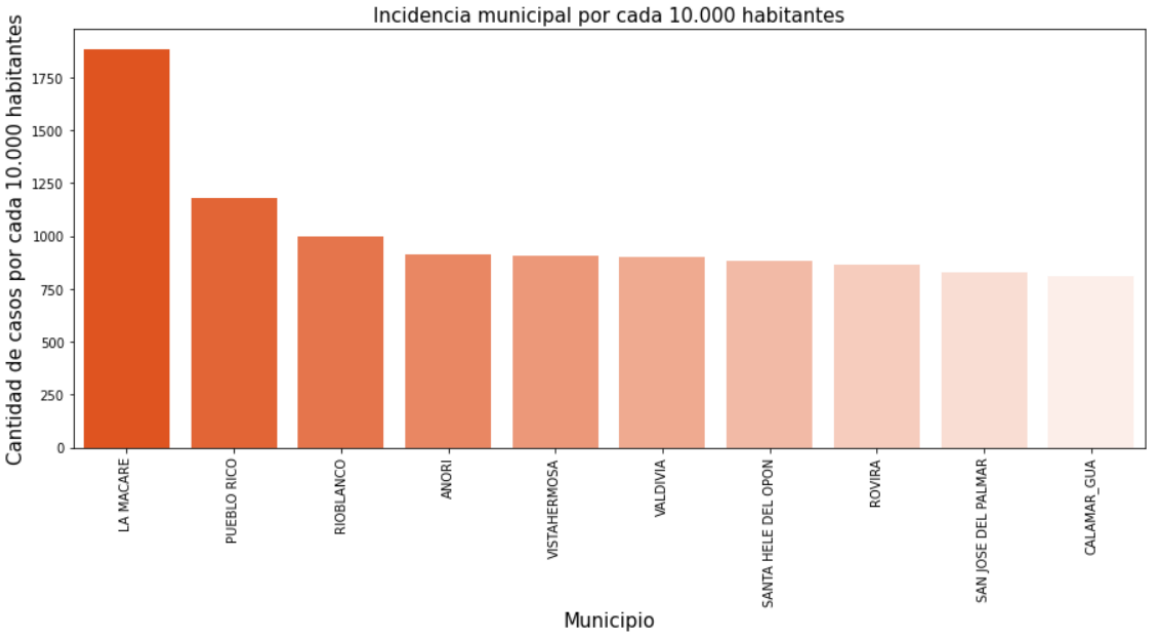
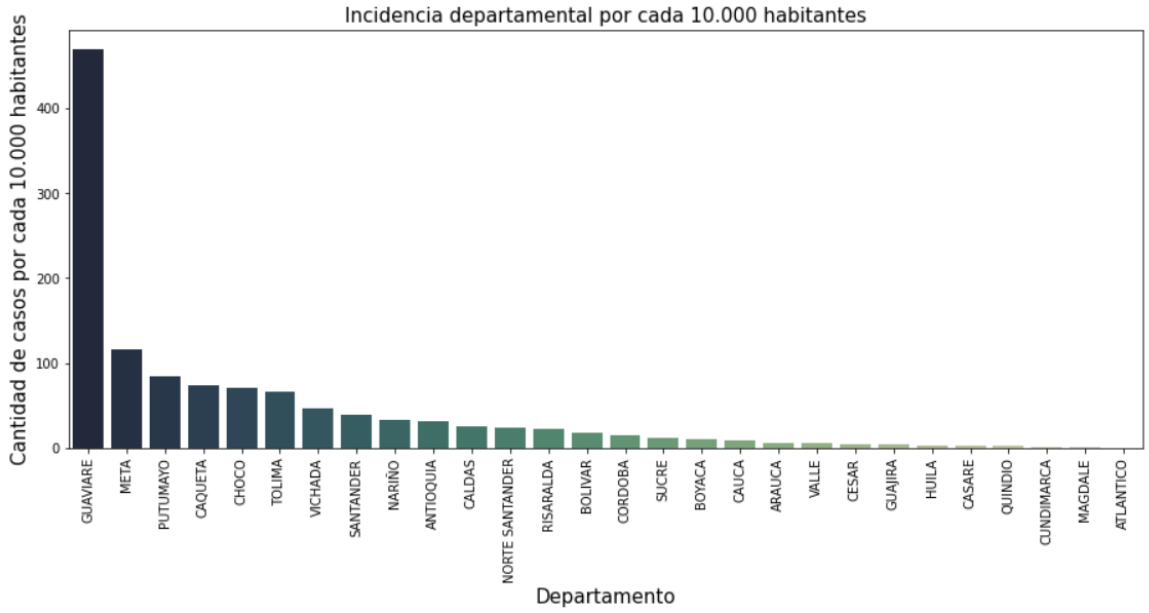


Figura 8: Departamentos con mayor incidencia



En el *dataset* semanal la cantidad de municipios que no reportaron ni un sólo caso de LC en los 10 años fue de 215 municipios (19.65%). En este *dataset* la variable respuesta

presentó un 92.95 % de registros igual a cero. Por otro lado, en el conjunto de datos con periodicidad mensual, el 83.98% de los registros reportaban cero casos.

Los registros con valores faltantes se concentraron en los corregimientos departamentales, estos corregimientos tuvieron que ser analizados en un *dataset* aparte, denominado *dataset* de corregimientos, debido a que presentaban 18 variables socioeconómicas con datos faltantes. En el conjunto de datos de localidades en departamentos con corregimientos, la localidad con más casos reportados fue Mitú y el corregimiento que más casos presentó fue Barrancominas (ver Figura 9). Mientras que el departamento que más casos presentó fue el Vaupés (ver Figura 10).

Figura 9: Porcentaje de casos en el *dataset* de localidades en departamentos con corregimientos

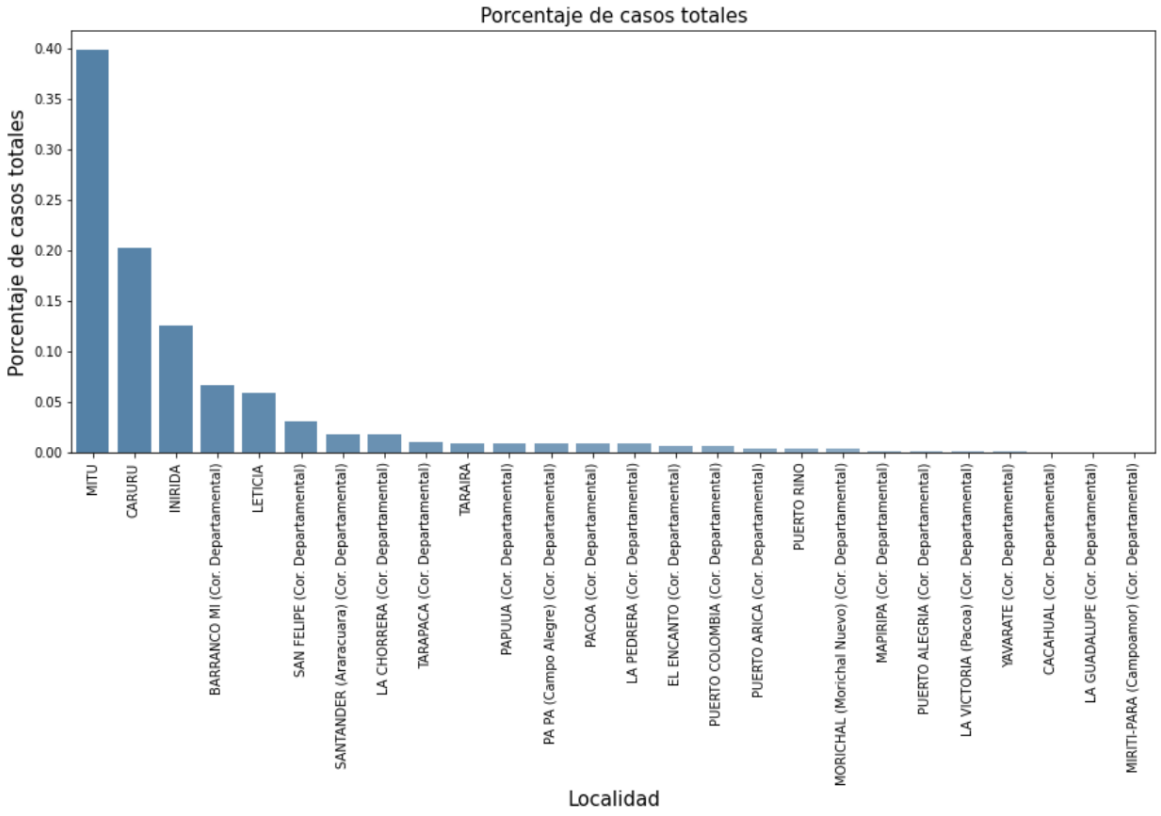
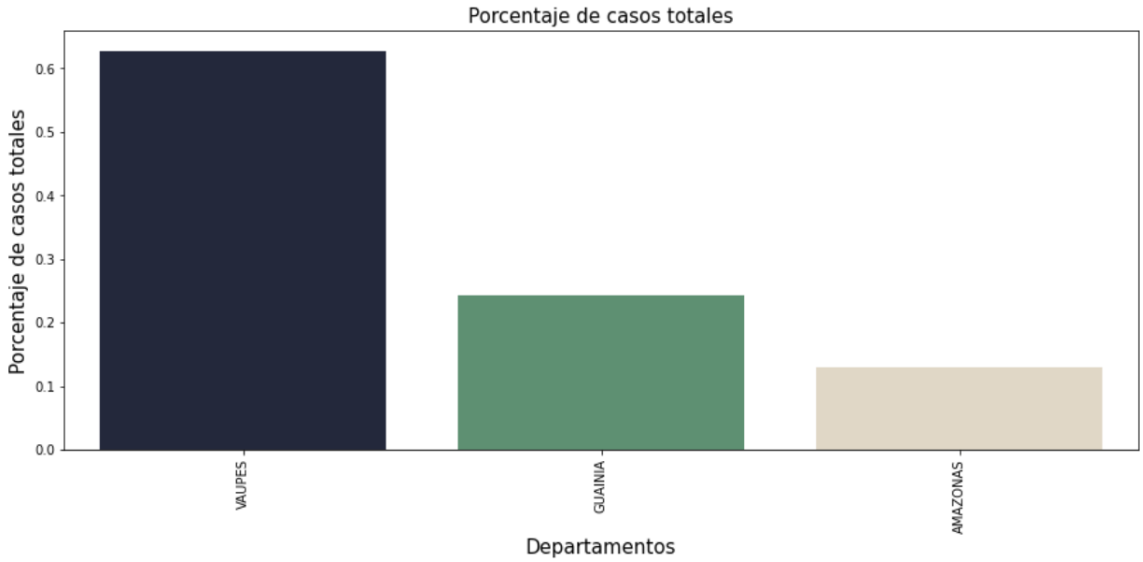


Figura 10: Porcentaje de casos por departamento en el *dataset* de corregimientos



Las figuras 11 a 16 muestran los histogramas para las variables temperatura, precipitación e índice EVI, respectivamente, tanto para el *dataset* semanal como para el *dataset mensual.*

Para el histograma de temperatura presentado en las Figuras 11 y 12 se puede observar que la mayoría de los registros de temperatura se encuentra entre 15 y 26 °C, con una mayor frecuencia de registros de 18°C en el *dataset* semanal (ver Figura 11) y en el *dataset* mensual 25°C (ver Figura 12). En los histogramas de precipitación (Figura 13 y Figura 14) se puede ver que el mayor número de observaciones correspondió a semanas con precipitación menor a 100 mm o meses con menos de 200 mm de precipitación. Para el caso del índice EVI (Figura 15 y Figura 16) el mayor valor de frecuencia se encuentra entre 0.4 y 0.5, para los dos *datasets*.

Figura 11: Histograma de temperatura para el *dataset* semanal

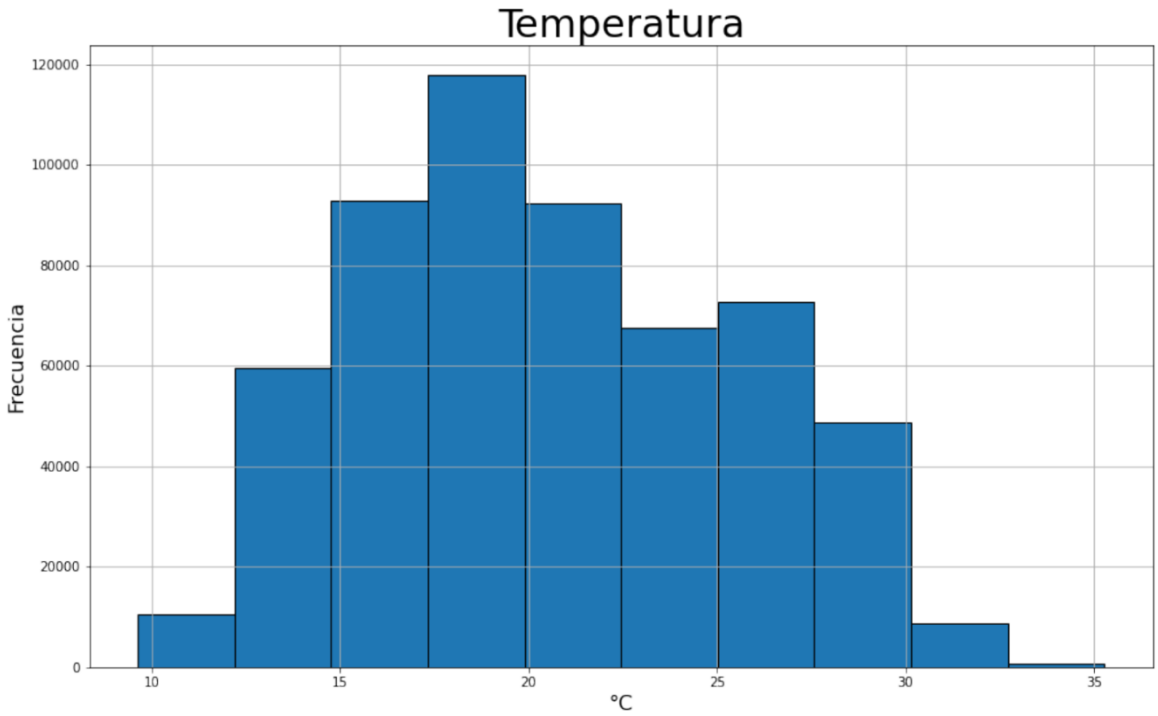


Figura 12: Histograma de temperatura para el *dataset mensual*

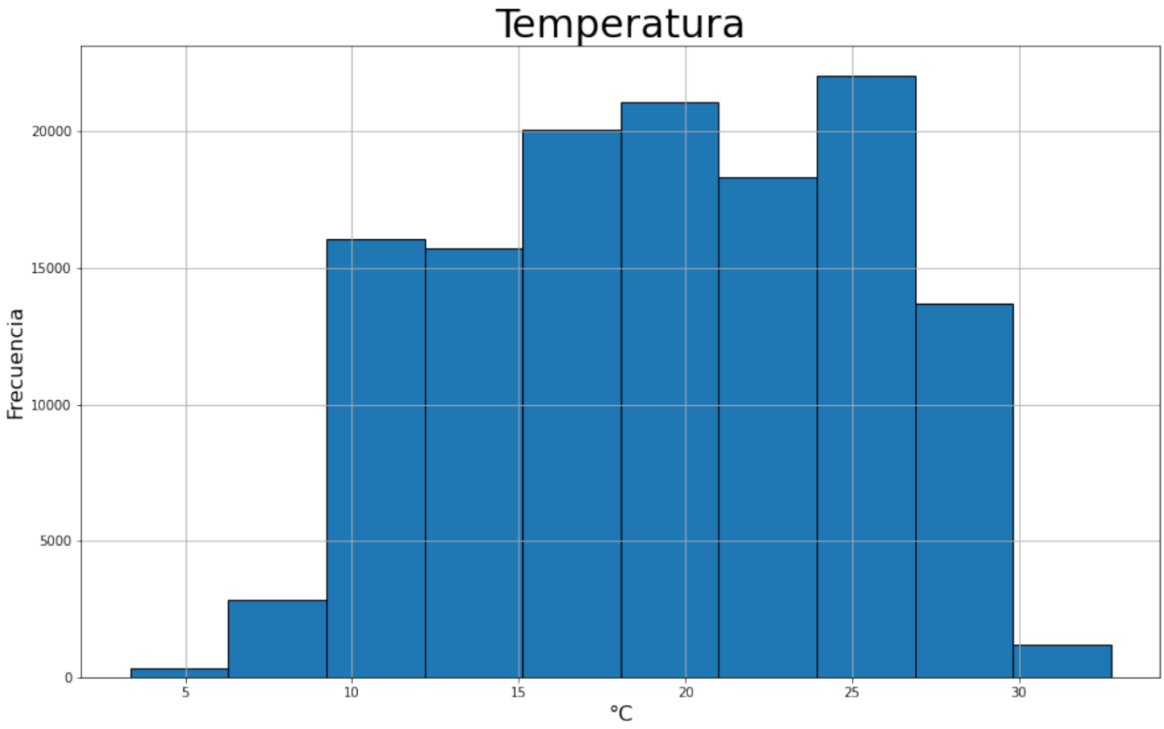


Figura 13: Histograma de precipitación para el *dataset semanal*

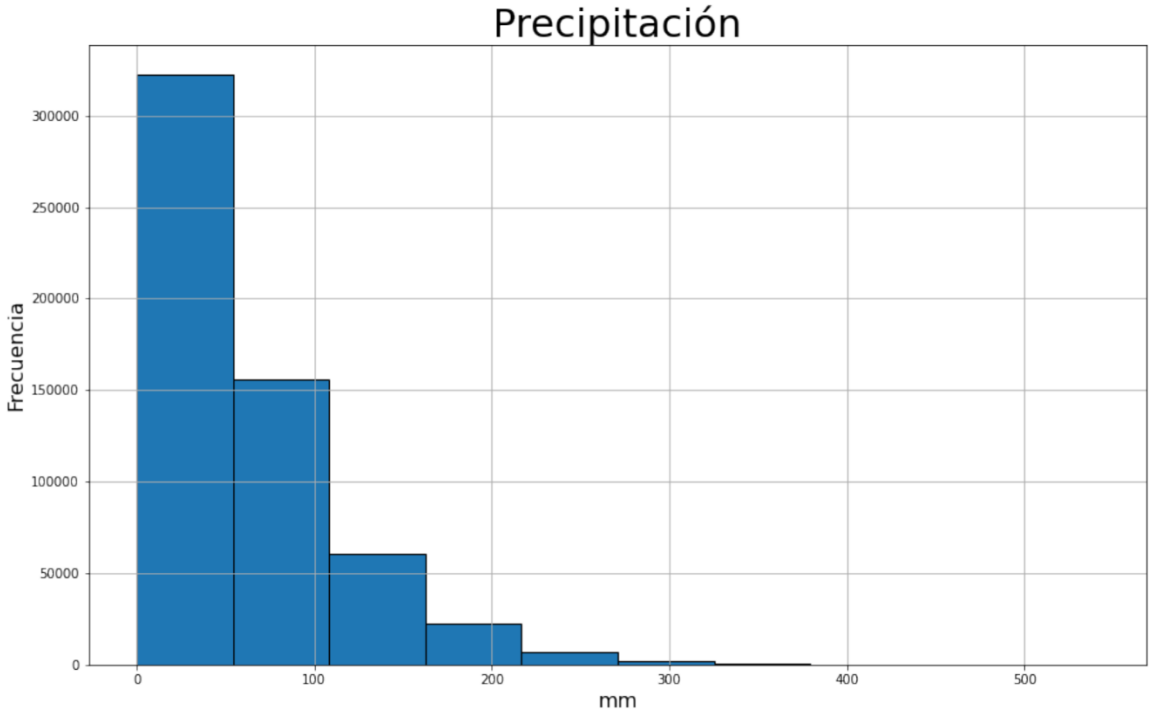


Figura 14: Histograma de precipitación para el *dataset mensual*

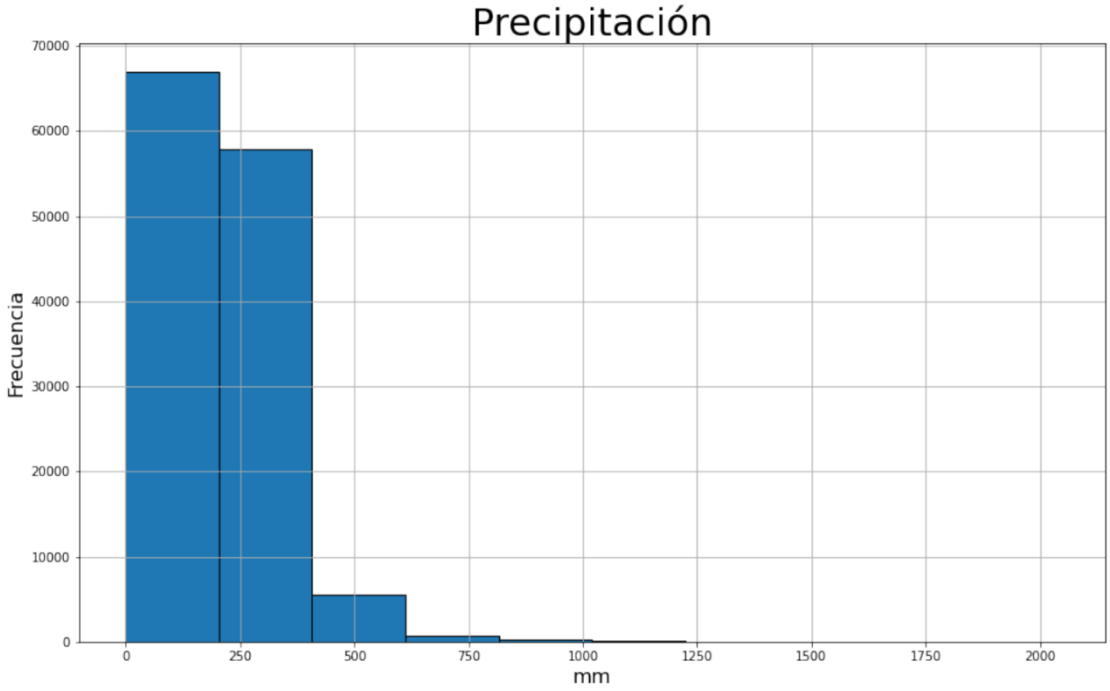


Figura 15: Histograma de índice de vegetación EVI para el *dataset semanal*

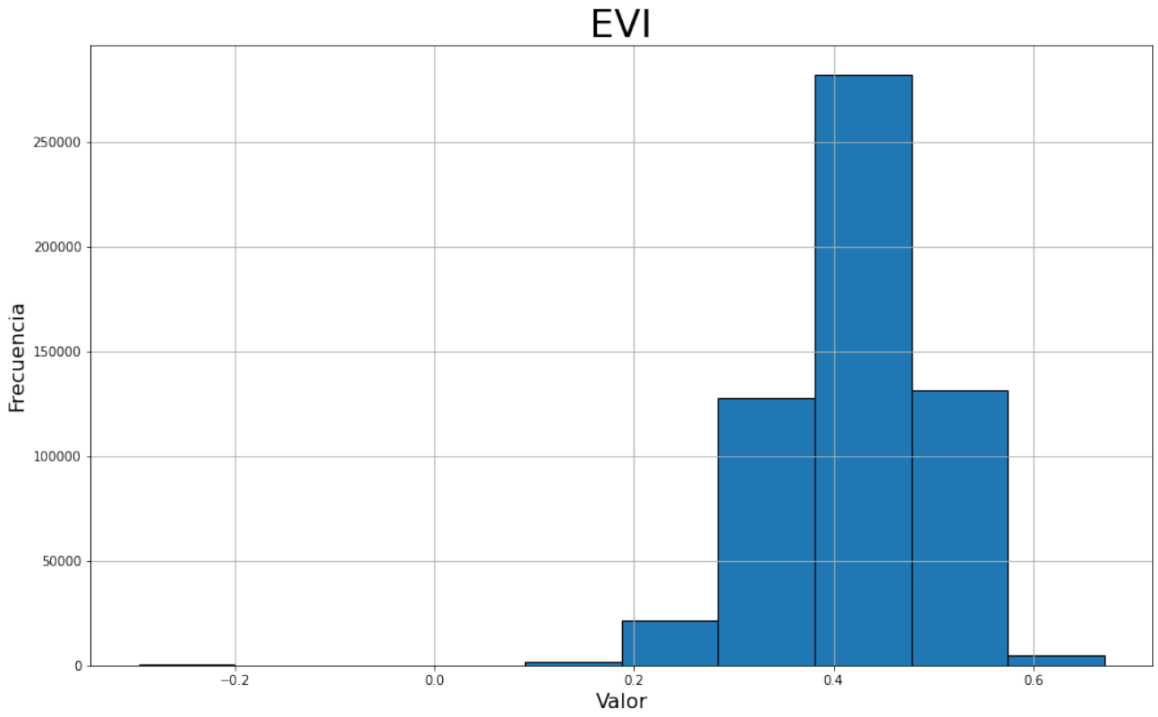
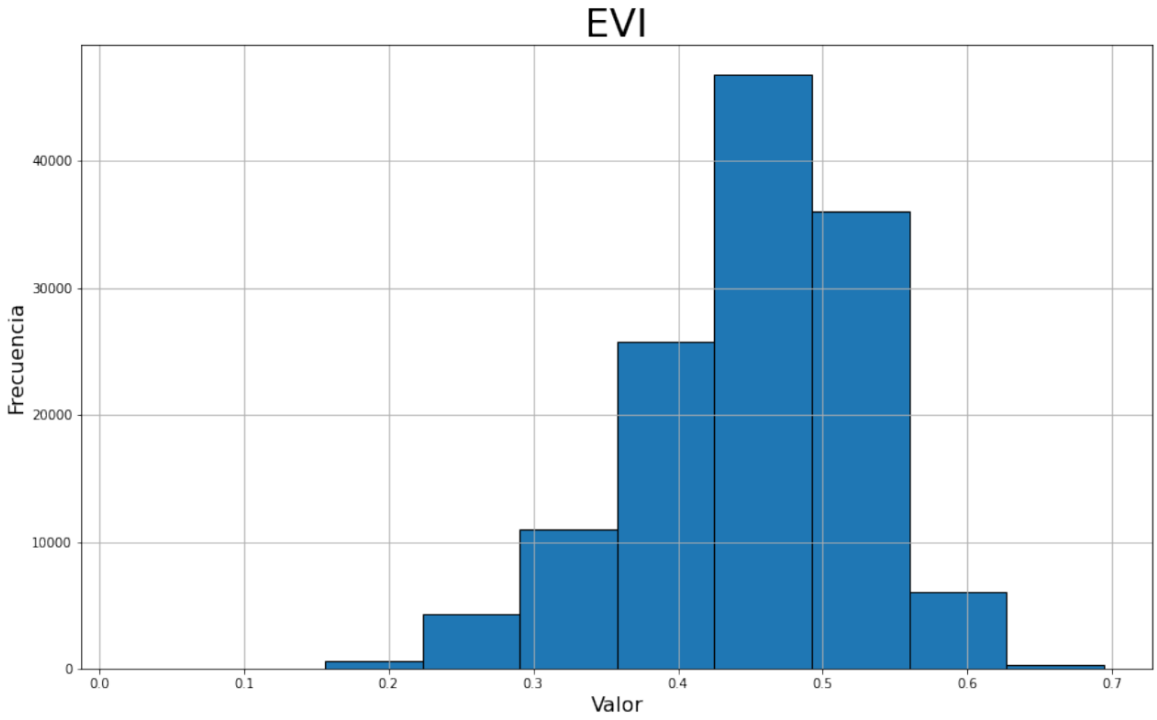


Figura 16: Histograma de índice de vegetación EVI para el *dataset mensual*



Los histogramas de casos muestran que la mayoría de registros se concentran en valores entre 0 y 20 casos para el *dataset* semanal (Figura 17) y 0 y 50 casos para el *dataset* mensual (Figura 18).

Figura 17: Histograma de casos de leishmaniasis para el *dataset semanal*

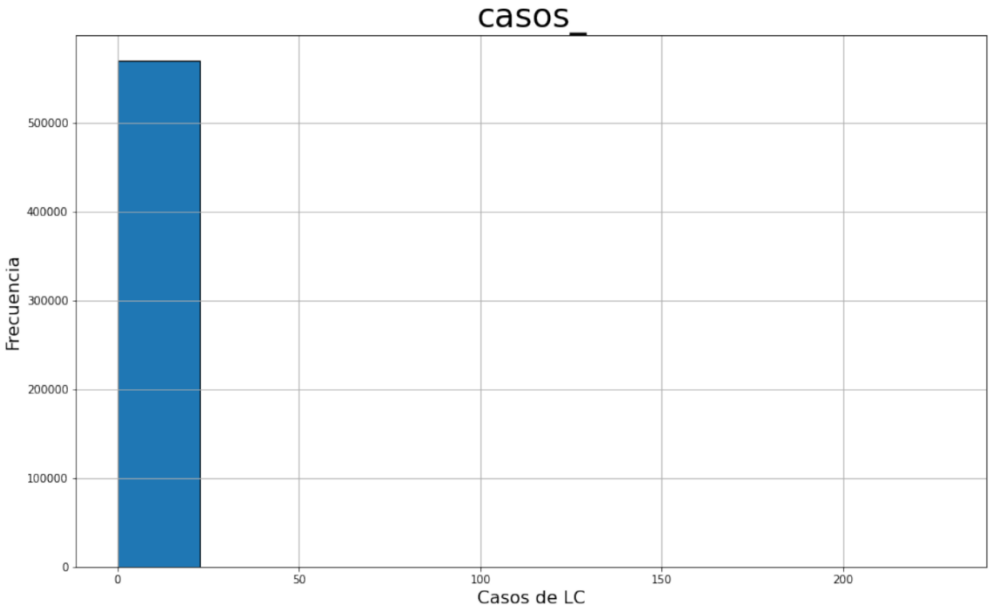
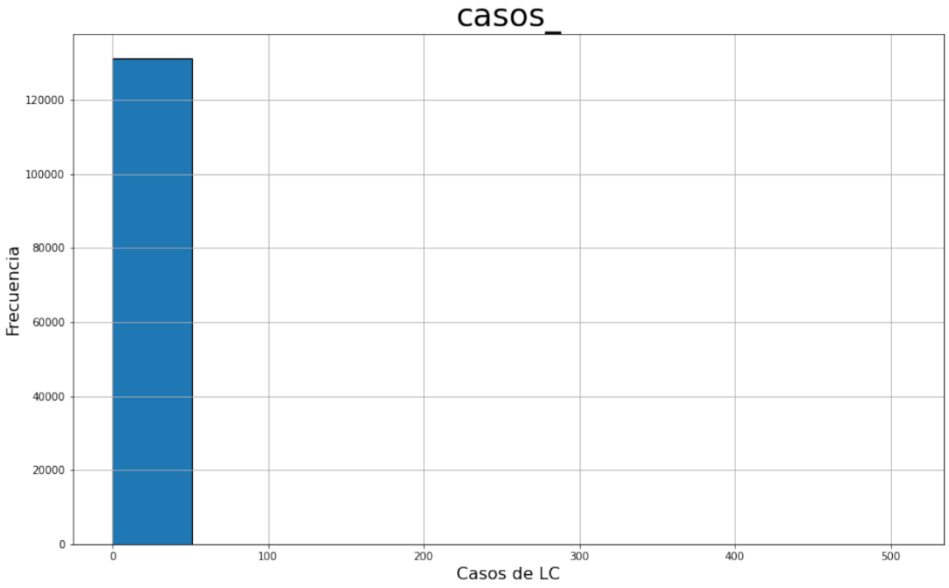


Figura 18: Histograma de casos de leishmaniasis para el *dataset mensual*



En los *boxplots* de la Figura 19 y Figura 20 de casos, se detalla como para la variable objetivo, la mayor cantidad de registros se concentran en cero, sin embargo, hay gran cantidad entre 0 y 60 casos para el *dataset* semanal (Figura 19) y entre 0 y 150 casos para el *dataset* mensual (Figura 20). El valor máximo de casos reportados en una misma semana fue superior a 200 (Figura 19), y a nivel mensual, el valor máximo fue superior a 500 (Figura 20).

Figura 19: *Boxplot* de casos para el *dataset* semanal

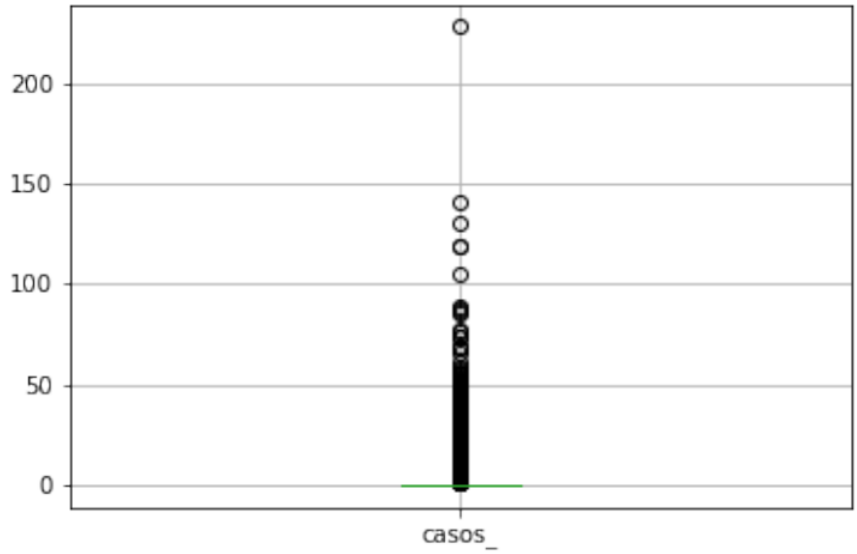
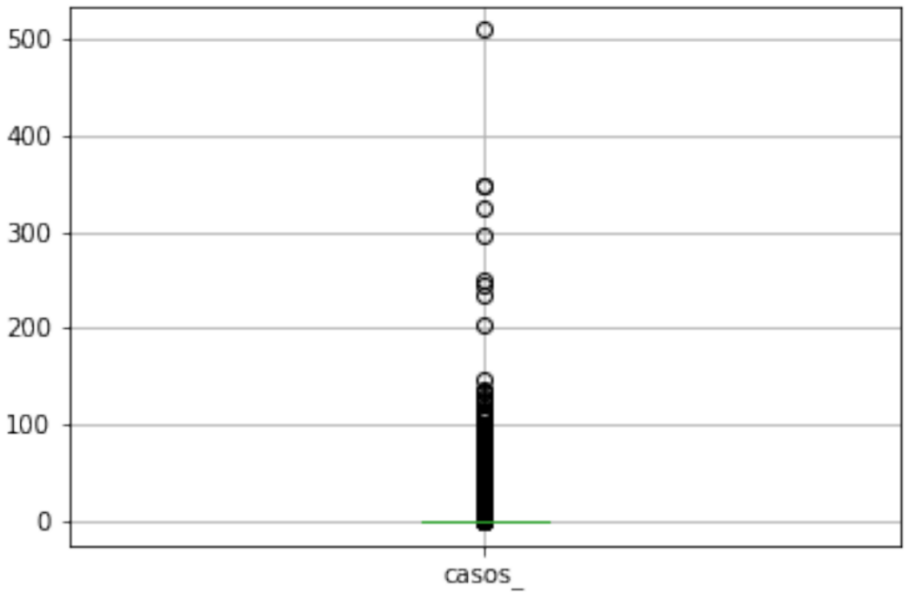


Figura 20: *Boxplot* de casos para el *dataset* mensual



La mayor cantidad de casos se concentra en valores altos del índice EVI (Figura 21 y Figura 22). Mientras que con menor precipitación se presentó un mayor número de casos de LC (Figuras 23 y 24).

Figura 21: Diagrama de dispersión de casos vs. índice de vegetación (EVI) para el *dataset* semanal

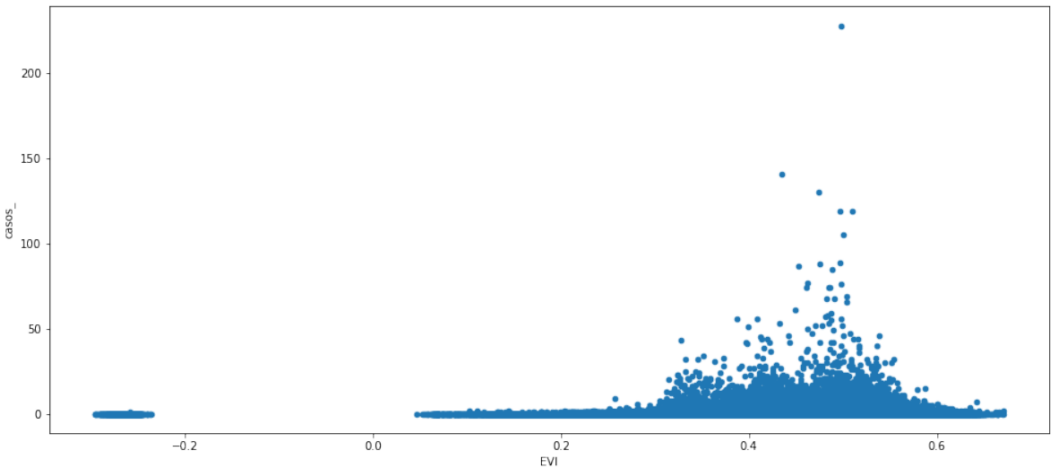


Figura 22: Diagrama de dispersión de casos vs. índice de vegetación (EVI) para el *dataset* mensual

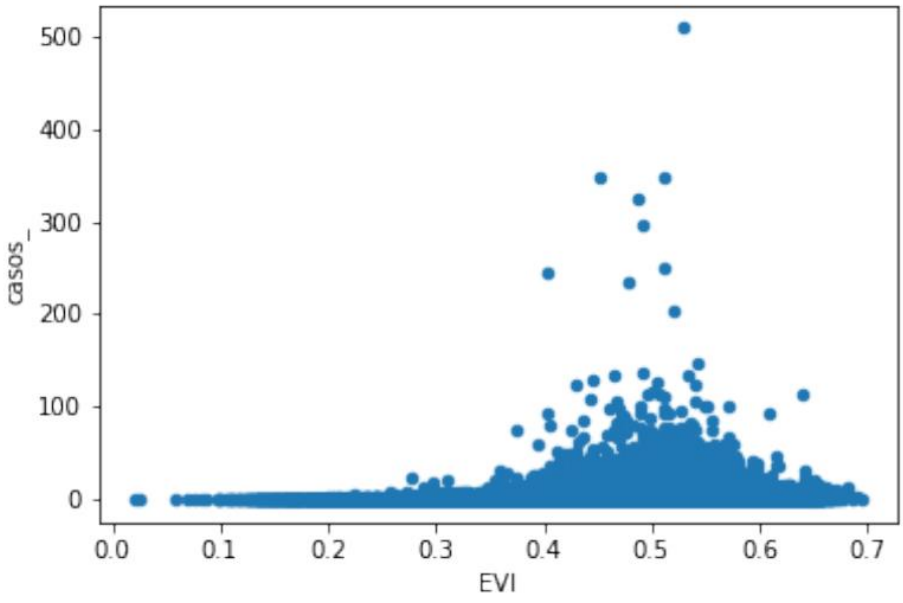


Figura 23: Diagrama de dispersión de casos vs. precipitación para el *dataset* semanal

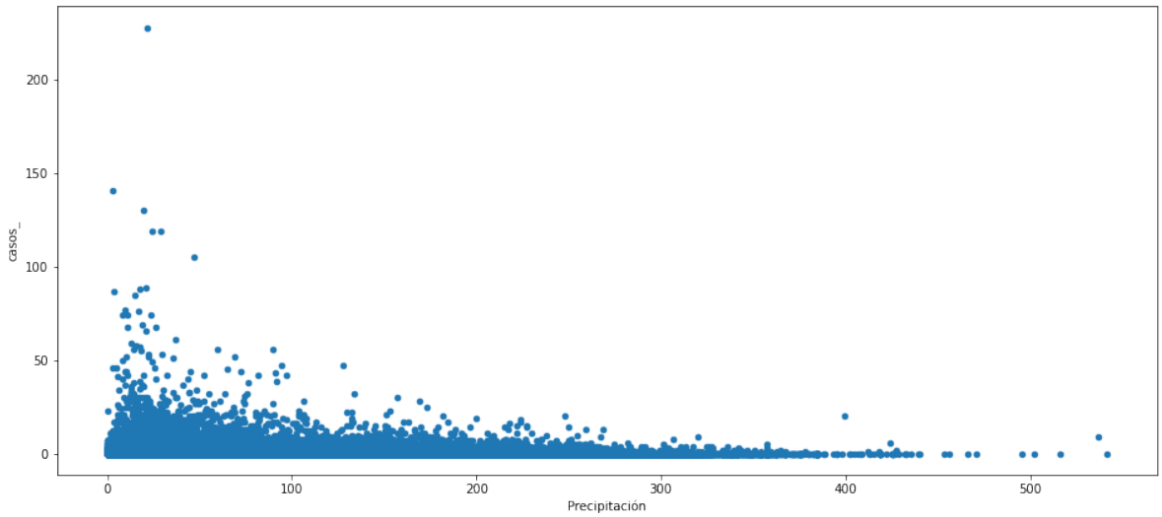
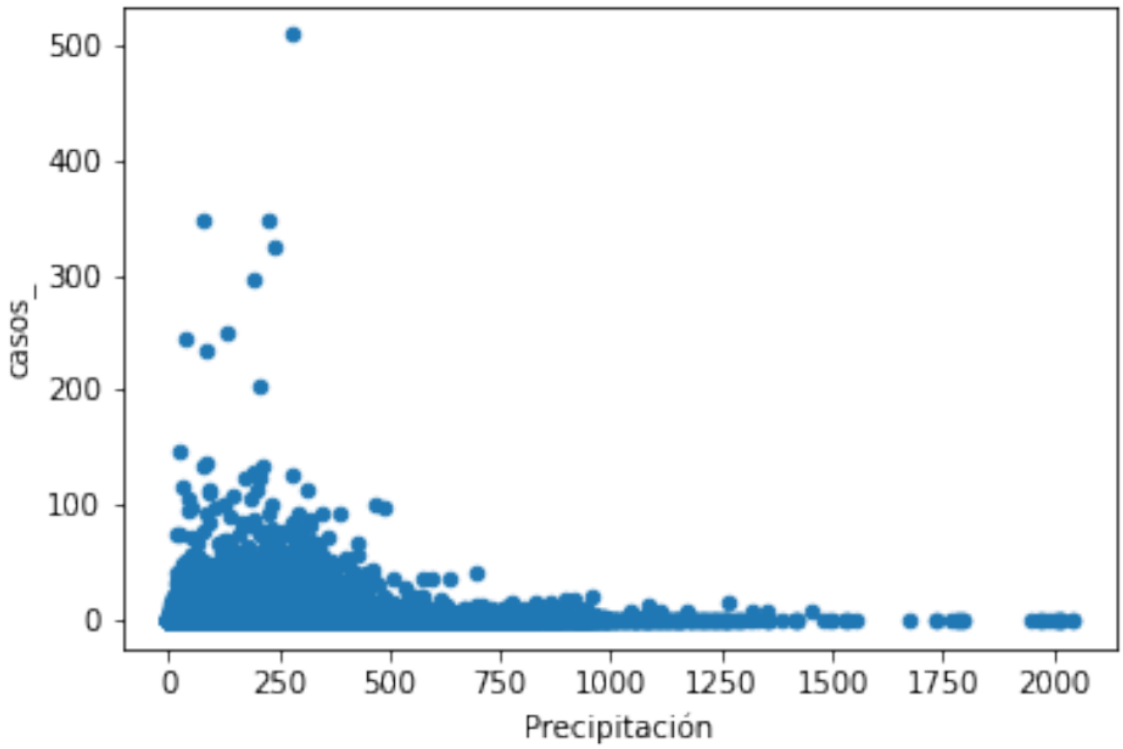


Figura 24: Diagrama de dispersión de casos vs. precipitación para el *dataset* mensual



La mayoría de los casos reportados se presentaron en el rango de temperatura entre 18°C y 28°C. El reporte más alto de casos se presentó en temperaturas por encima de 23°C y hasta 28°C para el *dataset* semanal (Figura 25) y entre 16°C y 26°C para el *dataset* mensual (Figura 26).

Figura 25: Diagrama de dispersión de casos vs. temperatura para el *dataset* semanal

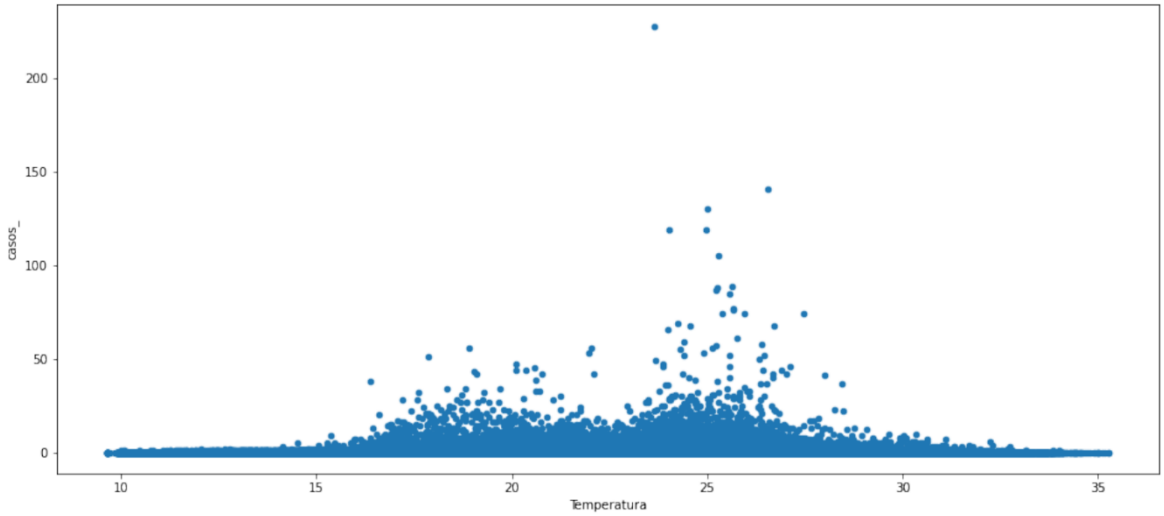
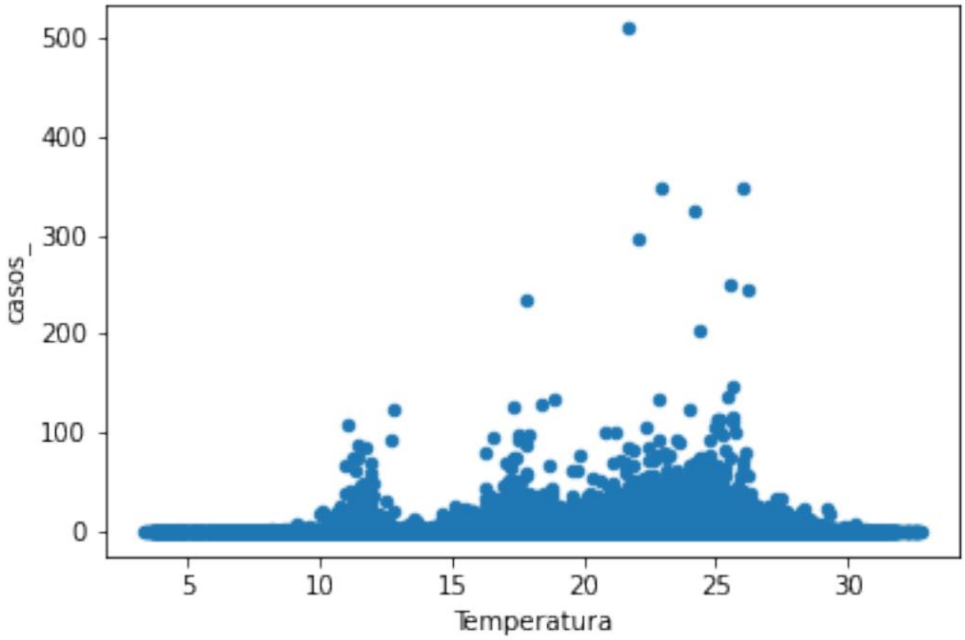


Figura 26: Diagrama de dispersión de casos vs. temperatura para el *dataset* mensual



También se hizo un análisis de las series de tiempo de las variables de observación y la variable respuesta, tanto para el municipio con más casos reportados (La Macarena), como para el departamento de Antioquia, que fue el departamento con más casos (ver Anexo 2). En el Anexo 2.1 se observa como los casos reportados para el municipio de La Macarena en el Meta tuvieron un pico entre la semana 100 a la 180, lo que corresponde al año 2009. En el Anexo 2.2 se ve cómo en general la temperatura presentó valores similares durante

todo el periodo de estudio, tomando valores entre 20°C y 30°C. Si se compara con el Anexo 2.1, se puede evidenciar que las semanas donde ocurrieron más casos corresponden a los valores más altos de temperatura presentados (ver Anexo 2.2). Respecto de la precipitación en La Macarena, se observa un comportamiento constante de la lluvia, excepto en el año 2017 donde hubo un aumento considerable de la misma (Anexo 2.3). El gráfico del EVI (Anexo 2.4), permite detallar que los años cercanos a 2017 presentaron un mayor aumento en el índice EVI, sin embargo, esto no parece aumentar la cantidad de casos en dicho municipio.

El Anexo 2.5 muestra la serie de tiempo de casos de Antioquia. El pico más alto de casos fue entre las semanas consecutivas 160 y 170 que corresponden a los meses de febrero, marzo y abril del año 2009 (Anexo 2.5). Referente a la temperatura (Anexo 2.6), se observa que en las mismas semanas donde hubo mayor cantidad de casos en el departamento, también se presentó un incremento considerable de temperatura, superando los 21 grados centígrados. Igualmente, en Antioquia durante esas mismas semanas consecutivas se presentó una disminución de la precipitación, seguida de un aumento de los casos reportados. Finalmente, la serie de tiempo del índice EVI (Anexo 2.8), inició el periodo de estudio con un valor por debajo de 0.43, seguido de una disminución hasta 0.38 y a partir del año 2008 presentó un comportamiento prácticamente cíclico alrededor de 0.45 durante el resto del periodo de estudio.

6.3. Preprocesamiento, reducción de dimensionalidad y tarea de minería de datos

En esta sección se presentan las diferentes técnicas utilizadas para reducción de dimensionalidad, el método de estandarización utilizado y las tareas de minería de datos seleccionadas.

6.3.1. Transformación de datos y reducción de la dimensionalidad.

Para el conjunto de datos mensual y semanal se realizó una estandarización mediante el Z-Score, de acuerdo con la siguiente fórmula:

$$Z = \frac{X - \mu}{\theta}$$

donde X hace referencia al valor en cuestión al que se le va a realizar la estandarización, μ es la media de los datos y θ la desviación estándar de los mismos.

Para la reducción de la dimensionalidad se probaron varias técnicas, una de ellas fue la matriz de correlación de Pearson, con esta técnica utilizando un valor de correlación mayor a 0.6, se encontró que, de las **65 variables iniciales**, 20 tuvieron una correlación moderada con otras 20, de las cuales se escogieron las 20 que menor valor de correlación tuvieron con la variable respuesta. Estas variables fueron: dimensión urbana, dimensión de la calidad de vida, ingresos totales per cápita, número de personas secuestradas, tasa de homicidios (por cada 100,000 habitantes), porcentaje de zona antrópica, porcentaje de bosques, porcentaje de cuerpos de agua, porcentaje de zonas húmedas, máxima precipitación, máxima temperatura, promedio de temperatura, mínimos metros sobre el nivel de mar, metros sobre el nivel de mar, precipitación, rango de metros sobre el nivel de mar, porcentaje de zonas secas, promedio de la temperatura, porcentaje de zonas acuáticas y porcentaje de zonas susceptibles de inundación. En los modelos implementados se tuvo en cuenta tanto el conjunto de datos completo con todas las variables como el conjunto sin las 20 variables correlacionadas.

A parte de lo anteriormente mencionado, se eliminaron de los dos *datasets* (mensual y semanal) de forma definitiva variables que no eran de interés o no podían aportar información relevante para los modelos, que tenían muchos datos faltantes o inconsistencias en los datos. Entre estas variables eliminadas está: código DANE-periodo, región, subregión, promedio de temperatura, promedio de precipitación y el índice de

vegetación NDVI. Es así como sin tener en cuenta la columna del nombre del departamento y sin contar la columna del nombre del municipio, se obtuvo un conjunto con **46 variables contextuales, 3 variables de observación** y **la variable respuesta.**

Además, se implementó la técnica de análisis de componentes principales (PCA) como otra opción de reducción de dimensionalidad, obteniendo un nuevo conjunto de datos que se guardó para ser usado en los futuros modelos implementados. También, se implementaron algoritmos no supervisados de agrupación (Anexo 3) para diferentes particiones del *dataset* (sólo variable respuesta, únicamente variables contextuales, todas las variables) los cuales no mostraron un patrón evidente de agrupación que fuera útil para el objetivo de esta investigación, por lo que se implementaron únicamente algoritmos de aprendizaje supervisado.

6.4. Definición del modelo predictivo

Se implementaron varios modelos de clasificación, posteriormente algunos de regresión y, finalmente, de regresión para series de tiempo.

En este capítulo se mostrará la implementación de cada una de las técnicas de aprendizaje de máquina escogidas, árboles de decisión, *naive bayes, máquinas de soporte vectorial* y redes neuronales. Se evaluó el desempeño de cada una de las técnicas anteriormente descritas, se seleccionaron las técnicas que presentaron mejor desempeño de acuerdo con las métricas con las que se evalúo. Finalmente, se realizaron las pruebas necesarias para evaluar el comportamiento de los modelos predictivos.

6.4.1. Implementación de los modelos de aprendizaje de máquina

En esta sección se muestran las diferentes implementaciones de técnicas de *machine learning* usadas en esta investigación (árboles de decisión, *naive bayes*, máquinas de soporte vectorial y redes neuronales), tanto para algoritmos de clasificación como para los algoritmos de regresión y de regresión para series de tiempo.

6.4.1.1. Modelos de clasificación

Primero se hizo la creación de las clases de la variable respuesta. Segundo, se dividió el conjunto de datos para entrenamiento, validación y prueba. Posteriormente, se estandarizaron los datos, se balancearon las clases de la variable respuesta, después se entrenó el modelo, se probó con los datos de validación y, finalmente, se analizaron las métricas de desempeño con los datos de prueba.

El primer paso realizado con todos los modelos de clasificación implementados fue la creación de las clases en la variable respuesta, en este caso se hicieron 2 clases, una con los registros que presentaron cero casos de LC (clase 0) y la otra clase los registros que presentaron 1 o más casos (clase 1). Con esta división fue claro que se tenía un conjunto de datos muy desbalanceado, tanto para el *dataset* mensual como para el semanal. En el caso del mensual se tenía más del 86% de registros con cero casos reportados y para el semanal más del 92% con cero casos.

El *dataset* se dividió en un subconjunto de entrenamiento (75%) y otro de prueba (25%), después, se hizo la estandarización z-score. Debido al desbalance en las clases de la variable binaria respuesta, se procedió a implementar diferentes técnicas de balanceo de clases para poder realizar mejor la tarea de clasificación.

Las técnicas para balancear las clases implementadas fueron:

Submuestreo (*undersampling*): esta técnica simplemente consiste en eliminar registros al azar de la clase con mayor cantidad de registros (clase 0), para que quede de mismo tamaño que la clase con menor cantidad de registros (clase 1).

Sobremuestreo (*oversampling*): consiste en igualar la cantidad de registros en ambas clases, para ello se repitieron registros al azar de la clase 1 hasta que se completó la misma cantidad de registros de la clase 0.

Combinación de *undersampling* y *oversampling* (*SMOTETomek*): consiste en un equilibrio entre las dos anteriormente mencionadas, es decir, se realiza tanto un *undersampling* como un *oversampling* combinados.

Es importante tener presente que el balanceo se hizo sólo sobre el conjunto de entrenamiento, el de prueba (*test*) se dejó intacto. A continuación, se muestran varios

resultados de modelos de clasificación con diferentes técnicas de *machine learning* y distintos métodos de balanceo de clases. Se observa que el mejor modelo teniendo en cuenta la métrica *Macro f1-score* es el modelo con la técnica *XGBoost* de clasificación sin balancear las clases y con el *dataset* semanal obteniendo un valor de 0.72 (ver Tabla 4). Con el *dataset* mensual, nuevamente el mejor resultado se obtuvo con el algoritmo *XGBoost* de clasificación, teniendo en cuenta la métrica *Macro f1-score* (ver Tabla 5). Es importante mencionar que ya desde este momento se empieza a ver que el modelo tiene mejor desempeño de clasificación con el *dataset* mensual.

Tabla 4: Diferentes modelos de clasificación con el *dataset* semanal sin balanceo de clases

Técnica de inteligencia artificial	F1-score clase 0	F1-score clase 1	Macro Recall	Macro F1-score
XGBClassifier*	0.97	0.47	0.67	0.72
Naive Bayes	0.88	0.34	0.76	0.61
Árboles de decisión	0.97	0.44	0.66	0.71
Máquinas de soporte vectorial	0.94	0.61	0.74	0.77
Perceptrón multicapa	0.94	0.63	0.77	0.79

*Modelo con mejor desempeño

Tabla 5: Diferentes modelos de clasificación con el *dataset* mensual sin balanceo de clases

Técnica de inteligencia artificial	F1-score clase 0	F1-score clase 1	Macro Recall	Macro F1-score
XGBClassifier*	0.94	0.65	0.77	0.80
Naive Bayes	0.88	0.52	0.75	0.70
Árboles de decisión	0.93	0.61	0.75	0.77
Máquinas de soporte vectorial	0.89	0.54	0.75	0.71
Perceptrón multicapa	0.94	0.63	0.77	0.79

*Modelo con mejor desempeño

En la Tabla 6 se pueden observar los diferentes modelos de *machine learning* y balance de clases que se implementaron con el *dataset* semanal. La métrica de desempeño que más se ajustó al objetivo en este trabajo fue el *Macro f1-score*, la cual consiste en un promedio entre el *f1-score* de cada clase. Con base en dicho criterio de desempeño, el mejor resultado para el *dataset* semanal se obtuvo con el algoritmo *XGBoost* y la técnica de balanceo *oversampling,* presentando un *Macro f1-score* de 0.74. El desempeño de este modelo se vio disminuido por la precisión y el *recall* de la clase 1 (al menos un caso reportado).

Tabla 6: Diferentes modelos de clasificación con el *dataset* semanal

Técnica de inteligencia artificial	Método de balanceo	F1-score clase 0	F1-score clase 1	Macro Recall	Macro F1-score
XGBClassifier	*undersampling*	0.90	0.42	0.84	0.66
Naive Bayes	*undersampling*	0.85	0.30	0.76	0.57
Árboles de decisión	*undersampling*	0.86	0.35	0.82	0.60
Máquinas de soporte vectorial	*undersampling*	0.90	0.40	0.83	0.65
Perceptrón multicapa	*undersampling*	0.91	0.43	0.84	0.67
XGBClassifier*	*oversampling*	0.95	0.53	0.82	0.74
Naive Bayes	*oversampling*	0.85	0.30	0.76	0.57
Árboles de decisión	*oversampling*	0.88	0.37	0.82	0.62
Máquinas de soporte vectorial	*oversampling*	0.88	0.37	0.80	0.62
Perceptrón multicapa	*oversampling*	0.89	0.41	0.85	0.65
XGBClassifier	*smoteTomek*	0.93	0.46	0.84	0.69
Naive Bayes	*smoteTomek*	0.85	0.30	0.76	0.57
Árboles de decisión	*smoteTomek*	0.88	0.38	0.82	0.63
Máquinas de soporte vectorial	*smoteTomek*	0.91	0.43	0.84	0.67
Perceptrón multicapa	*smoteTomek*	0.91	0.43	0.85	0.67

*Modelos con mejor desempeño

La Tabla 7 contiene los resultados del mejor modelo de clasificación obtenido, con la técnica de inteligencia artificial XGBoost de clasificación, el *dataset* mensual y con el método *smoteTomek* para el balanceo de las clases.

Tabla 7: Diferentes modelos de clasificación con el *dataset* mensual

Técnica de inteligencia artificial	Método de balanceo	F1-score clase 0	F1-score clase 1	Macro Recall	Macro F1-score
XGBClassifier	*undersampling*	0.89	0.62	0.84	0.76
Naive Bayes	*undersampling*	0.86	0.50	0.75	0.68
Árboles de decisión	*undersampling*	0.88	0.58	0.80	0.73
Máquinas de soporte vectorial	*undersampling*	0.89	0.61	0.83	0.75
Perceptrón multicapa	*undersampling*	0.90	0.63	0.84	0.76
XGBClassifier	*oversampling*	0.90	0.64	0.85	0.77
Naive Bayes	*oversampling*	0.85	0.30	0.76	0.57
Árboles de decisión	*oversampling*	0.89	0.58	0.80	0.74
Máquinas de soporte vectorial	*oversampling*	0.91	0.43	0.84	0.67
Perceptrón multicapa	*oversampling*	0.90	0.63	0.84	0.76
XGBClassifier*	*smoteTomek*	0.93	0.70	0.82	0.82
Naive Bayes	*smoteTomek*	0.85	0.50	0.75	0.67
Árboles de decisión	*smoteTomek*	0.89	0.58	0.80	0.73
Máquinas de soporte vectorial	*smoteTomek*	0.91	0.43	0.84	0.67
Perceptrón multicapa	*smoteTomek*	0.90	0.62	0.84	0.76

*Modelo con mejor desempeño

La Tabla 8 muestra el detalle de los parámetros seleccionados para el mejor modelo con el algoritmo *XGBoost*, el *dataset* mensual y balanceado con la técnica *smoteTomek.*

Además, se estimaron las métricas del área bajo la curva ROC (AUC) y el área bajo la curva precisión-recall (PR), para el modelo de clasificación con el mejor desempeño (Tabla 9). Ambas métricas confirman el buen desempeño del modelo seleccionado.

Debe resaltarse que normalmente para conjuntos de datos con clases balanceadas se suele utilizar la métrica de *AUC ROC*, pero para este conjunto en cuestión, se incluyó la métrica de *AUC PR*, que es útil para conjuntos de datos desbalanceados. Las figuras 27 y 28 presentan las gráficas para la curva ROC y la curva PR respectivamente.

Tabla 8: Parámetros del modelo de clasificación seleccionado

Parámetro	Valor
Tasa de aprendizaje (Learning rate)	0.1
Profundidad máxima (max depth)	30
Porcentaje de características por árbol (colsample bytree)	0.4
Gamma	0.4
Número de árboles (n_stimators)	1000

Tabla 9: Métricas adicionales del modelo de clasificación seleccionado

Métrica	Valor
ROC_AUC	0.924
PR_AUC	0.725

Figura 27: Curva ROC AUC

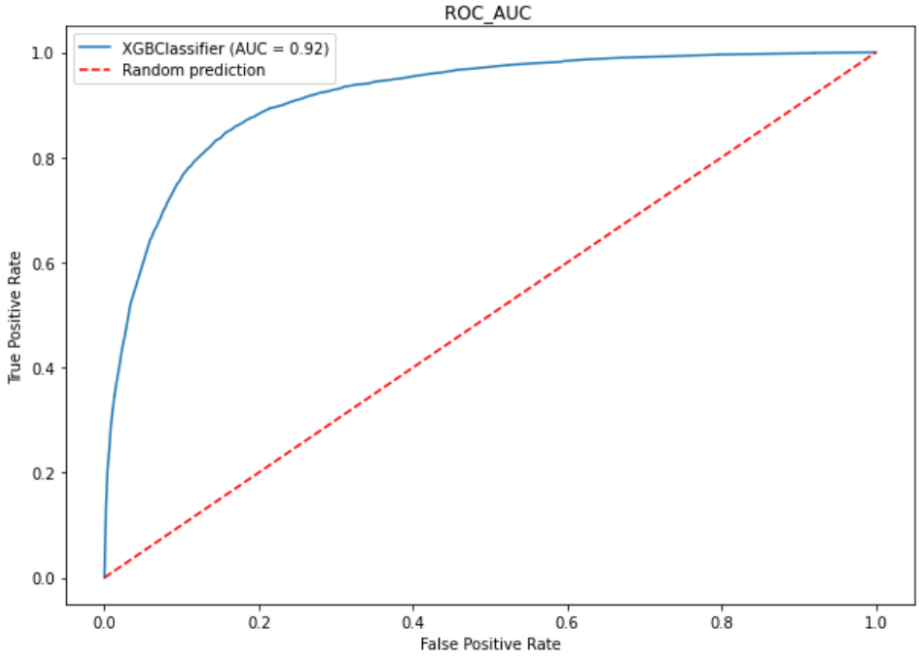
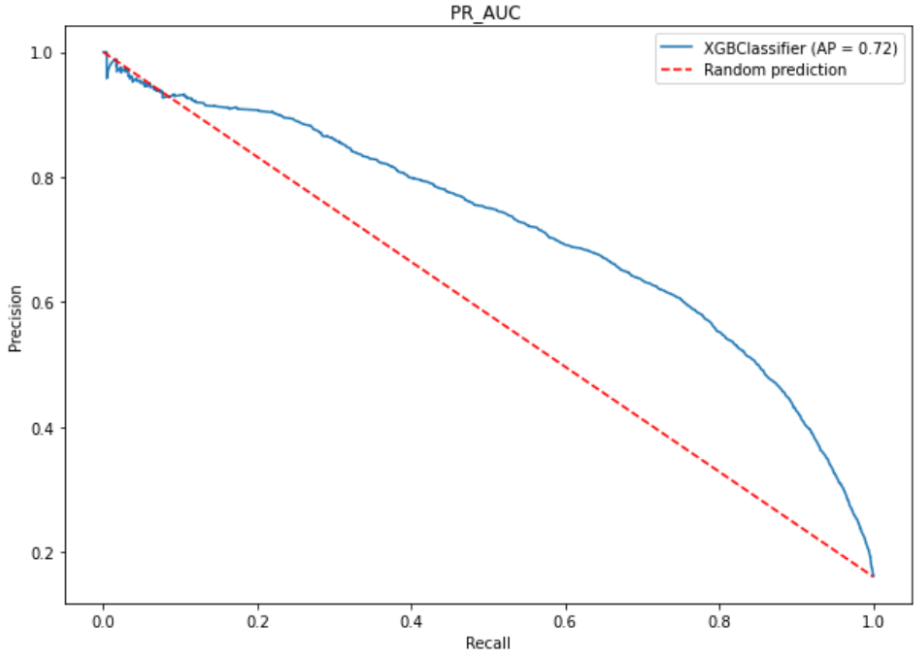


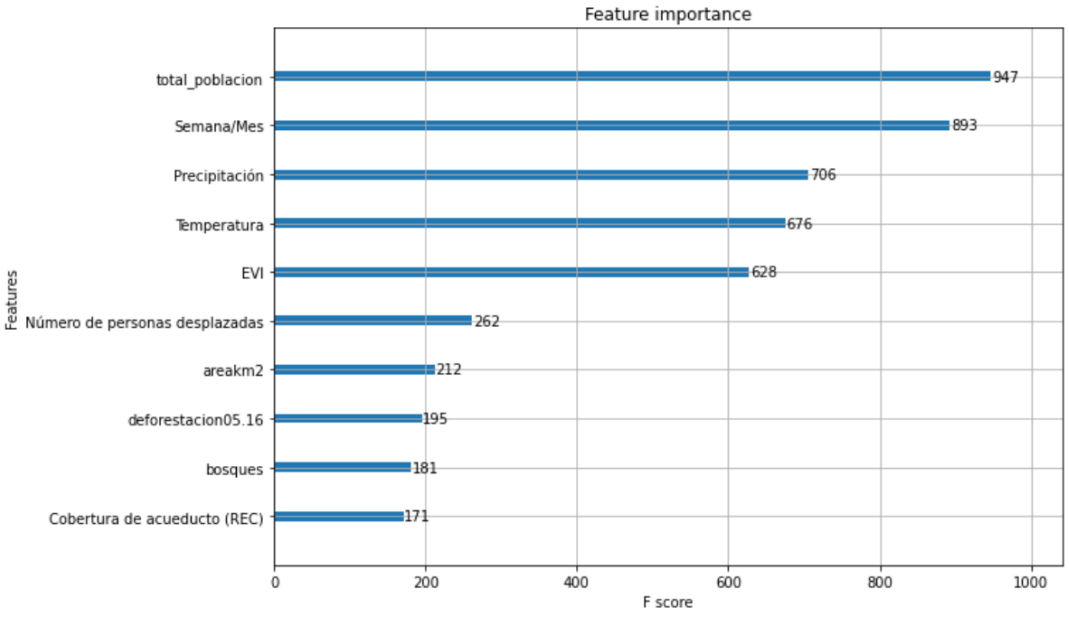
Figura 28: Curva PR AUC



Al comparar los resultados de la métrica *ROC_AUC* obtenidos en este estudio con los obtenidos por King (2004) usando un modelo de regresión logística para todo el país, se observa que el presente estudio obtuvo mejores resultados (0.929 vs. 0.72). Esto pudo obedecer tanto a la cantidad y tipo de variables utilizadas, como al mejor desempeño en la tarea de clasificación por parte de la técnica de *machine learning*.

Por otra parte, en esta investigación utilizando el algoritmo *XGBoost* para clasificación se encontró que las 10 variables más importante para la predicción de casos de LC fueron: el total de la población del municipio, el mes, la precipitación, la temperatura, el EVI, el número de personas desplazadas, el área del municipio en kilómetros cuadrados, el porcentaje de deforestación, el porcentaje de bosques y la cobertura de acueducto (Figura 29). La variable mes, al ser una variable seleccionada como importante para el modelo de clasificación, empezó a dar luces de que un modelo para este conjunto de datos se aprovecharía mejor implementando series de tiempo.

Figura 29: Las 10 variables seleccionadas como importantes para la clasificación binaria de casos de LC



Hasta donde se investigó el estado del arte, no existen trabajos previos donde se implementen técnicas de inteligencia artificial para estimar la ocurrencia de casos o incidencia de LC, sin embargo, se hace una comparación con los resultados obtenidos con autores que utilizaron diversos métodos estadísticos. Por ejemplo, King (2004) utilizando

un modelo con regresión logística para predecir la variación geográfica de LC, encontró que las variables más importantes fueron: la cobertura del suelo (25 variables categóricas) y la elevación sobre el nivel de mar; una de las variables que guarda relación con las encontradas mediante el algoritmo *XGBoost* es la variable de bosques, incluida como una de las coberturas del suelo incluidas por King (2004). También Valderrama y colaboradores (2010) utilizando un modelo condicional autorregresivo de Poisson para la correlación espacial, identificaron que la temperatura y la cobertura de bosques estaban asociadas de forma positiva con la incidencia de LC, ambas variables están incluidas dentro de las 10 variables más importantes encontradas en esta investigación mediante el algoritmo *XGBoost* para clasificación (Figura 29).

6.4.1.2. Modelos de regresión

Los modelos de regresión implementados muestran los diferentes resultados obtenidos con los *datasets* semanal y mensual, teniendo presente que se realizaron modelos incluyendo los casos donde el valor es cero y modelos únicamente con el conjunto de datos con valores de casos mayores que cero (Tabla 10), esto para encontrar diferentes resultados buscando una mejora de los modelos.

Para la explicación de la Tabla 10 es importante tener presente que el promedio de la variable casos, en el *dataset* donde fueron excluidos los ceros fue de 4.298 casos y en el *dataset* con registros de la variable respuesta que incluían ceros fue de 0.688. El algoritmo con mejor desempeño de acuerdo con la métrica de *Mean absolute error,* fue el perceptrón multicapa, tanto para la situación en la que se tenían en cuenta los ceros, como haciendo caso omiso de ellos (Tabla 10).

Tabla 10: Diferentes modelos de regresión con el *dataset* semanal

Técnica de inteligencia artificial	Con ceros	Mean absolute error	Mean squared error	R2 score	Explained variance
XGBRegressor	No	1.787	14.560	0.156	0.178
Perceptrón multicapa*	No	1.438	13.861	0.197	0.199
Máquinas de soporte vectorial	No	1.248	15.531	0.0.998	0.115
XGBRegressor	Sí	0.410	1.013	0.191	0.224
Perceptrón multicapa*	Sí	0.188	0.914	0.270	0.271
Máquinas de soporte vectorial	Sí	0.946	2.317	0.163	0.196

*Modelos con mejor desempeño

Los modelos implementados con el *dataset* mensual obtuvieron un mejor resultado de regresión con el perceptrón multicapa, en el *dataset* que incluía valores de casos igual a cero (Tabla 11). Para un promedio de 0.688 de casos en dicho *dataset,* el modelo cometió un error medio absoluto de 0.731 y obtuvo un *R2 Score* de 0.324.

Tabla 11

Diferentes modelos de regresión con *dataset* mensual

Técnica de inteligencia artificial	Con ceros	Mean absolute error	Mean squared error	R2 score	Explained variance
XGBRegressor	No	4.402	105.472	0.170	0.196
Perceptrón multicapa*	No	1.326	13.921	0.193	0.194
Máquinas de soporte vectorial	No	2.621	106.039	0.165	0.173
XGBRegressor	Sí	1.493	18.365	0.216	0.240
Perceptrón multicapa*	Sí	0.731	15.837	0.324	0.324
Máquinas de soporte vectorial	Sí	1.351	10.178	0.198	0.201

*Modelos con mejor desempeño

6.4.1.3. Modelos de regresión con *dataset* departamental

Se muestran dos resultados de implementación, uno con el *dataset* mensual y otro con el semanal utilizando la técnica de regresión mediante *XGBoost*. El *dataset* semanal presentó una media de casos de 6.184 casos, y el mensual un promedio de 28.899 casos. En ambos *datasets,* el *mean absolute error* fue menor al promedio de casos y el *R2-Score* para el *dataset* mensual presentó un valor de 0.736 (Tabla 12).

De igual forma, la Figura 30 presenta el diagrama de la predicción vs el valor real, lo que permite observar el desempeño del modelo con los datos de prueba. Utilizando *XGBoost* para regresión, se encontró que las 10 variables correlacionadas positivamente con la predicción de casos de LC fueron: el total de la población del departamento, la

precipitación, el EVI, la temperatura, el mes, los cultivos transitorios, el área del departamento en kilómetros cuadrados, la vegetación secundaria, el número de personas desplazadas y la cobertura de acueducto (Figura 31). Nuevamente acá la variable del mes aparece como importante para la predicción de casos de LC, lo cual como se mencionó antes, permite reforzar la importancia de realizar un trabajo con series de tiempo.

Tabla 12: Modelos de regresión con *dataset* departamental con *XGBoost*

Técnica de dataset	Mean absolute error	Mean squared error	R2 score	Explained variance
Semanal	2.824	36.094	0.740	0.740
Mensual*	10.047	377.713	0.807	0.807

*Modelo con mejor desempeño

Figura 30: Real vs predicción modelo de regresión departamental a nivel mensual

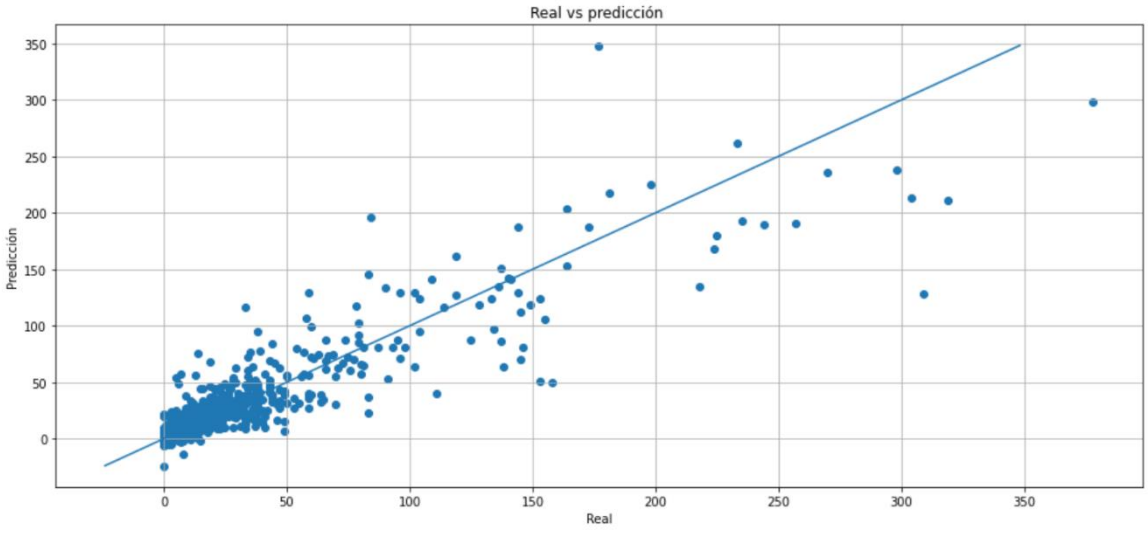
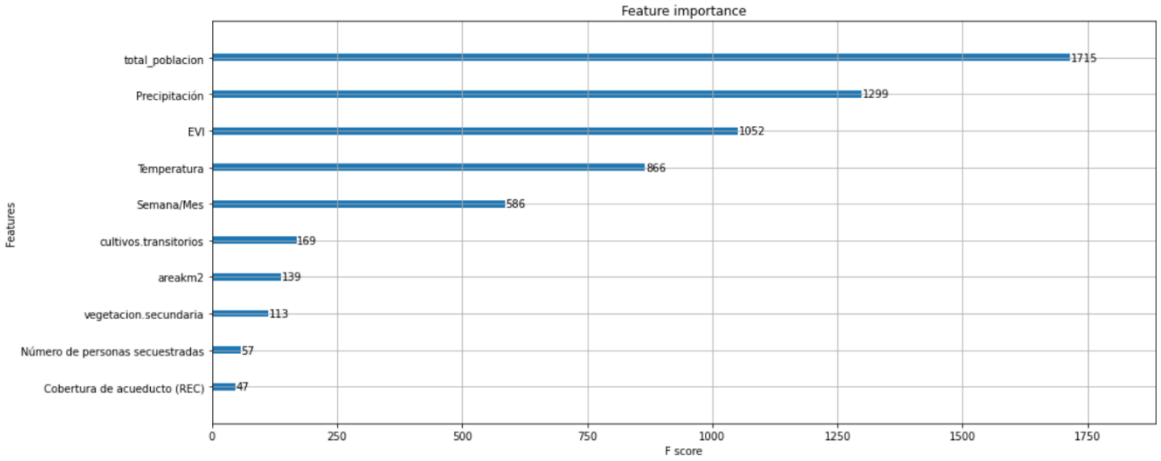


Figura 31: Las 10 variables más asociadas positivamente con los casos de LC en regresión



Gutiérrez y colaboradores (2018) utilizando un modelo de regresión binomial negativa para la incidencia de LC en los departamentos de Norte de Santander y Santander, identificaron que las variables que estaban asociadas de forma positiva con la incidencia fueron: la cobertura de bosques, las zonas agrícolas heterogéneas y los cultivos permanentes; ninguna de las variables reportadas por dichos autores fueron identificadas como significativas en el modelo *XGBoost* implementado para la regresión del *dataset* mensual a nivel departamental en este trabajo. Por otra parte, Pérez y colaboradores (2016) usando un análisis espacio temporal condicional autorregresivo de Poisson de efectos aleatorios para la identificación de los factores de riesgo ambiental para la LC, encontraron que las variables más importantes fueron: los bosques lluviosos, la vegetación secundaria, la temperatura y la precipitación; estos resultados coinciden en las variables de temperatura, precipitación y vegetación secundaria encontradas en esta investigación (Figura 31).

6.4.1.4. Modelos de regresión para series de tiempo en el set departamental

Se utilizó la función de autocorrelación para identificar el valor más alto de correlación de la serie de tiempo de la variable casos de cada departamento. A continuación, se muestra

a manera de ejemplo la función de autocorrelación para el departamento de Antioquia que fue el departamento con mayor cantidad de casos reportados (ver Figura 32 y Figura 33).

La Figura 32 muestra la serie de tiempo de la variable casos del departamento de Antioquia. El mayor pico de casos corresponde a más de 150 casos en el primer trimestre de 2010.

Figura 32: Serie de tiempo de Antioquia

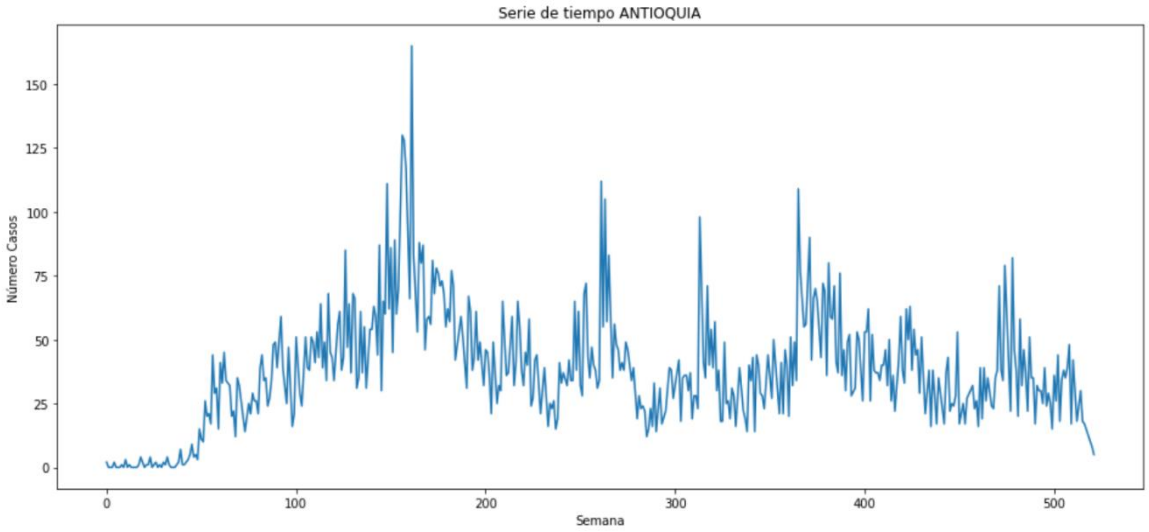
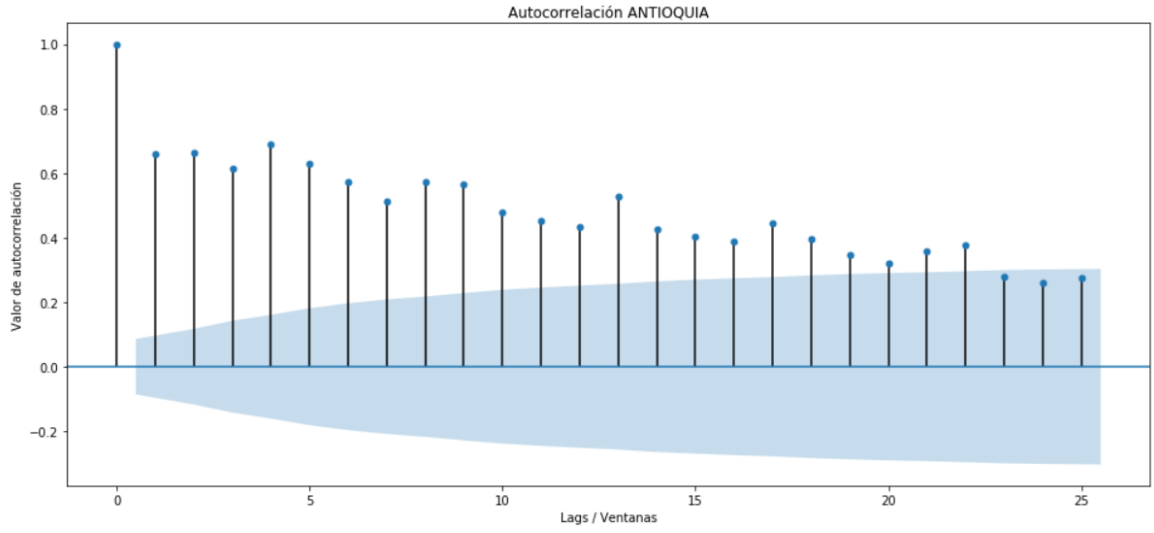


Figura 33: Función de autocorrelación de la variable casos para Antioquia



El análisis de la función de autocorrelación para cada departamento sirvió como guía para la inclusión de los retrasos en los modelos de regresión que predicen el número de casos

a partir de los valores previos de la variable respuesta y de las variables de observación (temperatura, precipitación y EVI).

Los mejores valores en ventanas de tiempo (*lags*) para cada departamento son presentados en el Anexo 4. Todos los resultados obtenidos en esta sección y presentados a continuación, fueron realizados con el *dataset* **departamental a nivel semanal** y teniendo presente que dentro de este conjunto de datos no están presentes los departamentos de Amazonas, Vaupés y Guainía, debido a que tienen corregimientos municipales y fueron excluidos desde el principio de la investigación.

Una vez se identificó la cantidad de retrasos a incorporar por cada departamento, se procedió a realizar los modelos de regresión para series de tiempo por cada departamento. Debido a que el conjunto de datos para cada departamento de forma individual fue de 522 observaciones (semanas analizadas en cada departamento), el desempeño de dichos modelos fue deficiente, en comparación con la compilación de todos los departamentos en un único modelo que contenía 14,616 observaciones, donde el conjunto de prueba correspondió al 20% final de la serie de cada departamento.

Los resultados para cada uno de los modelos departamentales no son presentados en este documento, sin embargo, en el Anexo 5 se muestra a manera de ejemplo los resultados deficientes de la implementación de los modelos *XGBoost* y *MLP*, para el departamento de Antioquia, teniendo como mejores métricas un *Mean absolute error* de 4.4 y un *R2 score* de 0.181 en el modelo con el algoritmo *XGBoost* (ver Anexo 5.6).

Una vez se obtuvieron los resultados a nivel departamental y después de analizar los resultados individuales de cada departamento, se procedió a realizar un modelo global, como se mencionó anteriormente, es decir, un modelo nacional entrenado con todos los departamentos y evaluado con un 20% de cada departamento para así tener más información de la enfermedad y sus variables asociadas.

El conjunto de datos a nivel departamental fue dividido en entrenamiento y prueba. Para ello se tomaron los datos de cada departamento, el 80% inicial de la serie de tiempo se utilizó para entrenamiento y el 20% final para prueba.

Luego se procedió a la normalización con *Z-Score*. Una vez normalizado el conjunto de datos se continuó con la preparación para un modelo con series de tiempo. Para ello se definió la cantidad de *lags* (corrimientos o ventanas de tiempo) que fueron tenidas en cuenta para el proceso de entrenamiento. La escogencia de dicha cantidad de *lags* se hizo basado en dos métodos, el primero de ellos fue promediando las ventanas de tiempo que se obtuvieron por cada departamento y la segunda forma fue buscando el mejor valor de *lags* para los modelos, probando valores por encima y por debajo del promedio y buscando el mejor modelo para el mejor número de *lags*. Como resultado de esta exploración, se seleccionaron *6 lags,* para ser incorporados en los modelos de series de tiempo a nivel nacional.

Se realizaron tres modelos nacionales diferentes, uno con la variable únicamente de casos de LC, otro modelo con las variables de observación (incluida la variable de casos de LC) y finalmente un tercer modelo con todas las variables, las de observación, la variable respuesta y las 46 variables de contexto. Para cada uno de estos modelos se estimaron predicciones a 1 semana a futuro.

Modelo de sólo casos: *dataset* conformado únicamente por la variable casos, utilizando los 6 reportes de 6 semanas anteriores de casos de LC para predecir una semana a futuro.

Modelo con las variables de observación: este conjunto de datos estaba conformado por las 3 variables de observación (temperatura, precipitación y EVI) y la variable respuesta (casos de LC). Para la construcción de las matrices de *lags* se tuvo en cuenta 6 corrimientos por cada variable, es decir, en total 24 variables para predecir el número de casos de LC de la semana siguiente.

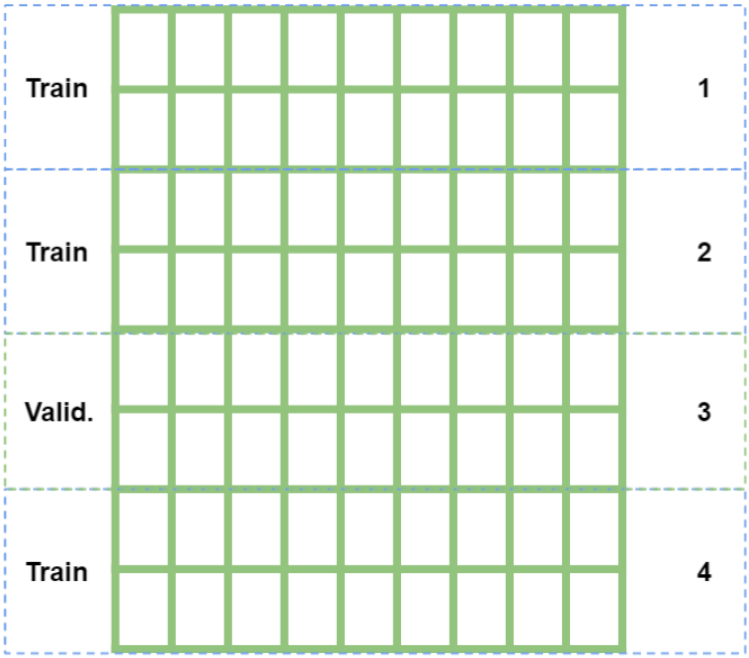
Modelo con las variables de observación y de contexto: para este modelo se utilizaron las 24 variables de observación descritas previamente y las 46 variables de contexto, que, debido al no ser de carácter temporal, no se requirió definir ventanas de tiempo.

Todo el afinamiento de los hiperparámetros realizado en este proyecto, se hizo con el algoritmo *RandomizedSerachCV* de la librería de *sklearn*.

Con el fin de obtener mejores resultados, mejor afinamiento de hiperparámetros y unas métricas más confiables, se implementó validación cruzada dividiendo el conjunto de entrenamiento en 4 pliegues (*folds)*. La Figura 34 muestra como se hizo el proceso de

validación cruzada usando 4 pliegues. En dicho proceso, el conjunto de entrenamiento se divide en 4 partes, se toman 3 para entrenamiento y 1 para validación, luego se cambia ese conjunto de validación por otro, y se toman los 3 restantes para entrenamiento y así sucesivamente hasta usar cada uno de los 4 conjuntos para validación con los tres restantes para entrenamiento.

Figura 34: Validación cruzada para el afinamiento de los hiperparámetros y entrenamiento del modelo



A continuación, se detallarán 5 aspectos importantes en los resultados del afinamiento de los hiperparámetros; primero, una tabla con el conjunto de hiperparámetros para el afinamiento junto con los parámetros seleccionados. Segundo, una gráfica de la función de pérdida del conjunto de validación que permite descartar el *underfitting* y el *overfitting*. Tercero, una figura que compara las series de tiempo observadas vs. las predichas, cuarto, una figura de la dispersión de los datos de la predicción vs. los valores reales y quinto una tabla con los resultados de las métricas de desempeño del algoritmo. Para el caso de los modelos con redes neuronales, se muestran sus resultados en la sección de anexos (ver Anexo 6), ya que el mejor desempeño en los modelos fue alcanzado por el algoritmo *XGBoost.*

***Tuning* modelo de sólo casos con *XGBoost*:**

Tabla 13: Conjunto de hiperparámetros *afinados* a nivel nacional con la variable de casos

Hiperparámetro	Valores evaluados	Valor seleccionado
Tasa de aprendizaje (Learning rate)	0.001, 0.01, 0.05, 0.10, 0.20, 0.50, 0.90	0.05
Profundidad máxima (max depth)	3, 4, 5, 6, 8, 9,10, 12, 15	9
Porcentaje de características por árbol (colsample bytree)	0.3, 0.4, 0.5, 0.7, 0.9	0.9
Gamma	0.0, 0.005, 0.009, 0.1, 0.2, 0.3, 0.4	0.09
Número de árboles (n_stimators)	100, 1000, 10000	1000

Figura 35: Función de perdida *mean absolute error* para el modelo nacional de sólo casos: eje *Y* representa el valor del mae y el eje *X* representa la cantidad de árboles creados.

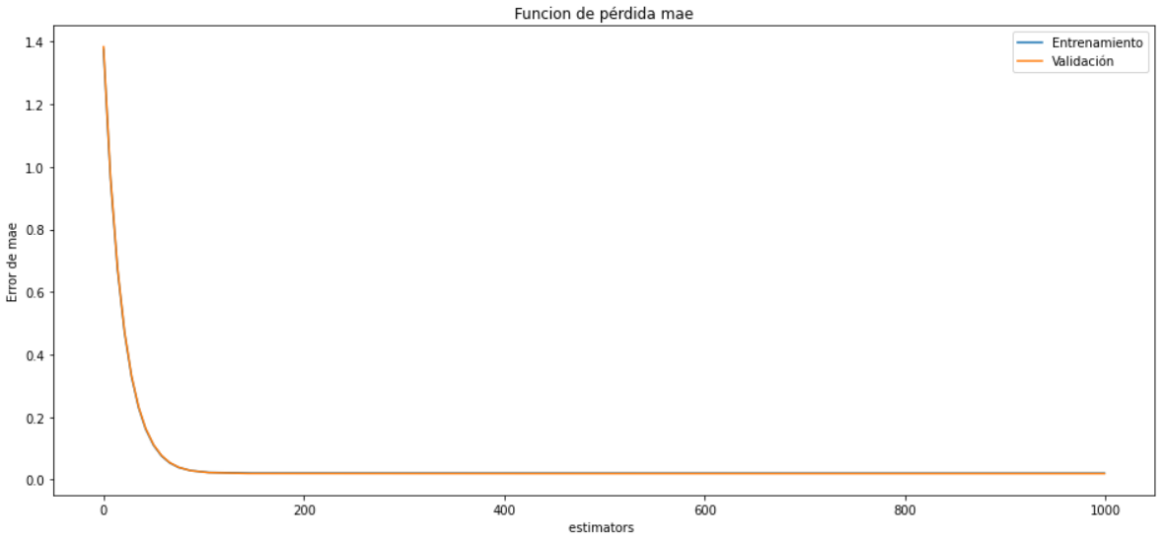


Figura 36: Real vs predicción del modelo nacional con sólo casos: el eje *Y* es el valor de casos de LC. El eje *X* hace referencia a 2,744 semanas de prueba, las cuales son las 98 últimas semanas de cada departamento.

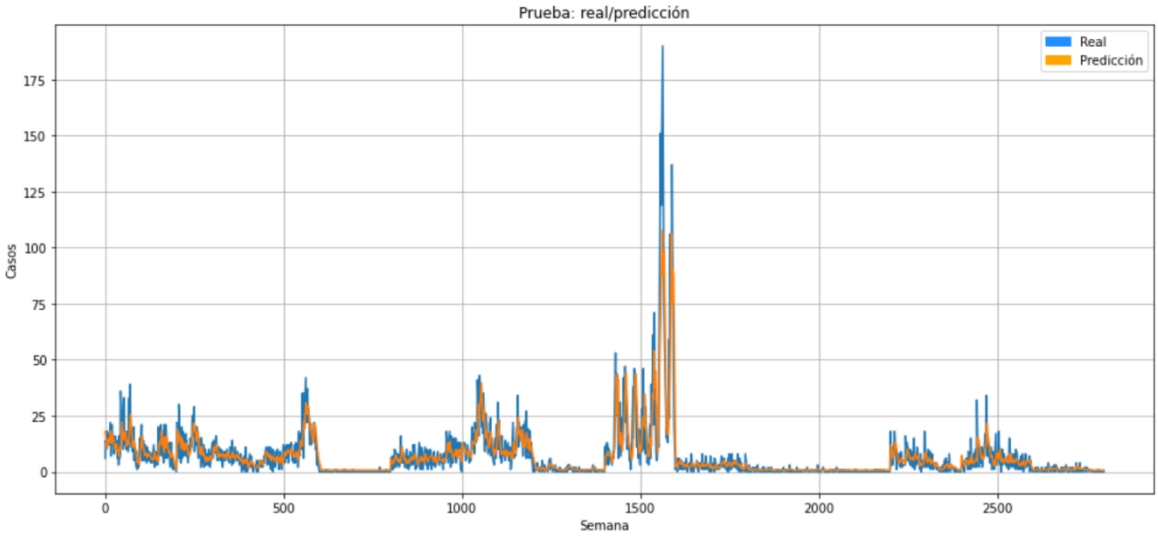


Figura 37: Comparación de valores observados vs. predichos del modelo nacional sólo casos: el eje *Y* es el valor predicho y el eje *X* el valor real

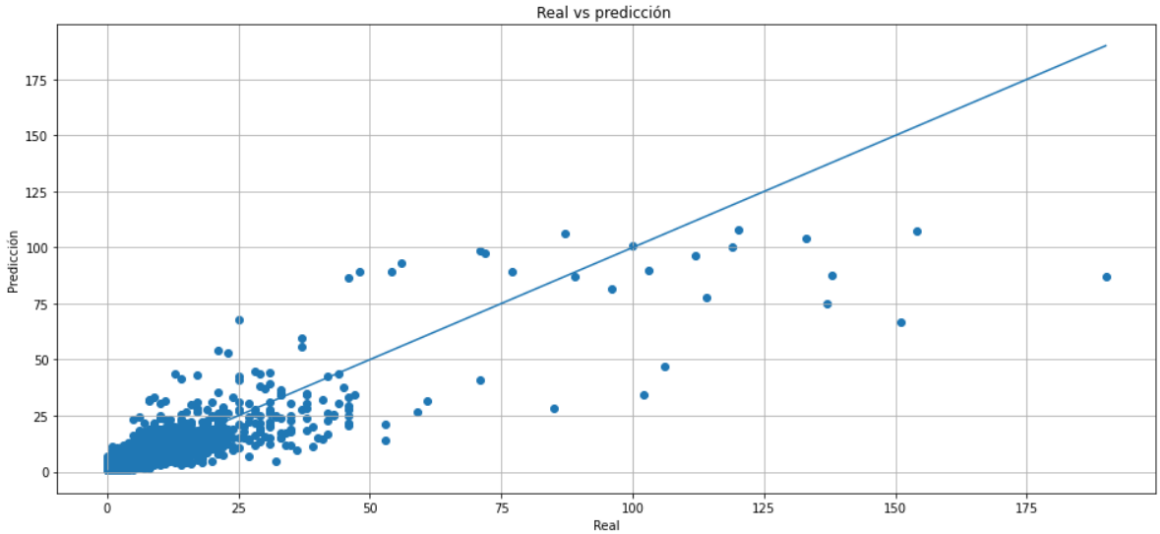


Tabla 14: Comparación de métricas resultantes a nivel nacional, de sólo casos

Métrica	XGBoost	MLP*	Modelo de media de casos	Modelo de persistencia
Mean absolute error	2.668	2.631	8.213	6.113
Mean squared error	31.400	36.661	112.734	83.872
R2 score	0.780	0.774	0.619	0.641
Explained variance score	0.781	0.778	0.627	0.646

*Para ver los resultados en detalle y el afinamiento del MLP referirse a los anexos 6.1,6.2,6.3,6.4 y 6.5.

***Tuning* del modelo XGBoost con variables de observación:**

Tabla 15: Conjunto de hiperparámetros *afinados* a nivel nacional con las variables de observación y la variable respuesta

Hiperparámetro	Valores evaluados	Valor seleccionado
Tasa de aprendizaje (Learning rate)	0.001, 0.01, 0.05, 0.10, 0.20, 0.50, 0.90	0.03
Profundidad máxima (max depth)	3, 4, 5, 6, 8, 9,10, 12, 15	8
Porcentaje de características por árbol (colsample bytree)	0.3, 0.4, 0.5, 0.7, 0.9	0.9
Gamma	0.0, 0.005, 0.009, 0.1, 0.2, 0.3, 0.4	0.09
Número de árboles (n_stimators)	100, 1000, 10000	1000

Figura 38: Función de perdida con *mean absolute error* para el modelo nacional con las variables de observación y la variable respuesta: eje *Y* representa el valor del mae y el eje *X* representa la cantidad de árboles creados.

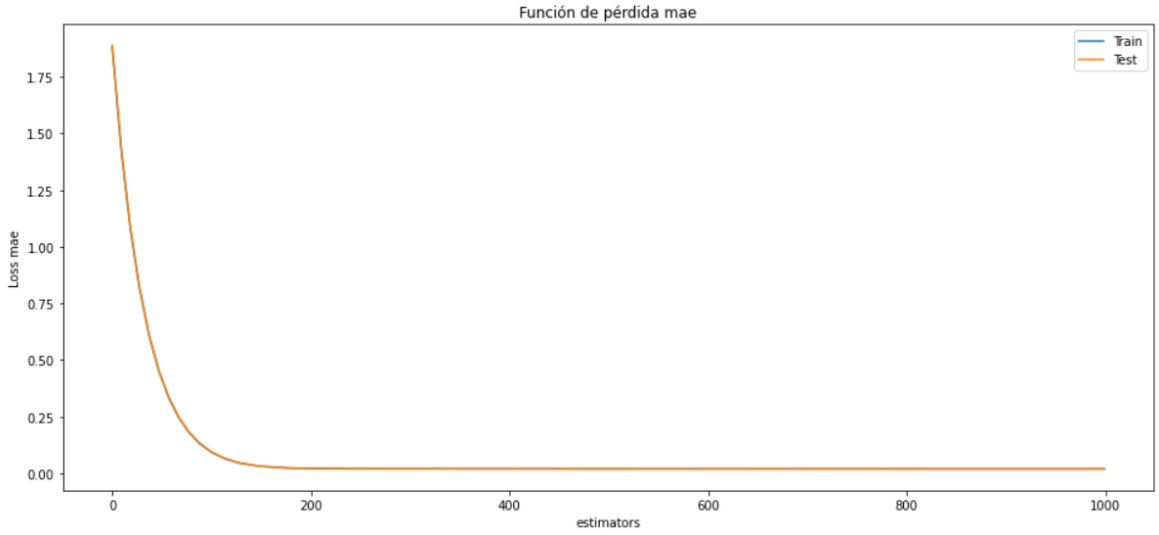


Figura 39: Valor real vs. predicción del modelo nacional con las variables de observación y la variable respuesta: el eje *Y* es el valor de casos de LC. El eje *X* hace referencia a 2,744 semanas de prueba, las cuales son las 98 últimas semanas de cada departamento.

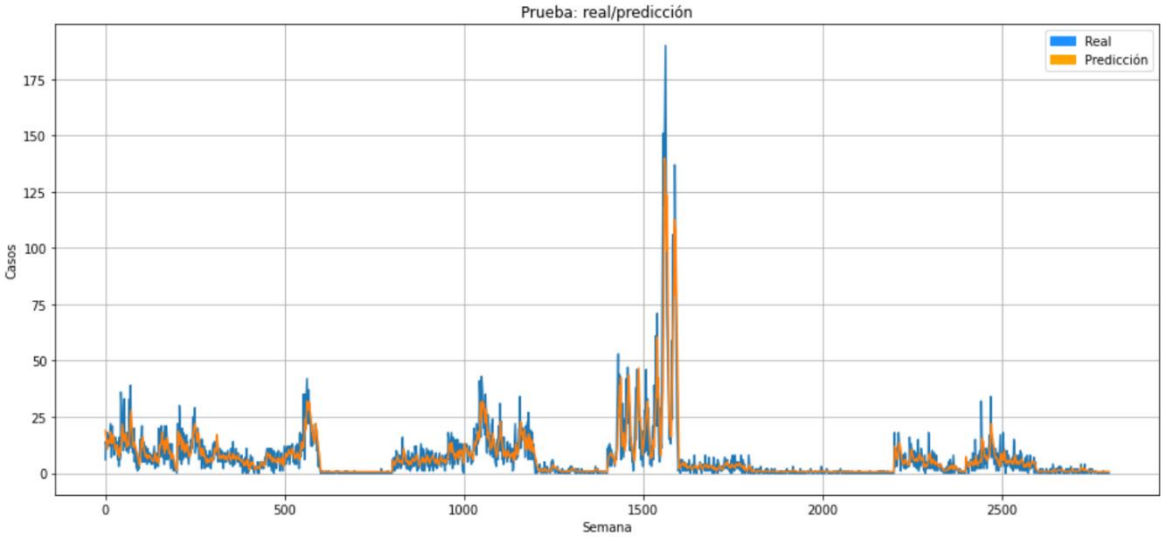


Figura 40: Comparación de valores observados vs predichos del modelo nacional con las variables de observación y la variable respuesta: el eje *Y* es el valor predicho y el eje *X* el valor real

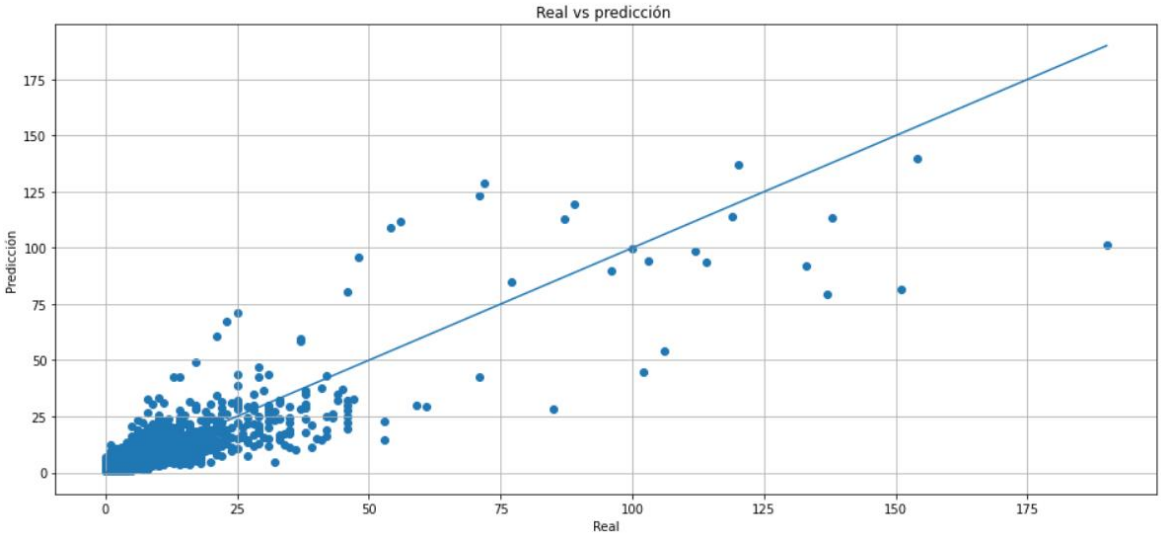


Tabla 16: Comparación de métricas resultantes a nivel nacional, con las 4 variables de observación

Métrica	XGBoost	MLP*	Modelo de media de casos	Modelo de persistencia
Mean absolute error	2.551	2.678	8.745	6.328
Mean squared error	30.408	32.825	113.142	84.017
R2 score	0.792	0.770	0.601	0.624
Explained variance score	0.797	0.771	0.619	0.631

*****Para ver los resultados en detalle y el afinamiento del MLP referirse a los anexos 6.6,6.7,6.8,6.9 y 6.10.

***Tuning* modelo XGBoost con las variables de observación y de contexto:**

Tabla 17: Conjunto de hiperparámetros *afinados* a nivel nacional con todas las variables

Hiperparámetro	Valores evaluados	Valor seleccionado
Tasa de aprendizaje (Learning rate)	0.001, 0.01, 0.05, 0.10, 0.20, 0.50, 0.90	0.05
Profundidad máxima (max depth)	3, 4, 5, 6, 8, 9,10, 12, 15	9
Porcentaje de características por árbol (colsample bytree)	0.3, 0.4, 0.5, 0.7, 0.9	0.9
Gamma	0.0, 0.005, 0.009, 0.1, 0.2, 0.3, 0.4	0.09
Número de árboles (n_stimators)	100, 1000, 10000	1000

Figura 41: Función de perdida *mean absolute error* para el modelo nacional con las variables de observación y de contexto: eje *Y* representa el valor del mae y el eje *X* representa la cantidad de árboles creados.

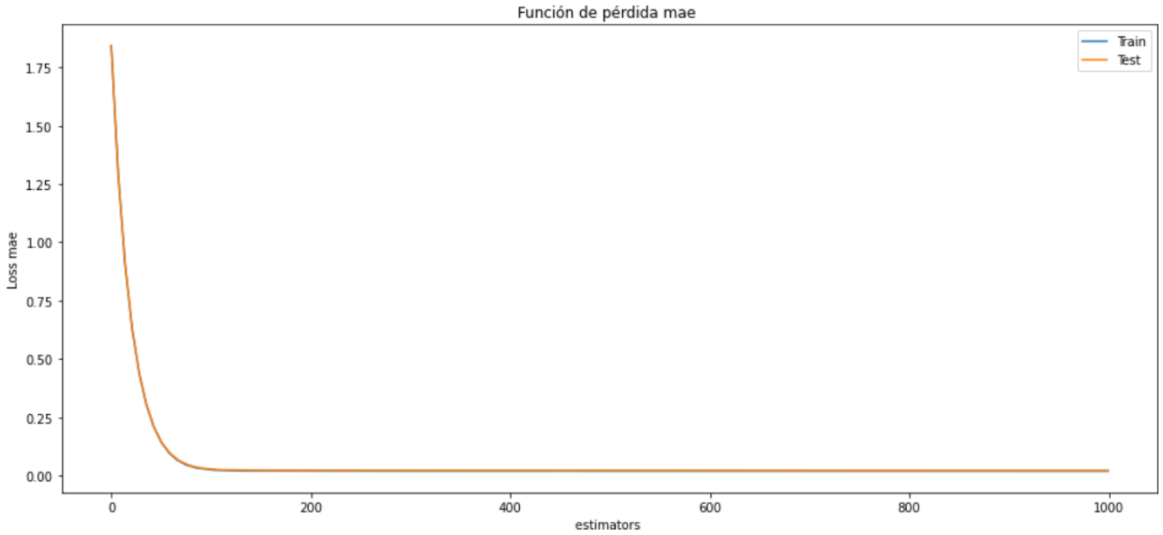


Figura 42: Real vs predicción del modelo nacional con las variables de observación y de contexto: el eje *Y* es el valor de casos de LC. El eje *X* hace referencia a 2,744 semanas de prueba, las cuales son las 98 últimas semanas de cada departamento.

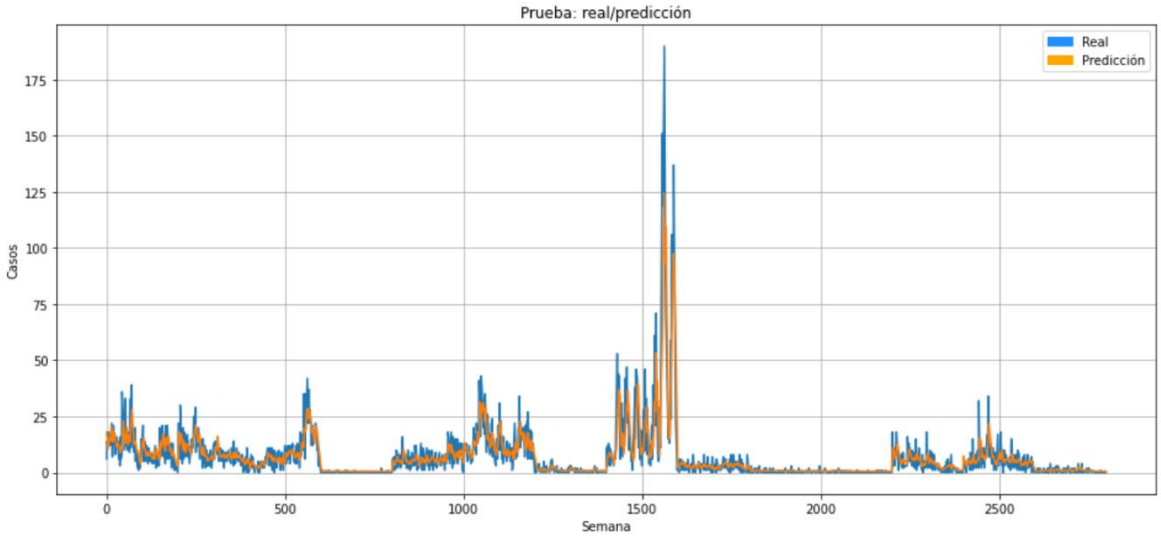


Figura 43: Comparación de valores observados vs. predichos del modelo nacional con las variables de observación y de contexto: el eje *Y* es el valor predicho y el eje *X* el valor real

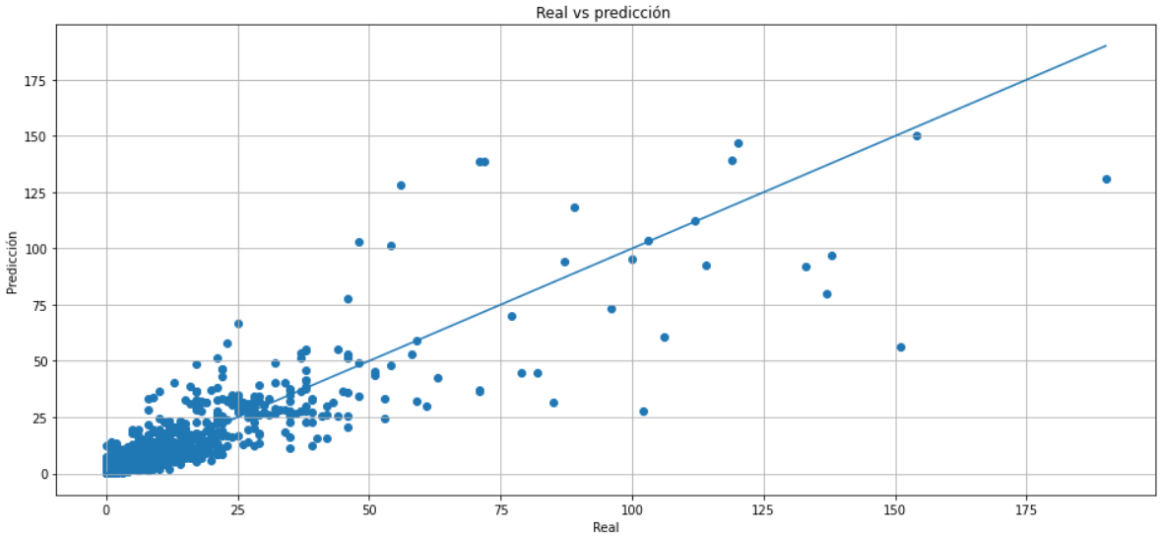


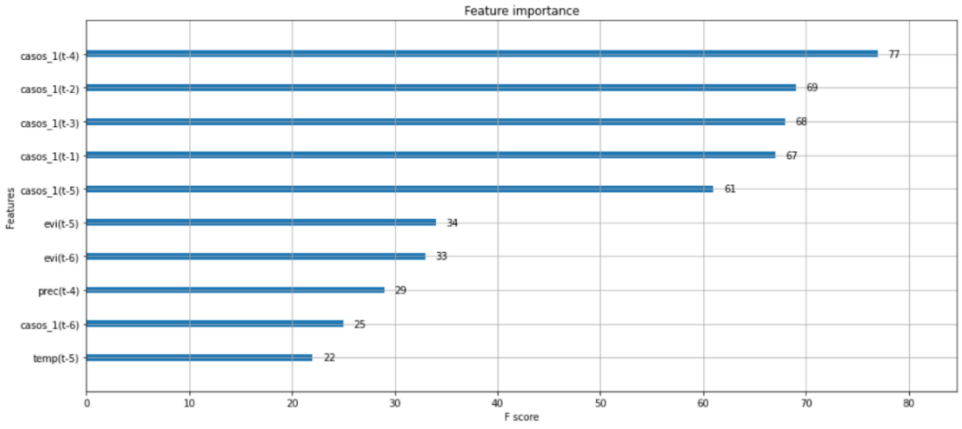
Tabla 18: Comparación de métricas resultantes a nivel nacional, con las variables de observación y de contexto

Métrica	XGBoost	MLP*	Modelo de media de casos	Modelo de persistencia
Mean absolute error	2.547	2.631	9.725	6.871
Mean squared error	30.206	36.661	128.196	84.017
R2 score	0.814	0.744	0.580	0.598
Explained variance score	0.817	0.748	0.591	0.602

*Para ver los resultados en detalle y el afinamiento del MLP referirse a los anexos 6.11,3.12,6.13,6.14 y 6.15.

Utilizando el algoritmo *XGBoost* para regresión con series de tiempo, se encontró que las 10 variables más importantes para la predicción de casos de LC fueron: casos de LC con 4 semanas de retraso, casos de LC con 2 semanas de retraso, casos de LC con 3 semanas de retraso, casos de LC con 1 semana de retraso, casos de LC con 5 semanas de retraso, EVI con 5 semanas de retraso, EVI con 6 semanas de retraso, precipitación con 4 semanas de retraso, casos de LC con 6 semanas de retraso y temperatura con 5 semas de retraso (Figura 44).

Figura 44: Las 10 variables más asociadas positivamente con los casos de LC en regresión para series de tiempo



Chaves y Pascual (2006) utilizando un estudio de series de tiempo (máxima entropía de la densidad espectral, ondulas cruzadas y filtro de *kalman*) de periodicidad mensual en relación con las variables climáticas, encontraron que la temperatura y el MEI (índice multivariado de El Niño) pueden predecir la dinámica de la incidencia de casos de LC hasta con un año de anticipación, lo cual consistente con los resultados de la presente investigación, particularmente en lo relacionado con la temperatura con 5 semanas de retraso, que fue identificada como variable significativa por dichos autores. También, Chaves y colaboradores (2014) usando análisis de series de tiempo con la función de autocorrelación, autocorrelación parcial, correlación cruzada y modelos mixtos generalizados de Poisson, encontraron que las variables que se encuentran correlacionadas positivamente con la incidencia de LC en escalas interanuales son el ENSO (El Niño Oscilación del sur), la lluvia y la temperatura; encontrando relación con el presente estudio en la precipitación con 4 semanas de retraso y la temperatura con 5 semanas de retraso (Figura 44).

6.4.2. Selección de técnicas con mejor desempeño

6.4.2.1. Modelo de clasificación

El mejor modelo de clasificación para el conjunto de datos municipal a nivel mensual mediante la técnica de balanceo de clases *smoteTomek* se obtuvo con el algoritmo *XGBoost* (Tabla 19).

Tabla 19: Modelo de clasificación seleccionado por su mejor desempeño.

Técnica de inteligencia artificial	Método de balanceo	F1-score clase 0	F1-score clase 1	Macro Recall	Macro F1-score
XGBClassifier	*smoteTomek*	0.93	0.70	0.82	0.82

6.4.2.2. Modelos de regresión

En la Tabla 20 se muestra el mejor modelo de regresión obtenido (*XGBoost*) sin series de tiempo, con las variables de observación y de contexto, a nivel mensual y con el *dataset* departamental.

Tabla 20: Modelo de regresión seleccionado sin series de tiempo

Resolución temporal	Mean absolute error	Mean squared error	R2 score	Explained variance
Mensual	10.047	377.713	0.807	0.807

Los modelos de series de tiempo con el algoritmo *XGBoost*, el *dataset* departamental a nivel nacional y a escala temporal de semana, mostraron el mejor desempeño de predicción (Tabla 21). Estos modelos fueron los utilizados para la aplicación web desarrollada.

Tabla 21: Modelo de regresión con series de tiempo seleccionado

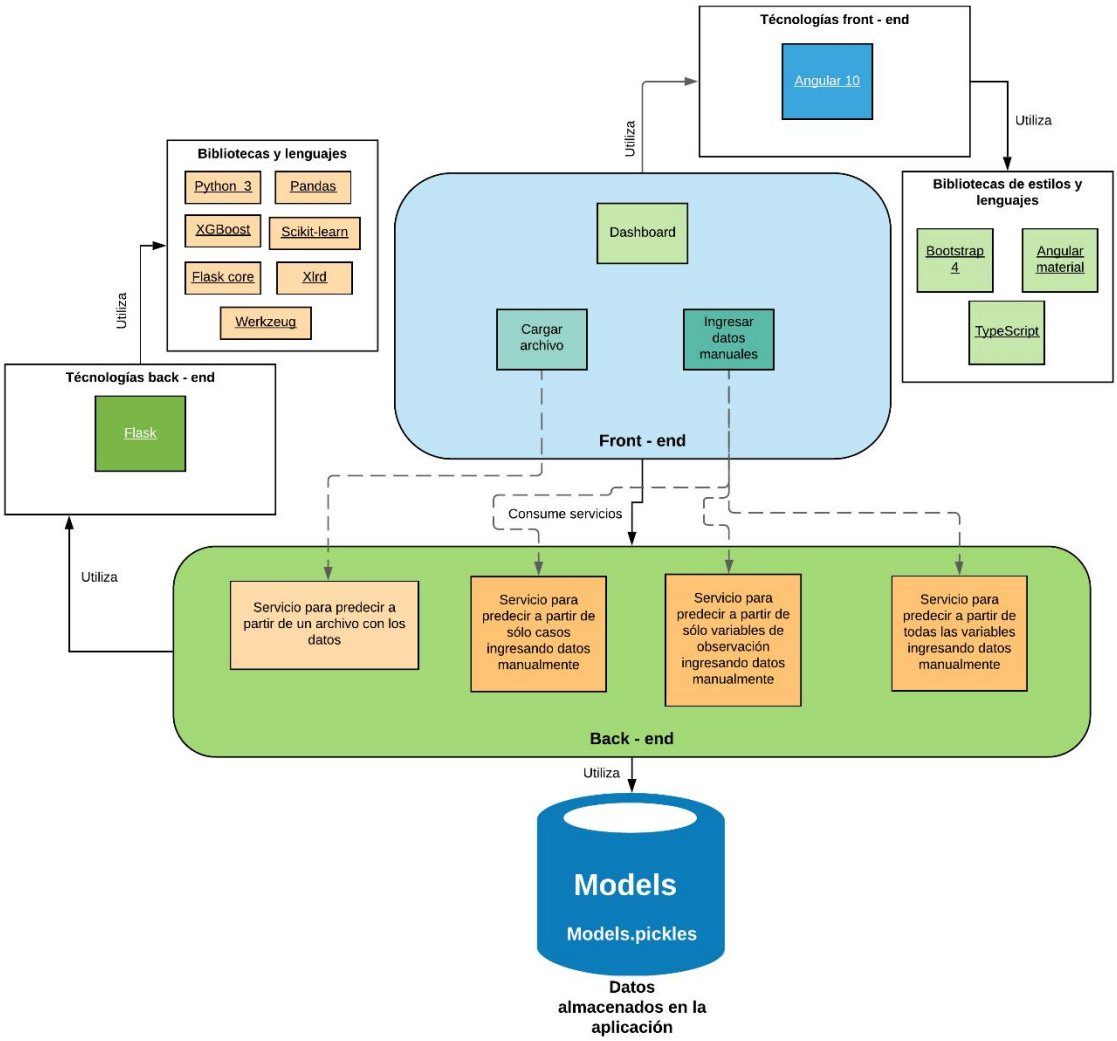
Datos que recibe	Mean absolute error	Mean squared error	R2 score	Explained variance
Sólo casos	2.668	31.400	0.780	0.781
Variables de observación	2.551	30.408	0.792	0.797
Todas las variables	2.547	30.206	0.814	0.817

Capítulo 7: Generación de aplicación web para usar los modelos

Como parte de esta investigación, se implementó una aplicación web que permite utilizar los modelos de series de tiempo de *machine learning* desarrollados, para que su uso sea más sencillo, interactivo y de mayor utilidad para usuarios tales como autoridades en salud pública. En seguida se muestra un diagrama de la arquitectura de la aplicación web realizada que permite dar una descripción del sistema.

En el diagrama mostrado en la Figura 45 se observa una aplicación web *fullstack*, realizada con angular para *front-end* con las funcionalidades e interfaz de usuario. Por el lado del *back-end*, este fue realizado en *framework flask* de Python y consistió en cuatro principales microservicios que se encargaban de seleccionar el modelo de *machine learning* correspondiente a la petición y la solicitud de parte del *front-end*; ya sea para predecir a partir de sólo casos ingresados manualmente, de sólo variables de observación ingresadas de manera manual, de variables de observación y de contexto digitadas manualmente o mediante un archivo de .xlsx con los datos. Finalmente, se tienen una carpeta en la aplicación, donde se tienen almacenados los diferentes modelos (*.pickles*), que se utiliza de adecuado a la petición del usuario.

Figura 45: Diagrama de la arquitectura de la aplicación web



La aplicación implementa tres modelos de series de tiempo para predecir los casos de LC a 1 semana a futuro (1. A partir de los casos anteriores, 2. A partir de las variables de observación y casos anteriores y 3. A partir de las variables contextuales, de observación y casos anteriores) usando el algoritmo *XGBoost.* Además, se agregaron 2 funcionalidades para predecir los casos a 2 y a 4 semanas a futuro.

En la aplicación desarrollada, los modelos implementados y su estandarización fueron almacenados como archivos tipo *pickles, c*on el objetivo de facilitar su uso y futuros cambios en la aplicación.

En el Anexo 7, se muestran unas capturas de pantalla de la aplicación web en funcionamiento. Para acceder a todo el código tanto de la aplicación web como de las diferentes implementaciones y modelos realizados, se puede consultar el siguiente repositorio en GitHub: https://github.com/TesisMaestriaCienciaDatos/ModeloPredictivoLC

Conclusiones y trabajo futuro

Conclusiones

Encontrar conocimiento útil es una etapa importante del proceso *KDD,* sin embargo, las etapas previas como la selección, preprocesamiento y transformación de los datos jugaron un papel muy importante en los resultados de este tipo de investigaciones; además, se debe tener mucho cuidado en estos pasos para garantizar un conjunto de datos consistente y confiable. Recolectar los datos directamente de fuentes como sensores remotos de la NASA y bases de datos nacionales permitió una mayor confianza en los datos. También, identificar valores erróneos o faltantes permitió construir un conjunto de datos sin ruido y que se estandarizó para ser usado en los algoritmos de predicción.

Uno de los aspectos relevantes encontrados en este trabajo tiene que ver con la gran cantidad de variables utilizadas y las diferentes técnicas implementadas con las variables del conjunto de datos, obteniendo en todos los casos mejores resultados con el algoritmo *XGBoost* el cual fue mucho más rápido en el entrenamiento y en la generación del modelo, además de presentar un mejor desempeño para el conjunto de datos de esta investigación.

El algoritmo *XGBoost* presentó mejores resultados de forma incremental para mayor cantidad de variables. Con el *dataset* que incluía las variables de observación y de contexto, dicho algoritmo presentó el mejor desempeño y asimismo no aumentó considerablemente la complejidad del mismo, a diferencia de las redes neuronales (MLP), las cuales disminuían de forma significativa su desempeño con el incremento de las variables de entrada, además de requerir más tiempo en ser entrenadas o generar el modelo.

El uso de las variables meteorológicas, ambientales y socioeconómicas usadas en este estudio son útiles para desarrollar modelos de predicción de LC. De igual forma, con el uso de técnicas de *machine learning* se pueden crear modelos para la predicción de casos de LC. Los resultados obtenidos en este estudio pueden ayudar a la implementación un de un sistema futuro de alerta temprana para la atención de brotes de LC a nivel departamental a través de las secretarías de salud.

Trabajo futuro

Respecto del trabajo que sigue con el presente proyecto, se sugiere explorar el desarrollo de un modelo de regresión con series de tiempo a escala municipal. De la misma forma, se sugiere incorporar más variables meteorológicas y socioeconómicas, al igual que intentar conseguir los datos a una mayor resolución espacial (p.e. vereda o barrio).

También, se recomienda realizar un estudio teniendo en cuenta los subreportes (datos de la enfermedad que por alguna razón no fueron reportados) de casos de LC para tener información más realista de la enfermedad e incorporar esta variable dentro de los modelos de *machine learning*. Finalmente, se recomienda probar con otras técnicas de *machine learning* que permitan trabajar de manera eficiente con series de tiempo.

Anexos

Anexo 1: Descripción y caracterización de las variables incorporadas

El campo dato corresponde a la información o concepto general de la variable en cuestión, la columna variable corresponde al nombre de la variable utilizada en el modelo, la columna descripción hace referencia al detalle y explicación de la variable y finalmente, el campo fuente en la tabla se refiere al origen o fuente de los valores de la variable.

Dato	Variable	Descripción	Fuente
Socioeconómicas			
Caracterización de entidad territorial DNP	Cobertura neta en educación secundaria	Variable porcentual, con valores entre 0 y 100. Representa el cubrimiento de educación secundaria en los municipios	Territata DNP
	Puntaje promedio de pruebas Saber 11 - Lectura crítica	Puntos obtenidos en promedio a nivel municipal en pruebas ICFES.	Territata DNP

	Tasa de mortalidad infantil en menores de 5 años	Mortalidad de niños por cada 1000	Terridata DNP
	Cobertura de vacunación pentavalente en menores de 1 año	Porcentaje entre 0 y 100 de vacunación en bebés	Terridata DNP
	Cobertura de acueducto (REC)	Porcentaje entre 0 y 100 de acueducto en el municipio	Terridata DNP
	Densidad poblacional	Número de habitantes por kilómetro cuadrado	Terridata DNP
	Ingresos totales per cápita	Miles de millones que recibe el municipio	Terridata DNP
	Tasa de homicidios (x cada 100.000 habitantes)	Número de homicidios por cada 100.000 habitantes en el municipio	Terridata DNP
	Número de personas secuestradas	Cantidad de secuestros a nivel municipal	Terridata DNP
	Número de personas desplazadas	Cantidad de personas desplazadas a nivel municipal	Terridata DNP

	Tasa de hurtos (x cada 100.000 habitantes)	Representa el número de hurtos municipales por cada 100.000 habitantes	Terridata DNP
	Crecimiento poblacional	Tasa de crecimiento de la población	Terridata DNP
Características propias del municipio DNP	Disparidades económicas	Es un valor numérico, entre más alto el número, más disparidad económica.	Tipologías DNP
	Hectáreas de cocaína	Es el número de hectáreas de coca por cada municipio	Tipologías DNP
	Dimensión urbana	Representa que tan desarrollado está el municipio, valor numérico	Tipologías DNP
	Dimensión económica	Tiene en cuenta los datos tributarios del municipio y el IPM con valores de 0 a 1	Tipologías DNP
	Dimensión de calidad de vida	Se calcula con base en los datos de cobertura de educación y vivienda municipal	Tipologías DNP

	Dimensión de seguridad	Es representada por un valor numérico entre 0 y 1 y tiene en cuenta la tasa de homicidios, secuestros, hurtos y coca municipal	Tipologías DNP
Geográficas			
Datos geográficos DANE	Código DANE-periodo	Variable que une la información entre el código DANE y el periodo (semana consecutiva o mes consecutivo)	DANE
	Región	Representa las diferentes regiones a las que pertenece cada municipio de Colombia	DANE
	Subregión	Representa las diferentes subregiones a las que pertenece cada municipio	DANE
	Departamento	Es el departamento al que está asociado cada municipio	DANE

	Municipio	La ubicación geográfica más específica a la que se tiene acceso.	DANE
	Área en km2	Representa el área del municipio, dada en Km2	DANE
Ambientales			
Características ambientales de los municipios IDEAM	Porcentaje de zona antrópica	Es el porcentaje de cobertura antrópica (hecha por el hombre) del municipio	IDEAM
	Porcentaje de bosques	Porcentaje de zonas de bosques que tiene el municipio	IDEAM
	Cultivos permanentes	Porcentaje de cultivos permanentes como cacao, palma africana, cítricos, etc.	IDEAM
	Cultivos transitorios	Porcentaje de cultivos como yuca, trigo, etc.	IDEAM

	Herbazales	Porcentaje de hierbas y arbustos	IDEAM
	Mosaico	Porcentaje de coberturas de mosaico	IDEAM
	Pastos	Porcentaje de pastos en el municipio	IDEAM
	Vegetación secundaria	Porcentaje de vegetación secundaria en los municipios, como potreros abandonados	IDEAM
	Zonas acuáticas	Porcentaje de zonas acuáticas en los municipios, como ríos, lagunas, etc.	IDEAM
	Otras coberturas	Es el porcentaje de coberturas no reconocidas por satélites, o tapadas por nubes	IDEAM
	Árido	Es un porcentaje de las zonas áridas del municipio	IDEAM

	Seco	Representa el porcentaje de las zonas secas del municipio	IDEAM
	Húmedo	Porcentaje de zonas húmedas	IDEAM
	Pluvial	Porcentaje de zonas pluviales	IDEAM
	Metros Sobre el nivel del mar (msnm)	Altura promedio en metros de los municipios sobre el nivel del mar	IDEAM
	Mínima altura sobre el nivel del mar (minmsnm)	Mínima altura de los municipios sobre el nivel del mar	IDEAM
	Máxima altura sobre el nivel del mar (maxmsnm)	Representa la máxima altura de los municipios sobre el nivel del mar	IDEAM
	Rango de metros sobre el nivel del mar (rango msnm)	Es la diferencia entre la altura máxima y la mínima	IDEAM
	Deforestación	Es el porcentaje de bosque destruido	IDEAM

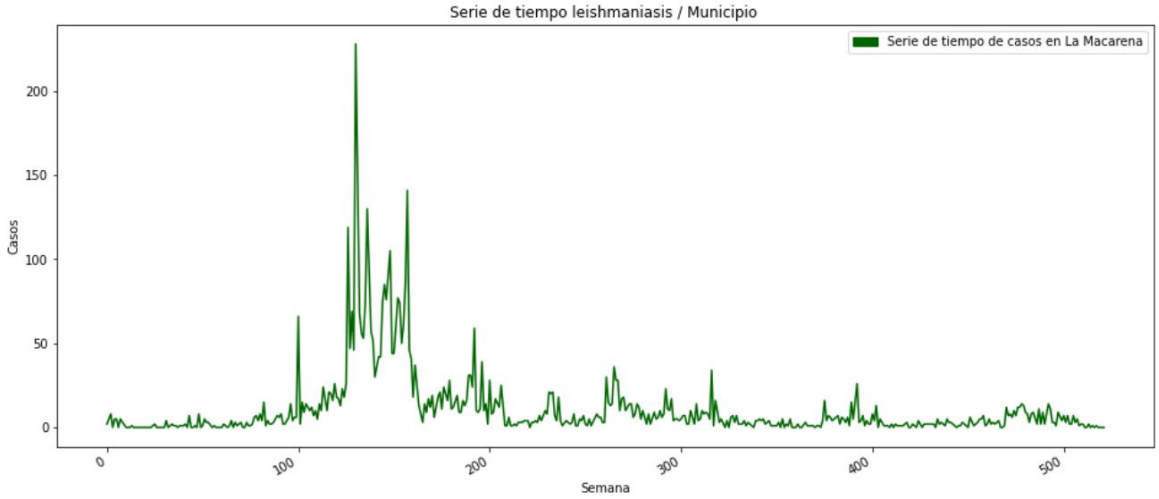
	Cuerpos de agua	Porcentaje de zonas del municipio donde se puede acumular agua	IDEAM
	Zonas susceptibles de inundación	Representa el porcentaje de las zonas de los municipios que se pueden inundar, por ejemplo, las cercanas a ríos	IDEAM
	Zonas inundables	Valor expresado en Km2 que representa las zonas inundables por municipio.	IDEAM
Observación			
Variables de observación climáticas	Temperatura	Es el valor de la temperatura de cada municipio tomada semanal y mensualmente en grados centígrados	NASA Giovanni
	Máxima temperatura	Valor máximo de la temperatura de un municipio en específico	NASA Giovanni

	Mínima temperatura	Valor mínimo de la temperatura de un municipio en específico	NASA Giovanni
	Precipitación	Cantidad de lluvia en el municipio, medida en milímetros (mm)	NASA Giovanni
	Máxima precipitación	Valor máximo de la precipitación de un municipio en específico	NASA Giovanni
	Mínima precipitación	Valor mínimo de la precipitación de un municipio en específico	NASA Giovanni
	Promedio temperatura	Es el promedio de la temperatura para el municipio en todo el periodo de estudio	NASA Giovanni
	Promedio precipitación	Representa el promedio de la precipitación para el municipio en todo el periodo de estudio	NASA Giovanni

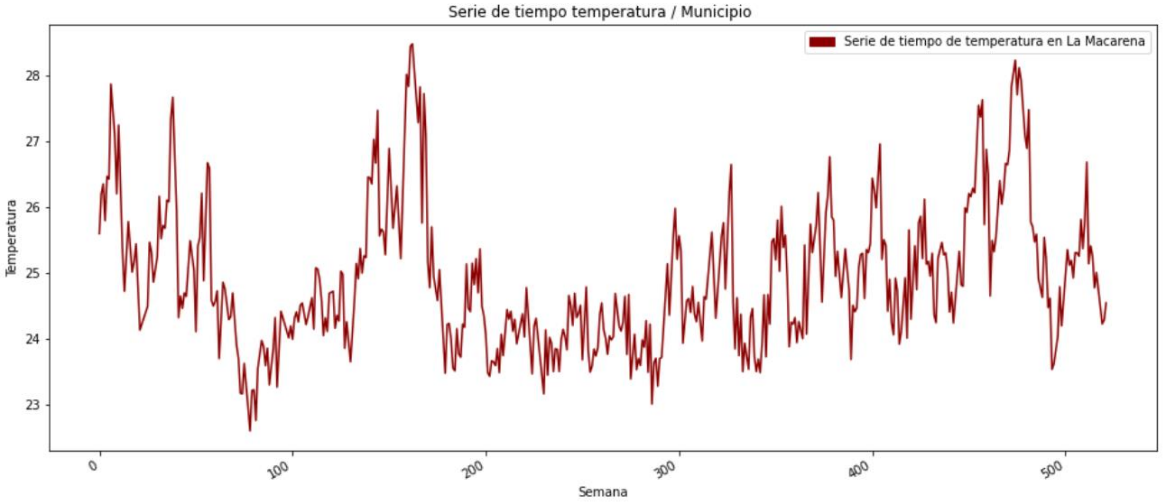
Vegetación UA	EVI	Es el índice de vegetación mejorado (valores mayores a cero)	Universidad de Arizona
	NVDI	Es el índice de vegetación de diferencia normalizada (valores mayores a cero)	Universidad de Arizona
SIVIGILA	Total de población	Representa la población del municipio, es un sólo dato en todo el periodo de estudio	SIVIGILA
	Casos	Es la variable de observación objetivo, es un valor numérico discreto y representa la cantidad de casos de leishmaniasis municipal	SIVIGILA

Anexo 2: Series de tiempo

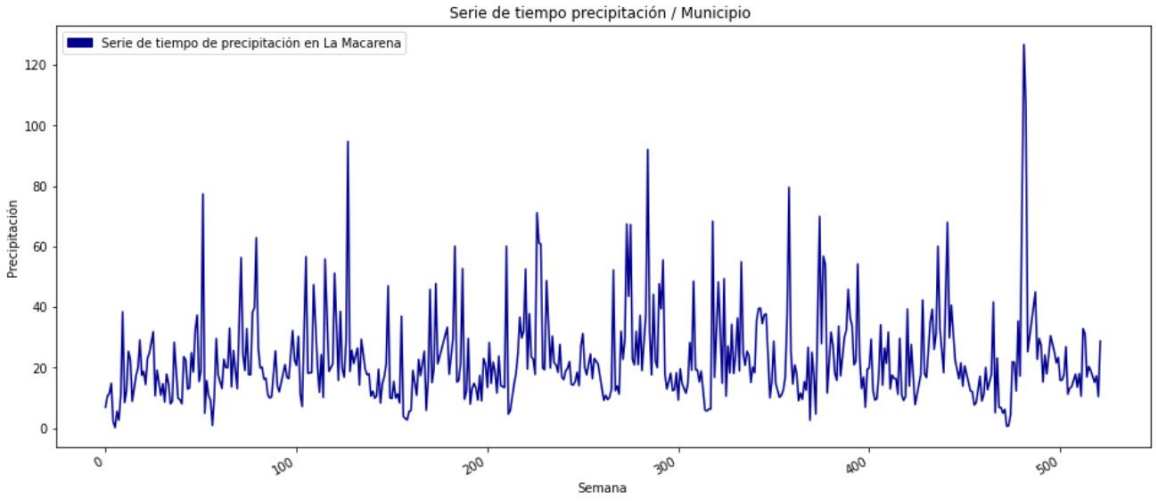
Anexo 2.1: Serie de tiempo de casos en La Macarena



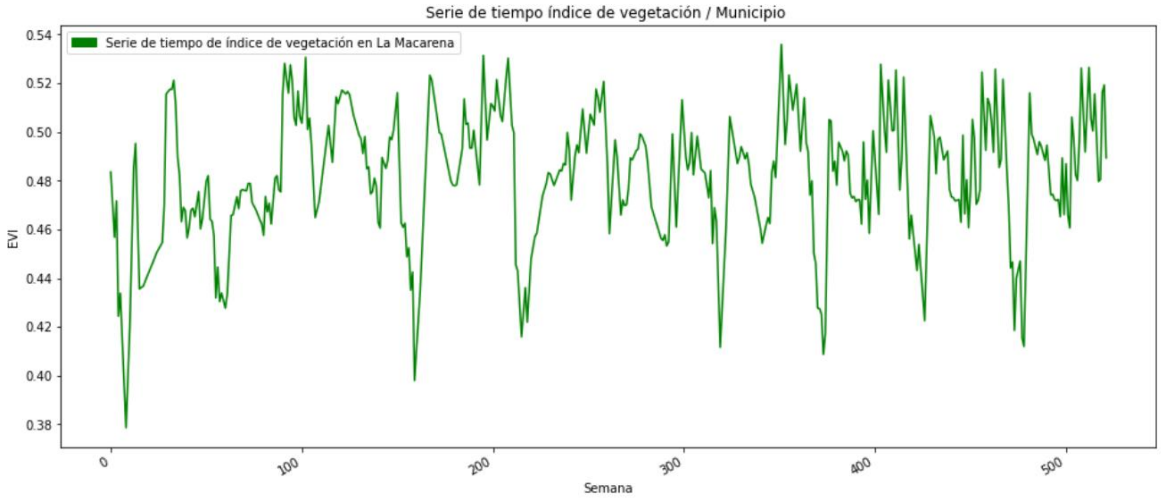
Anexo 2.2: Serie de tiempo de temperatura en La Macarena



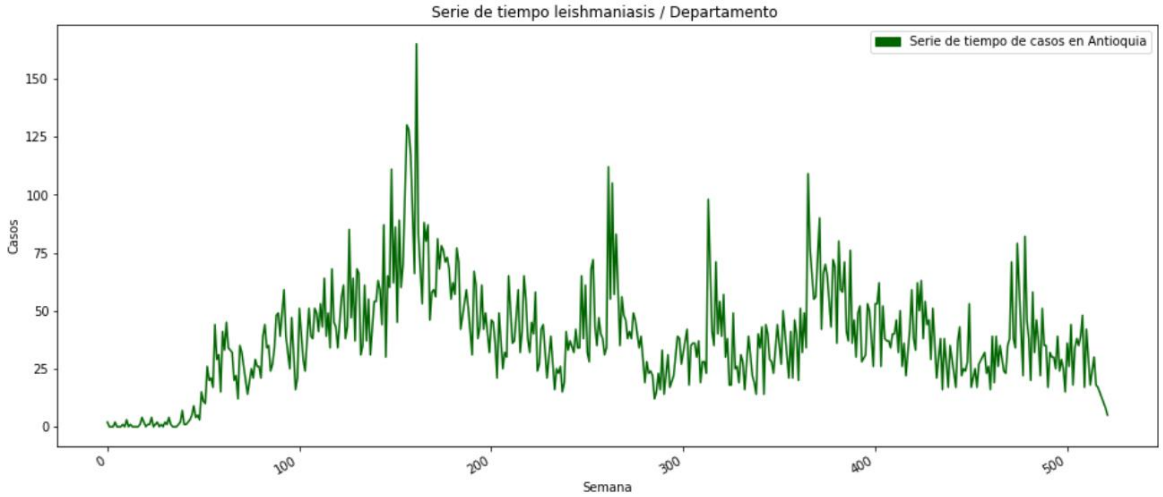
Anexo 2.3: Serie de tiempo de precipitación en La Macarena



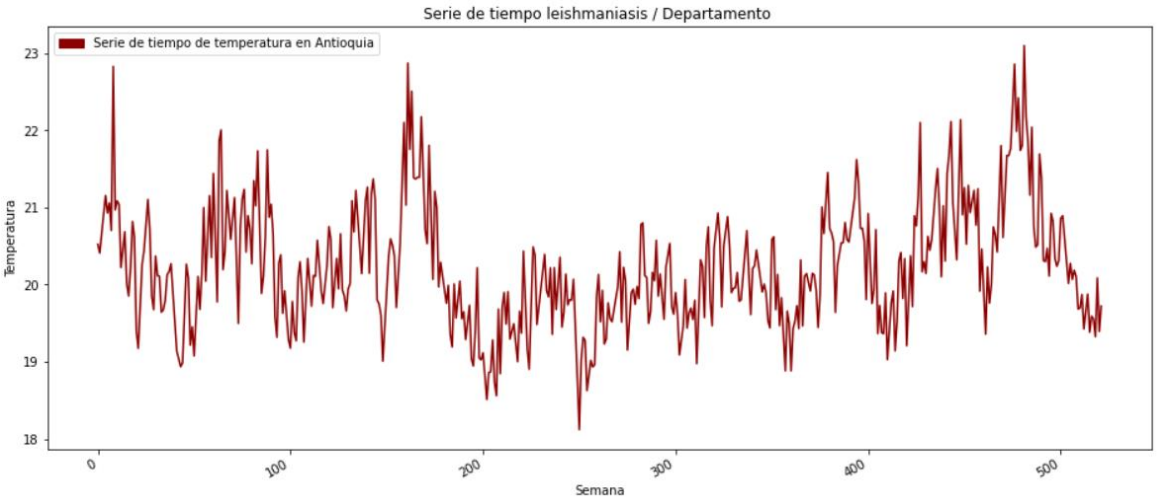
Anexo 2.4: Serie de tiempo de EVI en La Macarena



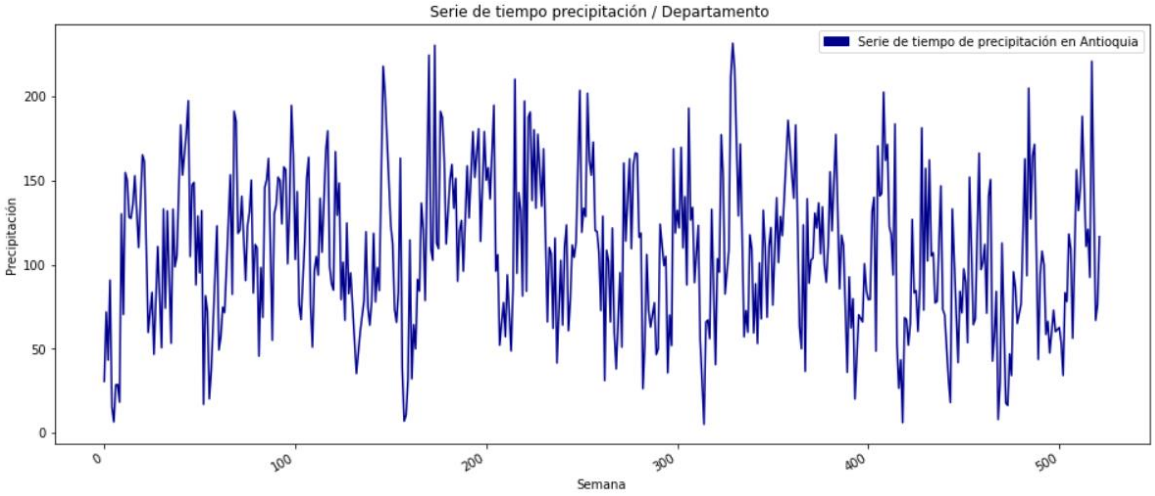
Anexo 2.5: Serie de tiempo de casos en Antioquia



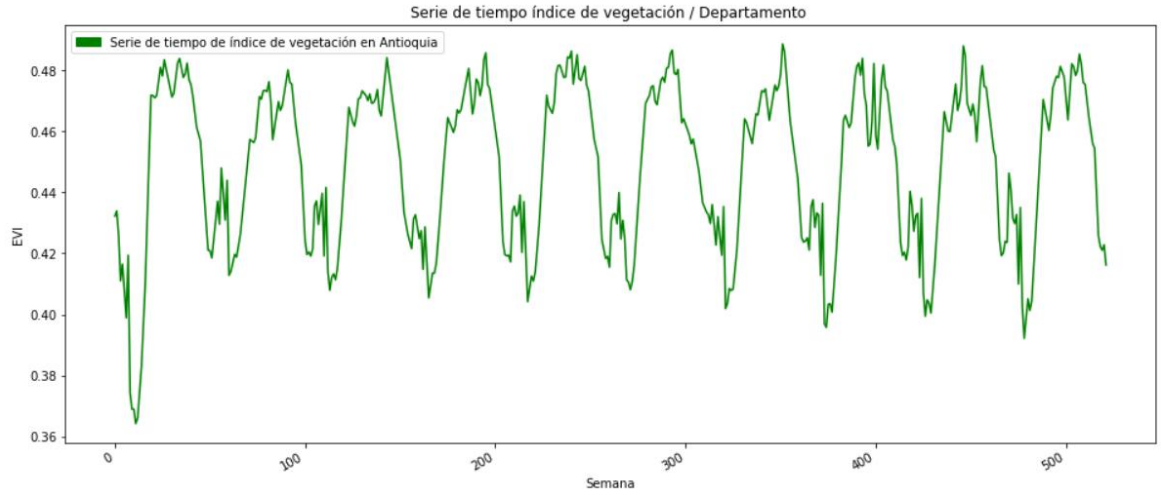
Anexo 2.6: Serie de tiempo de temperatura en Antioquia



Anexo 2.7: Serie de tiempo de precipitación en Antioquia



Anexo 2.8: Serie de tiempo de EVI en Antioquia



Anexo 3: *Clustering*

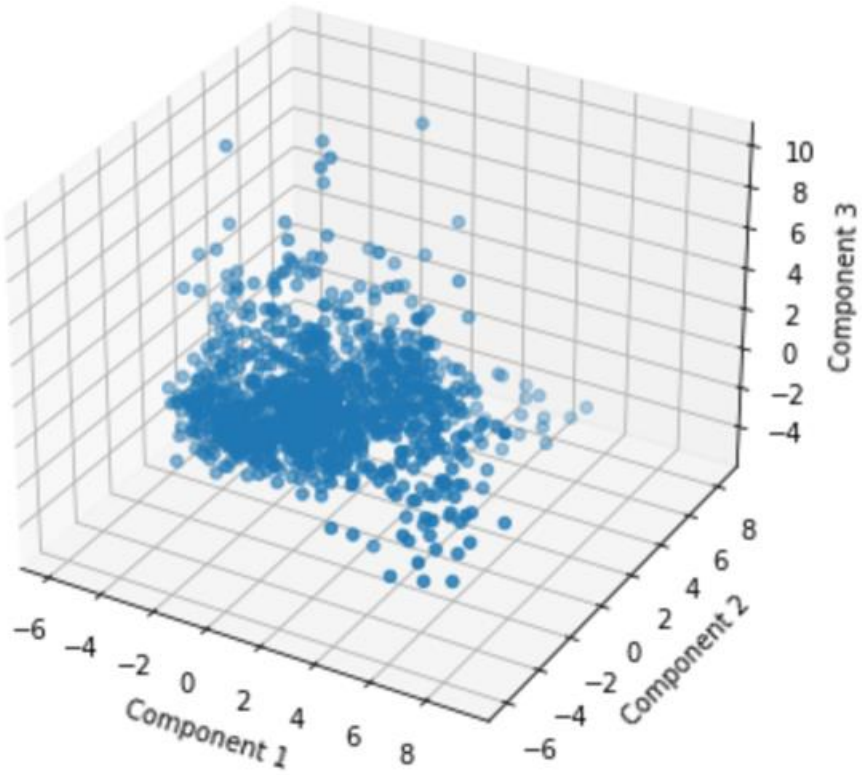
Para reducción de dimensionalidad se implementó PCA (Análisis de Componentes Principales). Esto se hizo mediante el algoritmo de PCA de *sklearn*, utilizando como argumento de entrada el porcentaje de explicabilidad deseado en el conjunto de datos. En este caso se seleccionó un valor de explicabilidad del 80%, con el que se obtuvo un conjunto de 15 componentes a partir de las 46 variables contextuales, las 3 de observación y la variable respuesta. El Anexo 3.1 muestra los 3 componentes principales.

Se realizó un agrupamiento (*clustering*) de los datos partiendo de los 15 componentes seleccionados en PCA, utilizando el algoritmo *k-means.* Se seleccionaron 3 *clusters*

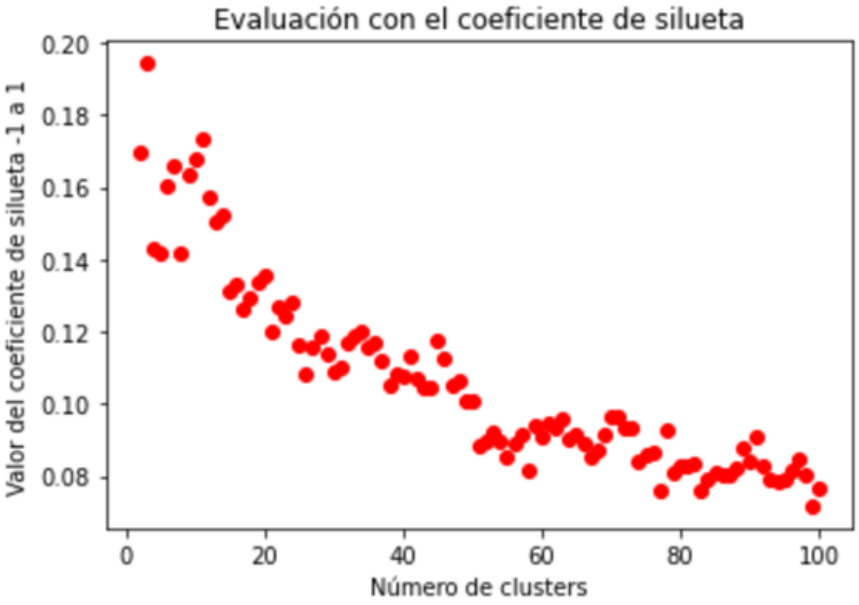
teniendo como criterio el coeficiente de silueta (0.181). En el Anexo 3.3 se puede detallar que no hay una clara división entre los 3 *clusters* y los elementos de cada conglomerado se superponen entre sí, sin haber una clara división entre los grupos lo que permitió concluir que no es muy viable trabajar con agrupación con este conjunto de datos.

El análisis de *clustering* con las variables de observación permitió tener 4 *clusters* (valor de silueta 0.34) (Anexo 3.4). Los grupos generados en este análisis y los análisis posteriores con *clustering* no produjeron resultados útiles para el objetivo de clasificar los municipios entre aquellos que tenían casos de LC y aquellos que no tenía casos.

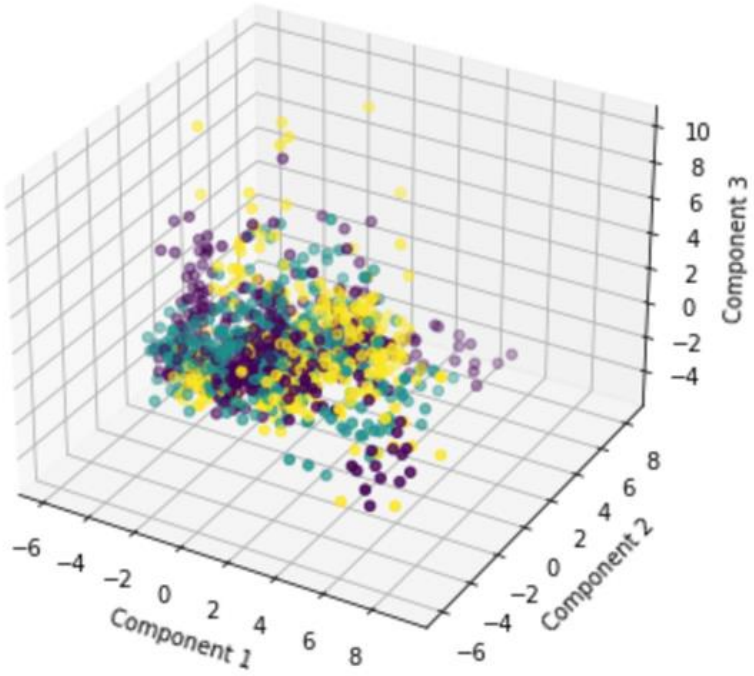
Anexo 3.1: PCA 3 principales componentes



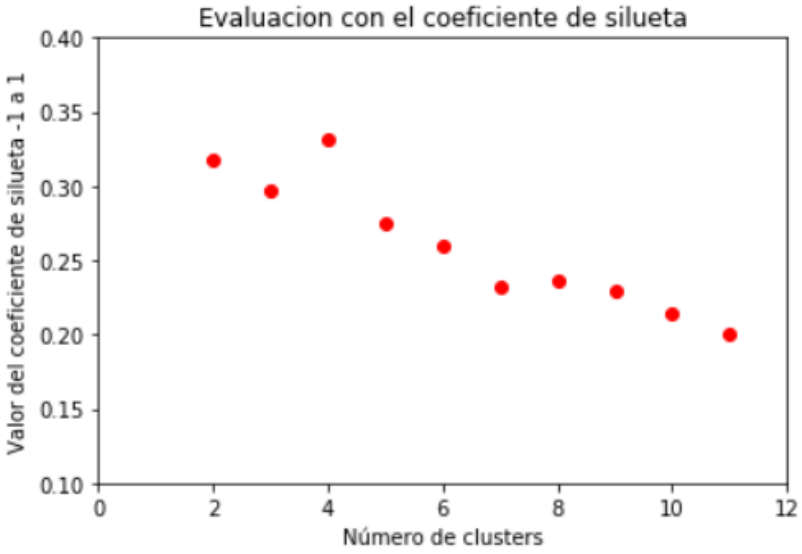
Anexo 3.2: Coeficiente de silueta todas las variables



Anexo 3.3: 3 *Clusters* con 3 componentes principales



Anexo 3.4: Coeficiente de silueta para las variables de observación



Anexo 4: Numero de retrasos incluidos en cada departamento a partir del análisis de correlación cruzada

Departamento	Número de retrasos
Antioquia	9
Atlántico	19
Cundinamarca	10
Bolívar	6
Boyacá	8
Caldas	9
Caquetá	7
Cauca	4
Cesar	9

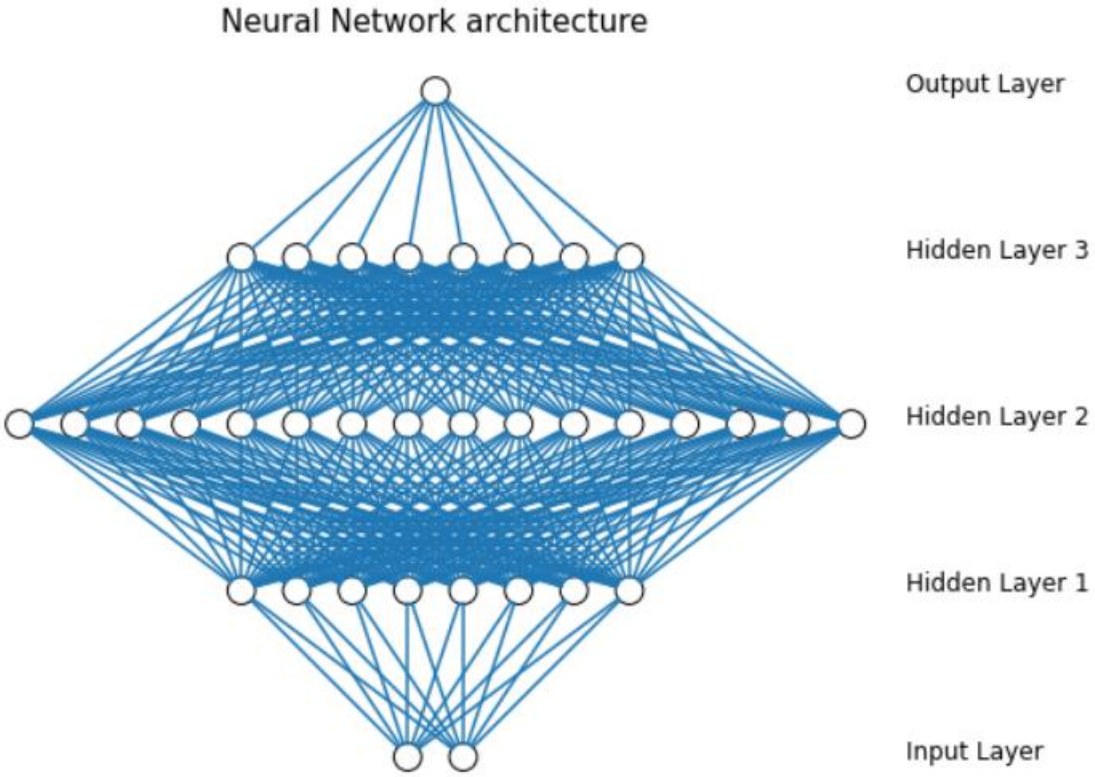
Córdoba	9
Chocó	9
Huila	9
Guajira	8
Magdalena	6
Meta	9
Nariño	8
Norte de Santander	8
Quindío	9
Risaralda	8
Santander	8
Sucre	8
Tolima	6
Valle de Cauca	9
Arauca	21
Casanare	4
Putumayo	7
Guaviare	9
Vichada	11

Anexo 5: Modelo con MLP para Antioquia

El modelo realizado con redes neuronales para el departamento de Antioquia fue perceptrón multicapa (*MLP* por sus siglas en inglés) con la arquitectura presentada en la Anexo 5.1. Se puede observar que la arquitectura de la red neuronal corresponde a una capa de entrada con 9 neuronas (no se muestra la imagen a escala real debido a que sería imperceptible su diseño), tres capas ocultas, la primera con 61 neuronas, la siguiente con 120 neuronas y la última con 57. Y al final, una capa de salida con 1 neurona, que equivale a la respuesta de cantidad de casos predichos (Anexo 5.1). Para llegar a la arquitectura previamente descrita, se hizo *tuning* (afinamiento) de los hiperparámetros de la red neuronal el conjunto de parámetros *afinados* fueron los siguientes mostrados en el Anexo 5.2. Después de realizar el *tuning* con el conjunto de datos (Anexo 5.2), en la tercera columna se muestra el resultado de los hiperparámetros seleccionados. Y finalmente el Anexo 5.3 muestra el real vs el predicho de la prueba del modelo *MLP.*

Por otro lado, también se hizo la implementación con *XGBoost*, encontrando resultados un poco mejores que los del perceptrón multicapa. En el Anexo 5.4 se detalla el universo de hiperparámetros escogidos para el *tuning* del modelo de regresión con *XGBoost*. Una vez realizado este *tuning,* el Anexo 5.5 muestra el real vs el predicho del modelo. Finalmente las métricas obtenidas por ambos modelos son comparadas en el Anexo 5.6.

Anexo 5.1: Arquitectura del perceptrón multicapa para la serie de tiempo de Antioquia.



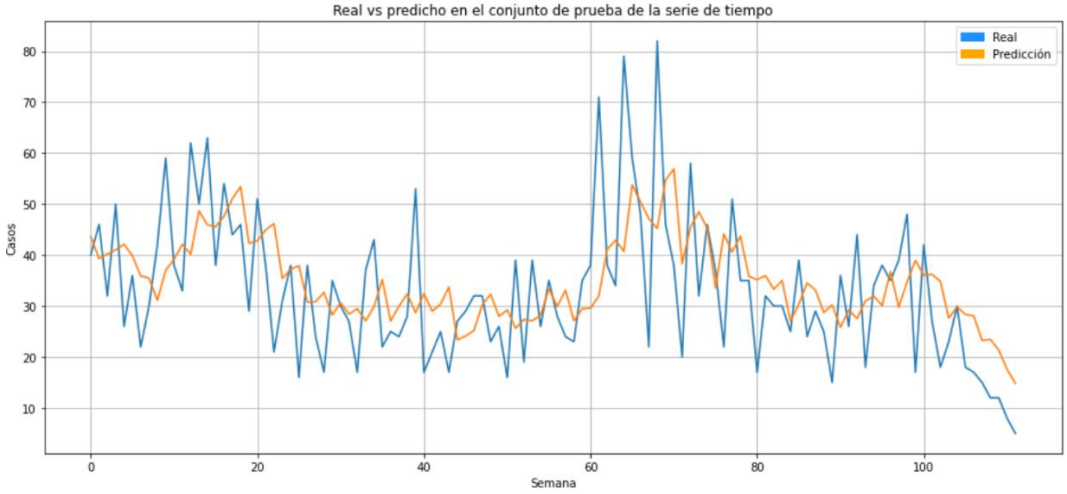
Arquitectura para la red neuronal del modelo de Antioquia, con 9 neuronas de entrada, 3 capas ocultas, la primera con 61 neuronas, la segunda con 120 y la tercera con 57 la cual se comunica con la neurona de salida.

Anexo 5.2: Conjunto de hiperparámetros del *MLP afinados* para Antioquia

Hiperparámetro	**Valores evaluados**	**Valor seleccionado**
Función de activación	'logistic', 'relu', 'tanh'	'relu'
Optimizador (*solver*)	'lbfgs', 'adam'	'adam'
Alpha	np.logspace (0.0001, 2)	1
Cantidad de capas ocultas	1, 2 y 3	3
Cantidad de neuronas	1-90	(61, 120, 57)

Tasa de aprendizaje (Learning rate)	'constant', 'invscaling', 'adaptive'	'invscaling'
Batch size	64,128, 256, 512	512
Máximo número de iteraciones	10,50,200,1000	200

Anexo 5.3: Real vs predicción del *MLP* para Antioquia: el eje *Y* es el valor de casos de LC. El eje *X* hace referencia a 112 semanas de prueba



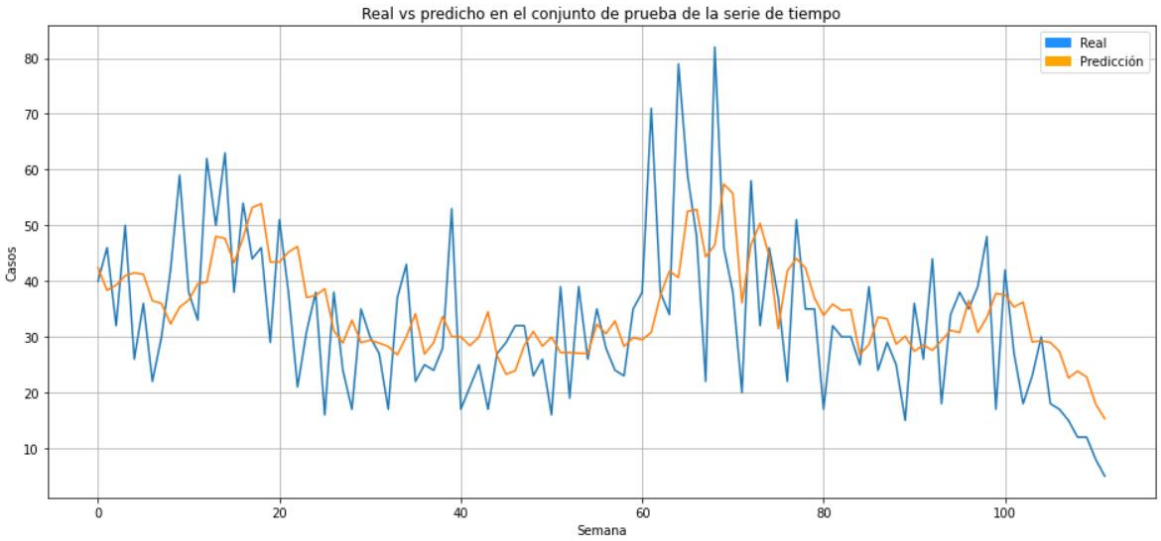
Anexo 5.4: Conjunto de hiperparámetros de XGBoost *afinados* para Antioquia

Hiperparámetro	Valores evaluados	Valor seleccionado
Tasa de aprendizaje (Learning rate)	0.001, 0.01, 0.05, 0.10, 0.20, 0.50, 0.90	0.001
Profundidad máxima (max depth)	3, 4, 5, 6, 8, 10, 12, 15	6
Porcentaje de características por árbol (colsample bytree)	0.3, 0.4, 0.5, 0.7	0.5

Gamma	0.0, 0.1, 0.2, 0.3, 0.4	0.3
Número de árboles (n_stimators)	100, 1000, 10000	10000

Presenta el universo de hiperparámetros del XGBoost y su selección después del afinamiento.

Anexo 5.5: Real vs predicción *XGBoots* para Antioquia: el eje *Y* es el valor de casos de LC. El eje *X* hace referencia a 112 semanas de prueba



Anexo 5.6: Comparación de las métricas de ambos modelos para Antioquia

Métrica	XGBoost	MLP
Mean absolute error	9.669	9.897
Mean squared error	149.980	165.124
R2 score	0.248	0.198
Explained variance score	0.263	0.201

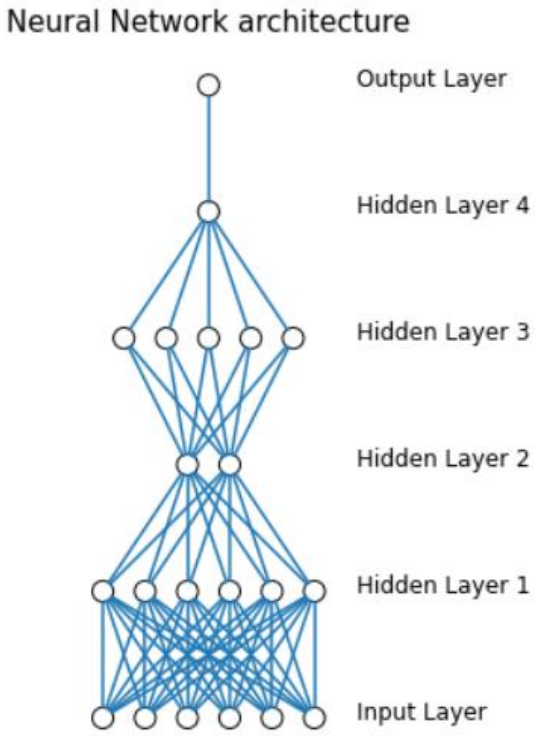
Anexo 6: Implementaciones a nivel nacional con *MLP*

Anexo 6.1: Conjunto de hiperparámetros *afinados* a nivel nacional de sólo casos para el *MLP*

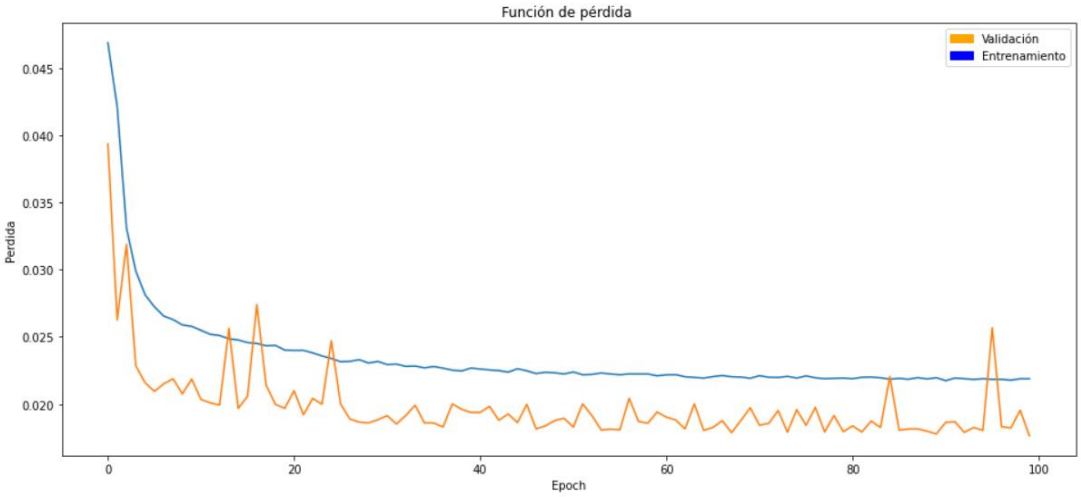
Hiperparámetro	Valores evaluados	Valor seleccionado
Función de activación	'logistic', 'softmax', 'softplus', 'softsign', 'relu', 'tanh', 'sigmoid', 'hard_sigmoid', 'linear'	'softplus'
Optimizador (*solver*)	'lbfgs', 'adam'	'adam'
Momentum	0.0, 0.2, 0.4, 0.6, 0.8, 0.9	0.4
Cantidad de capas ocultas	1, 2, 4, 5 y 6	4
Cantidad de neuronas	1-200	(6, 2, 5, 1)
Tasa de aprendizaje (Learning rate)	1e-1, 1e-2, 1e-3, 1e-4	0.1
Batch size	2,4,8,16,32,64,128,256,512,1024	8

Presenta el universo de hiperparámetros del perceptrón multicapa y su selección después del afinamiento.

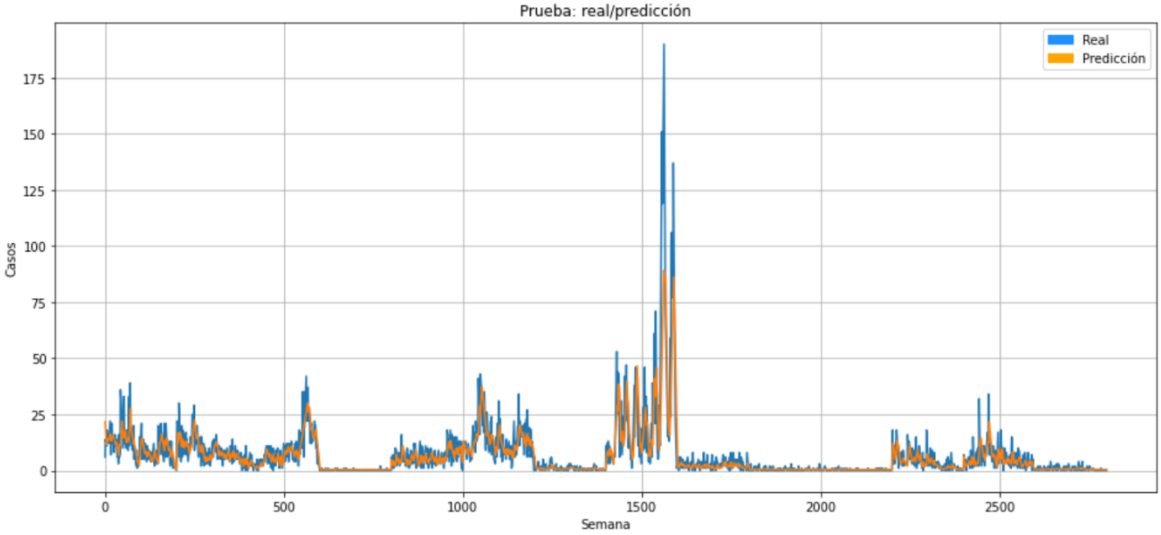
Anexo 6.2: Arquitectura red neuronal modelo nacional de sólo casos para el *MLP* Arquitectura para la red neuronal del modelo nacional de sólo casos con 6 neuronas de entrada, 4 capas ocultas, la primera con 6 neuronas, la segunda con 2, la tercera con 4 y la última con 1 neurona, la cual se comunica con la neurona de salida.



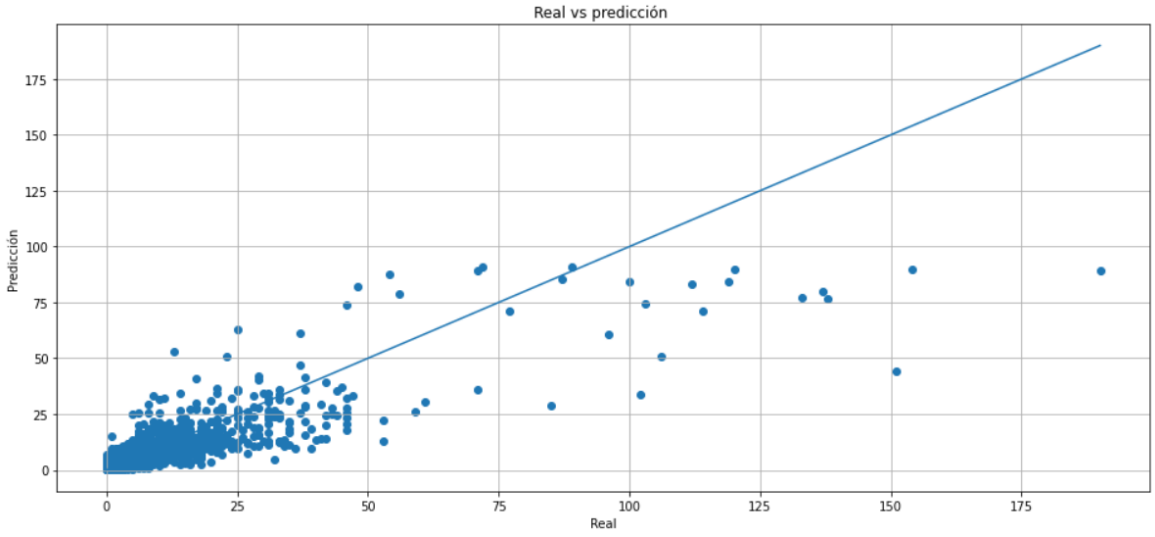
Anexo 6.3: Función de perdida con *mean absolute error* para el modelo nacional de sólo casos para el *MLP:* eje *Y* representa el valor del mae y el eje *X* representa la cantidad de ejecuciones del algoritmo



Anexo 6.4: Real vs predicción del modelo nacional de sólo casos para el *MLP:* el eje *Y* es el valor de casos de LC. El eje *X* hace referencia a 2,744 semanas de prueba, las cuales son las 98 últimas semanas de cada departamento.



Anexo 6.5: Comparación de valores observados vs predichos del modelo nacional con sólo casos LC para el *MLP:* el eje *Y* es el valor predicho y el eje *X* el valor real

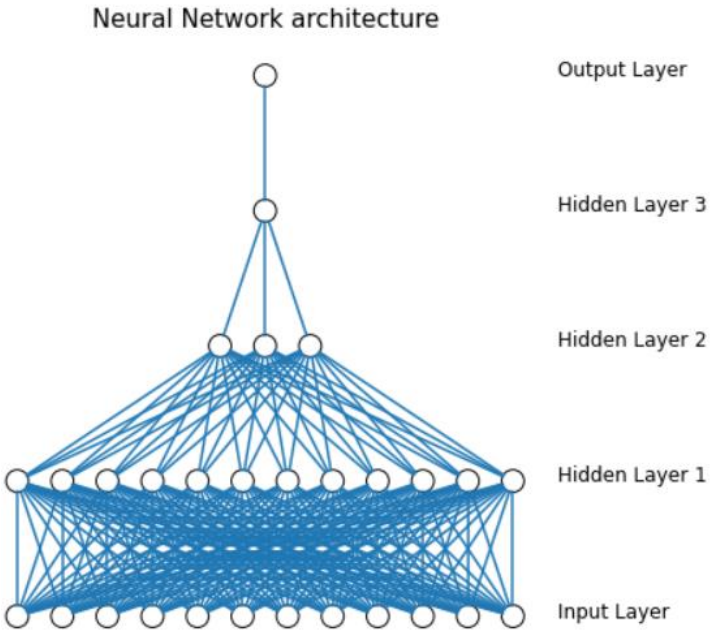


.

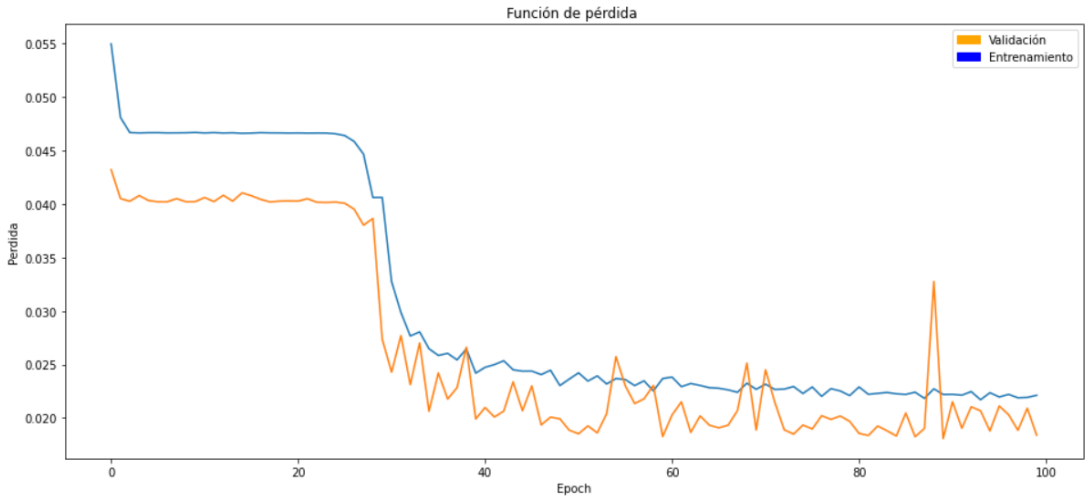
Anexo 6.6: Conjunto de hiperparámetros del perceptrón multicapa *afinados* a nivel nacional para las 4 variables de observación

Hiperparámetro	Valores evaluados	Valor seleccionado
Función de activación	'logistic', 'softmax', 'softplus', 'softsign', 'relu', 'tanh', 'sigmoid', 'hard_sigmoid', 'linear'	'softplus'
Optimizador (*solver*)	'lbfgs', 'adam'	'adam'
Momentum	0.0, 0.2, 0.4, 0.6, 0.8, 0.9	0.9
Cantidad de capas ocultas	1, 2, 4, 5 y 6	3
Cantidad de neuronas	1-200	(24, 5, 1)
Tasa de aprendizaje (Learning rate)	1e-1, 1e-2, 1e-3, 1e-4	0.1
Batch size	2,4,8,16,32,64,128,256,512,1024	16

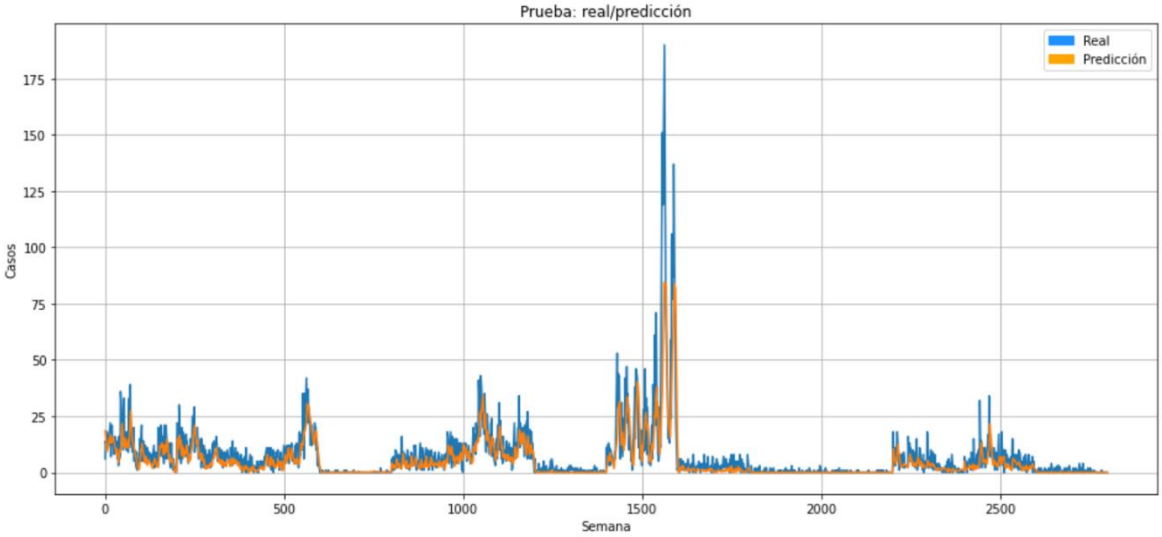
Anexo 6.7: Arquitectura red neuronal modelo nacional de variables de observación *MPL* Arquitectura para la red neuronal del modelo nacional de sólo casos con 24 neuronas de entrada, 3 capas ocultas, la primera con 24 neuronas, la segunda con 5 y la tercera con 1 neurona, la cual se comunica con la neurona de salida.



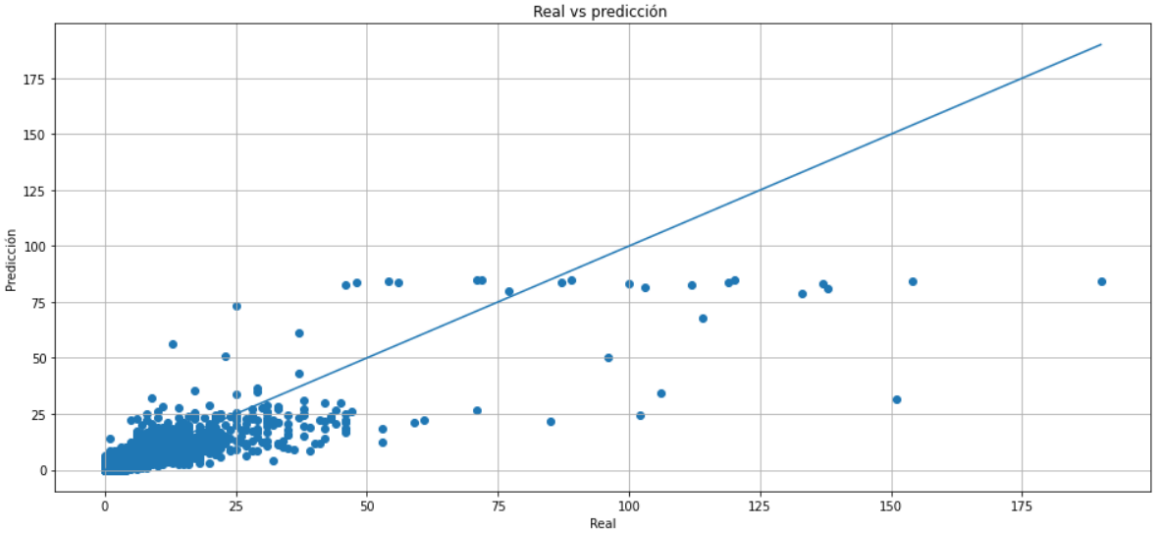
Anexo 6.8: Función de perdida con mean absolute error para el modelo nacional de variables de observación MLP: eje *Y* representa el valor del mae y el eje *X* representa la cantidad de árboles creados.



Anexo 6.9: Real vs predicción del modelo nacional de variables de observación y variable respuesta *MPL:* el eje *Y* es el valor de casos de LC. El eje *X* hace referencia a 2,744 semanas de prueba, las cuales son las 98 últimas semanas de cada departamento.



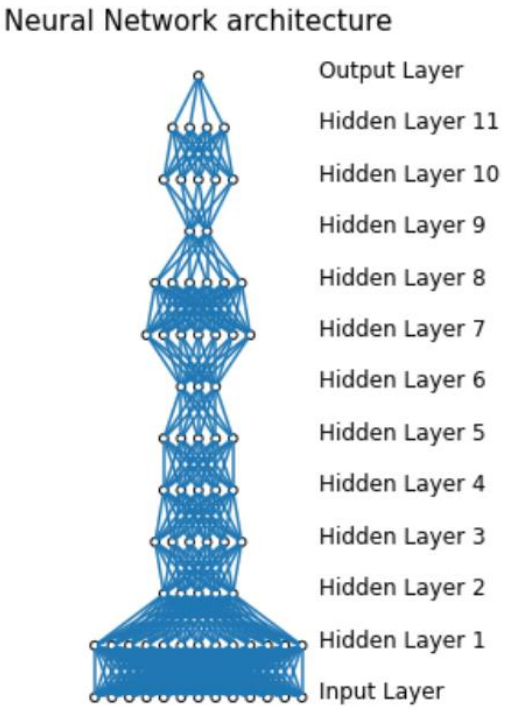
Anexo 6.10: Comparación de valores observados vs predichos del modelo nacional con las varioables de observación y la variable respuesta MLP: el eje *Y* es el valor predicho y el eje *X* el valor real



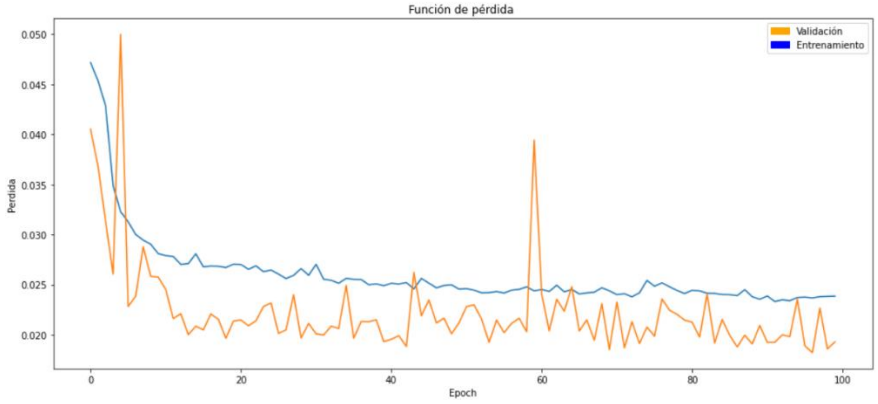
Anexo 6.11: Conjunto de hiperparámetros del *MLP afinados* a nivel nacional para las todas las variables

Hiperparámetro	Valores evaluados	Valor seleccionado
Función de activación	'logistic', 'softmax', 'softplus', 'softsign', 'relu', 'tanh', 'sigmoid', 'hard_sigmoid', 'linear'	'softsign
Optimizador (*solver*)	'lbfgs', 'adam'	'adam'
Momentum	0.0, 0.2, 0.4, 0.6, 0.8, 0.9	0.6
Cantidad de capas ocultas	1 - 18	11
Cantidad de neuronas	1-200	(68, 27, 29, 27, 28, 17, 34, 32, 11, 23, 20)
Tasa de aprendizaje (Learning rate)	1e-1, 1e-2, 1e-3, 1e-4	0.1
Batch size	2,4,8,16,32,64,128,256,512,1024	2

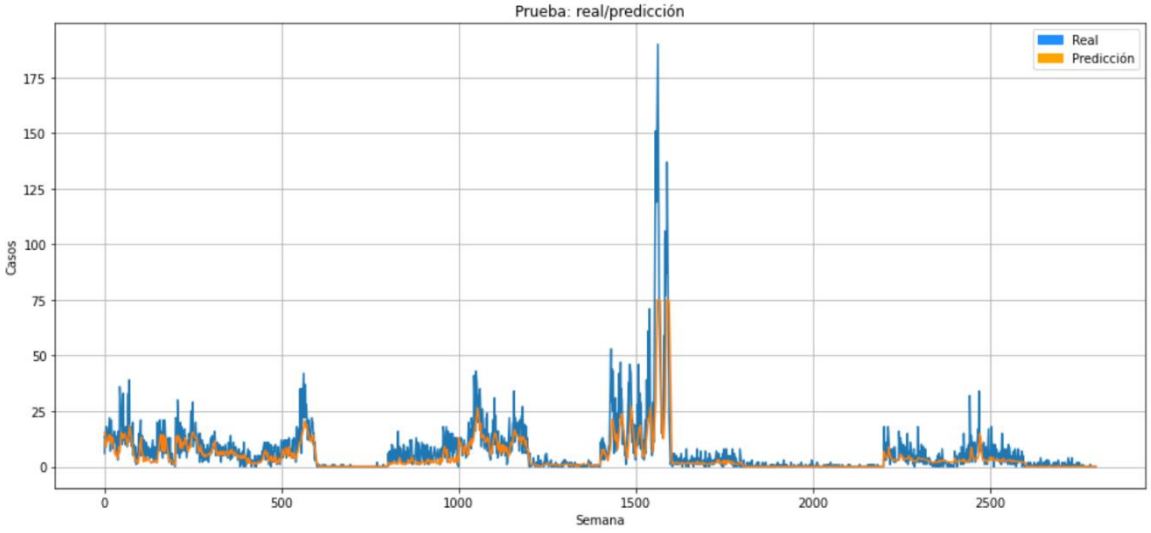
Anexo 6.12: Arquitectura red neuronal modelo nacional de todas las variables MPL Arquitectura para la red neuronal del modelo nacional de sólo casos con 68 neuronas de entrada, 11 capas ocultas, la primera con 68 neuronas, la segunda con 27, la tercera con 29, la cuarta con 27, la quinta con 28, la sexta con 17, la séptima con 34, la octava con 32, la novena con 11, la décima con 23 y la undécima con 20 neuronas, las cual se comunica con la neurona de salida.



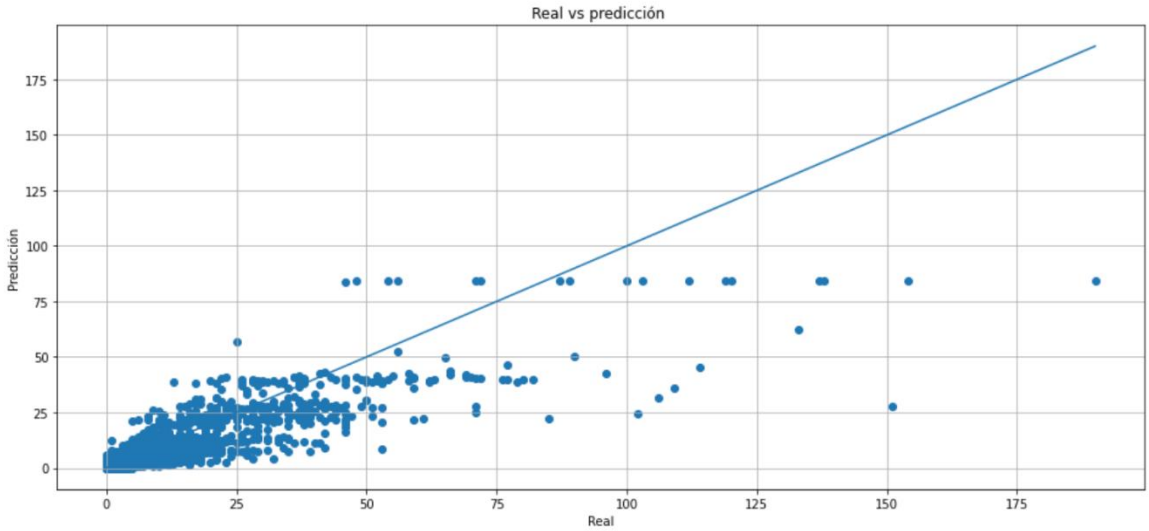
Anexo 6.13: Función de perdida con mean absolute error para el modelo nacional de todas las variables para el *MLP:* eje *Y* representa el valor del mae y el eje *X* representa la cantidad de árboles creados



Anexo 6.14: Real vs predicción del modelo nacional de todas las variables para el *MLP:* el eje *Y* es el valor de casos de LC. El eje *X* hace referencia a 2,744 semanas de prueba, las cuales son las 98 últimas semanas de cada departamento.



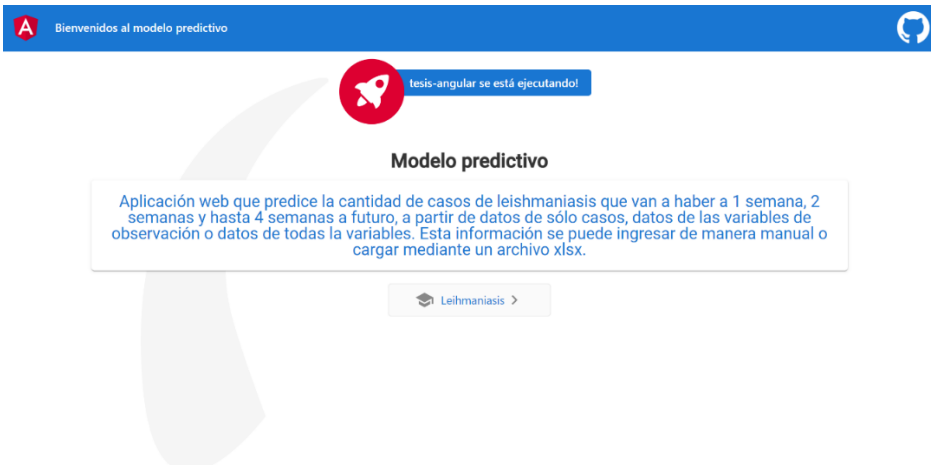
Anexo 6.15: Comparación de valores observados vs predichos del modelo nacional con todas las variables para el *MLP:* el eje *Y* es el valor predicho y el eje *X* el valor real



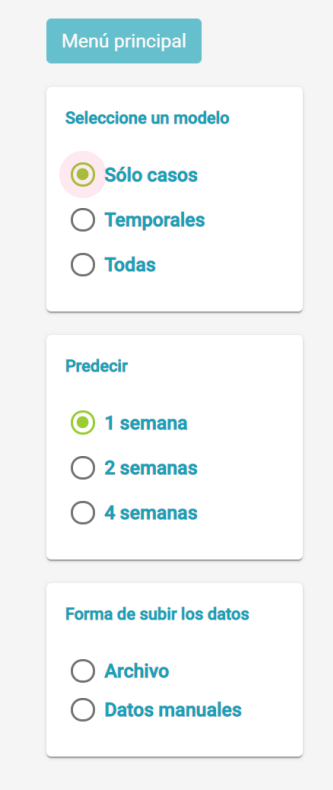
Anexo 7: Aplicación web

Se presentan las evidencias y capturas de la aplicación web que utiliza los modelos desarrollados

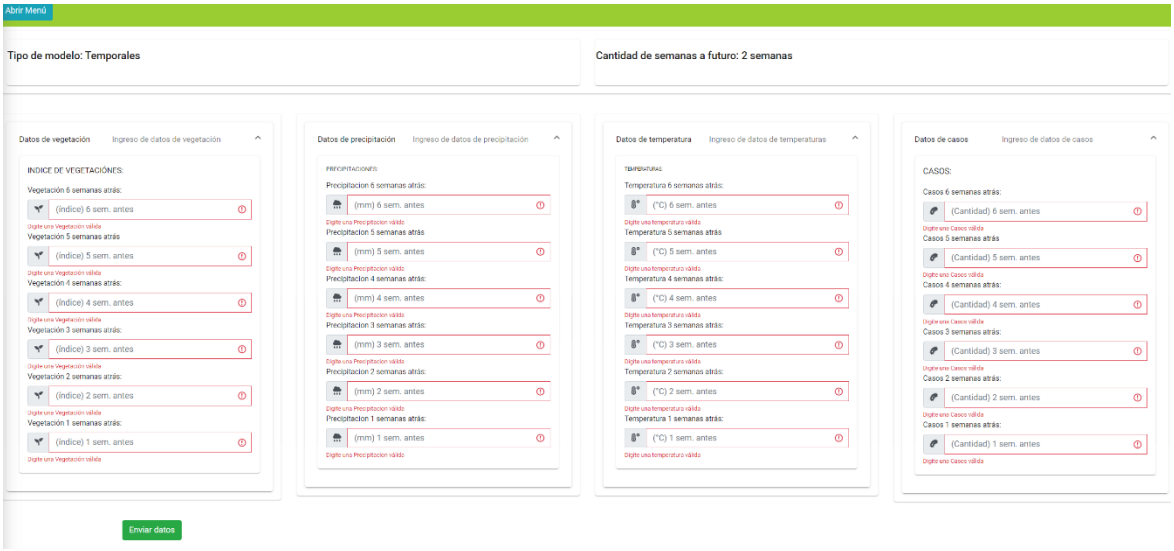
Anexo 7.1: Captura del *dashboard* de la aplicación web para predecir casos



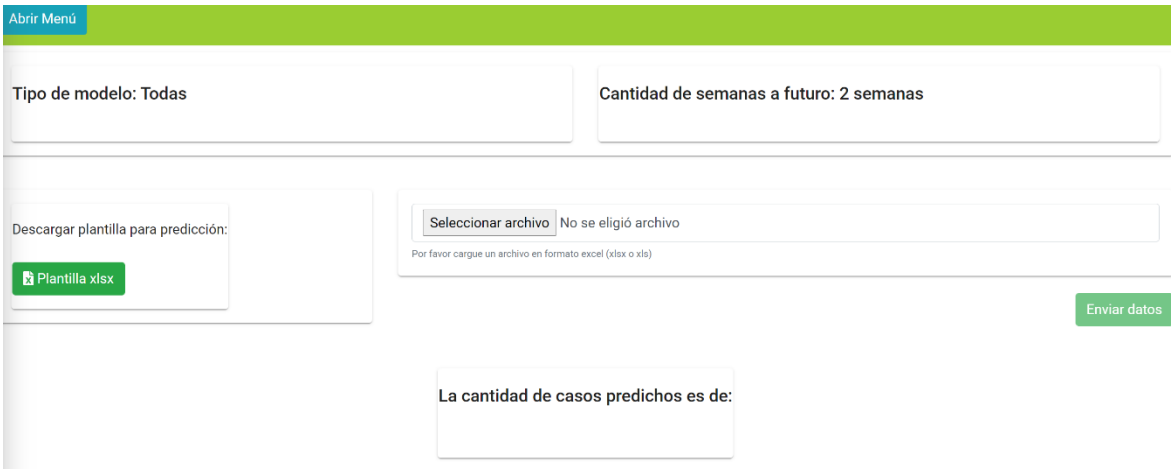
Anexo 7.2: Captura del *menú* con las diferentes opciones de la aplicación web



Anexo 7.3: Captura de la interfaz de la recepción de los datos de manera manual de la aplicación web



Anexo 7.4: Captura de la interfaz de la recepción de los datos mediante un archivo de la aplicación web



Bibliografía

[1] Agudelo, J. *Informe de evento de leishmaniasis.* Colombia, 2017.

[2] Zambrano, P. *Leishmaniasis.* Colombia: Bogotá, Instituto Nacional de Salud 2014.

[3] Sociedad colombiana de infectología. *Guía 2l. Guía de atención de la leishmaniasis.* Ministerio de la protección social, Colombia.

[4] Medical Care Development International. *Leishmaniosis: ciclo biológico.* Disponible en: https://www.mcdinternational.org/trainings/malaria/spanish/dpdx/HTML/Frames/G-L/Leishmaniasis/body_Leishmaniasis_pg1#Life%20Cycle

[5] Leishmaniosis: ciclo biológico de la leishmania y transmisión. 20 AV 16. Leishmaniosis. Disponible en: http://axonveterinaria.net/web_axoncomunicacion/auxiliarveterinario/20/AV_20_16-19_Leishmaniosis_ciclo_transmision.pdf

[6] Hernández, A., Gutiérrez, J., Xiao, Y., Branscum, A. & Cuadros, D. *Spatial epidemiology of cutaneous leishmaniasis in Colombia: socioeconomic and demographic factors associated with a growing epidemic.* The royal society tropical medicine & hygiene. 2019.

[7] Russell, S. & Norving, P. *Artificial Intelligence A Modern Appoach.* Prentice Hall, Third Edition. 2010

[8] Alberto, J. *Introducción al Análisis de series temporales.* Universidad Complutense de Madrid, marzo de 2007.

[9] Brownlee, J. *Hyperparameter Optimization With Random Search and Grid Search.* Machine Learning Mastery. (2020, sep. 19). [Online]. Disponible en: https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/

[10] Dong, W., Huang, Y., Lehane, B. & Ma, G. *XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring.* ELSEVIER, 2020.

[11] Sitiobigdata. *Gentle Introduction of XGBoost Library.* (2019, ene. 19). [Online]. Disponible en: https://sitiobigdata.com/2019/01/20/gentle-introduction-of-xgboost-library/#

[12] Towards data science. *The Ultimate Guide to AdaBoost, random forests and XGBoost.* (2020, mar. 16). [Online]. Disponible en: https://towardsdatascience.com/the-ultimate-guide-to-adaboost-random-forests-and-xgboost-7f9327061c4f

[13] King, R., Campbell-Lendrum, D., & Davies, C. *Predicting Geographic Variation in Cutaneous Leishmaniasis, Colombia,* 2004.

[14] Chaves, L. & Pascual, M. *Climate Cycles and Forecasts of Cutaneous Leishmaniasis, a Nonstationary Vector-Borne Disease.* PLos Medicine, agosto de 2006.

[15] Chaves, L. *Climate and recruitment limitation of hosts: the dynamics of American cutaneous leishmaniasis seen through semi-mechanistic seasonal models. Annals of Tropical Medicine & Parasitology*, diciembre de 2008.

[16] Lewnard, J. A., Jirmanus, L., Júnior, N. N., Machado, P. R., Glesby, M. J., Ko, A. I. & Weinberger, D. M. *Forecasting Temporal Dynamics of Cutaneous Leishmaniasis in Northeast Brazil*. PLoS Neglected Tropical Diseases, 2014.

[17] Sharafi, M., Ghaem, H., Tabatabaee, H. R., & Faramarzi, H. *Forecasting the number of zoonotic cutaneous leishmaniasis cases in south of Fars province, Iran using seasonal ARIMA time series method.* Asian Pacific Journal of Tropical Medicine, diciembre de 2016.

[18] Valderrama, C., Alexander, N., Ferro, C., Cadena, H., Marín, D., Holford, T., Munstermann, L. & Ocampo, C. *Environmental Risk Factors for the Incidence of American Cutaneous Leishmaniasis in a Sub-Andean Zone of Colombia (Chaparral, Tolima). Am. J. Trop. Med. Hyg., 82(2), 2010, pp. 243–250.*

[19] Chaves, L. F., Calzada, J. E., Valderrama, A., & Saldaña, A. *Cutaneous Leishmaniasis and Sand Fly Fluctuations Are Associated with El Niño in Panamá. PLoS Neglected Tropical Diseases, octubre de 2014.*

[20] Pérez, M., Ocampo, C., Valderrama, C. & Alexander, N. *Spatial modeling of cutaneous leishmaniasis in the Andean region of Colombia.* Mem Inst Oswaldo Cruz, Rio de Janeiro, Vol. 111(7): 433-442, julio de 2016.

[21] Gutiérrez, J., Martínez, R., Ramoni, J., Diaz, F., Gutiérrez, R., Ruiz, F., Botello, H., Gil, M., González, J. & Palencia, M. *Environmental and socio-economic determinants associated with the occurrence of cutaneous leishmaniasis in the northeast of Colombia.* Trans R Soc Trop Med Hyg 2018; 00: 1–8.

[22] Yuexin, W., Nishiura, H, Yiming, Y. & M. Saitoh. *Deep Learning for Epidemiological Predictions.* Short Research Papers II. MI, USA, julio de 2018.

[23] Zhao, N., Charland, K., Carabali, M., Nsoesie, E., Maher-Giroux, M., Rees, E., Yuan, M., Garcia C., Ramirez, G. & Zinszer, K. *Machine learning and dengue forecasting: Comparing random forests and*

artificial neural 2 networks for predicting dengue burdens at the national sub-national scale in Colombia. BioRxiv, enero de 2020.

[24] Fayyad, U. & Stolorz, P. *Data mining and KDD: Promise and challenges. Future Generation Computer Systems.* ELSEVIER, pp. 99-104, 1997

[25] Shafique, U., & Qaiser, H. A *Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA).* International Journal of Innovation and Scientific Research, Vol. 12 No. 1, pp. 217-22, noviembre de 2014.

[26] Maimon, O. & Rokach, L. *Data Mining and Knowledge Discovery Handbook.* Springer-Verretraso New York, Inc., 2nd ed., febrero 2018.