

**Modelo multidimensional de aprendizaje automático para determinar un índice de vulnerabilidad de la COVID-19**



*Tesis de Trabajo de Grado*

Modalidad: Trabajo de Investigación

**Juan Sebastián Realpe González.**

Código: 100616021378

**Paula Andrea Rosero Pérez.**

Código: 100616021403

*Director: MSc. Ricardo Salazar Cabrera*

*Co-Director: Ph.D. Diego Mauricio López Gutiérrez*

Universidad del Cauca

**Facultad de Ingeniería Electrónica y Telecomunicaciones**

**Departamento de Telemática**

Línea de Investigación: e-Salud

*Popayán, febrero 2023*

*A Dios por acompañarme y estar presente a lo largo de mi vida. A mi padre, por ser mi motivación y fortaleza, por sus enseñanzas, sabiduría, apoyo y amor incondicional. A mi madre, por su amor, confianza, apoyo, enseñanzas y arduo trabajo. A mis hermanos, quienes son mi motivación diaria. A Sebastián, por acompañarme a lo largo de este proceso, por creer en mí e impulsarme a cumplir mis sueños. A Alejandra, por sus cuidados y cariño. A mis amigos, por ser parte de mi formación, por compartir sus conocimientos y hacer de este trayecto una experiencia más agradable. A mi compañero Juan Sebastián, por ser parte de cada uno de mis proyectos a lo largo de la carrera y por su ayuda incondicional. A todos ¡Muchas gracias! Este trabajo es por y para ustedes.*

*Paula Andrea Rosero Pérez*

*Quiero primeramente darle gracias a Dios por permitirme cumplir este anhelado sueño de ser un profesional, de igual manera a mi madre Ana Cecilia González que a pesar de las adversidades estuvo incondicionalmente apoyándome en todo este proceso, gracias a mis abuelos, tíos y primos por ser un pilar fundamental en mi desarrollo personal y académico. También agradecer a mi compañero de vida, mi perro Doggie, por brindarme ese apoyo emocional en todo momento, a Paula Rosero con quien inicie este viaje profesional y ahora estamos cumpliendo este sueño juntos; por último, pero no menos importante, infinitamente gracias a todos mis amigos de Universidad porque mutuamente nos brindamos ese apoyo y compañerismo para salir adelante y lograr esta meta. ¡Gracias a todos ustedes!*

*Juan Sebastián Realpe González*

# Agradecimientos

Expresamos nuestros más sinceros agradecimientos al MSc. Ricardo Salazar Cabrera, director del trabajo de grado, por su guía en todo el desarrollo del trabajo, sus correcciones, sugerencias, tiempo, dedicación y apoyo. Gracias por entender las diferentes situaciones e inconvenientes que se nos presentaron en todo el proceso, así como por las enseñanzas brindadas.

Al PhD. Diego Mauricio López Gutiérrez gracias por sus aportes, tiempo y disposición para el desarrollo del trabajo, por acompañarnos desde la elaboración de la propuesta hasta la culminación de todo este proceso.

Al Ingeniero David Restrepo, nuestros sinceros agradecimientos por su tiempo, disposición, y contribuciones brindadas a partir de su experiencia para el desarrollo del trabajo, así como en la realización del artículo. A Charic Daniel Farinango Cuervo gracias por el tiempo y sus aportes en la elaboración del artículo.

También agradecemos a la Universidad del Cauca, especialmente a todos los profesores que durante nuestro camino por el Alma Mater, aportaron en nuestra formación profesional así como en nuestro crecimiento personal.

Por último, agradecemos inmensamente a nuestros familiares, quienes nos han acompañado durante toda nuestra estadía universitaria, brindándonos su apoyo emocional como económico, siendo así partícipes de este logro.

# Abstract

COVID-19 is an infectious disease caused by a virus named SARS-CoV-2, which was first reported on December 31, 2019, upon warning of a cluster of viral pneumonia cases in Wuhan, People's Republic of China. The COVID-19 epidemic was listed by the World Health Organization (WHO) as a Public Health Emergency of International Importance (PHEIC) on January 30, 2020. In Colombia, the first case was confirmed on March 6, 2020. Subsequently, on March 11, 2020, WHO declared that COVID-19 could be characterized as a pandemic. Consequently, DANE published an index of vulnerability to COVID-19 called "Index of vulnerability by block using demographic variables and comorbidities", which allows us to know the vulnerability per city block that people have in case of being infected with COVID-19. The data used were those obtained from the 2018 "Censo Nacional de Población y Vivienda" (CNPV) and the "Registro Individual de Prestación de Servicios de Salud" (RIPS). To identify the blocks with high levels of vulnerability, a K-means cluster analysis was proposed, and the vulnerability levels were classified as: low, medium-low, medium, medium-high, and high. Unfortunately, for the calculation of the proposed index, DANE did not consider multiple risk factors (that could increase the risk of COVID-19) found in the literature (such as: environmental factors such as temperature and precipitation). This work aims to identify the most relevant risk factors that influence COVID-19 infection and to propose a vulnerability index that considers these factors.

This work proposes two vulnerability indexes. The first is an index that considers information from the CNPV 2018 (which does not include information on comorbidities due to patient privacy). The second index was developed considering that the results of the first index had low results in the selected metrics. The second index is multidimensional, since it considers the values of the index already calculated by DANE and data from other variables found as risk factors (such as: unemployment rate, gross domestic product, and mobility variables, among others). The following phases were considered for the realization of both indexes: data collection, data preparation and, modeling and evaluation considering classification algorithms (for the first index) and regression algorithms (for the second index). The results of these phases showed that when considering different types of risk factors, there is a greater correlation between the variables and the incidence of COVID-19. With the second vulnerability index calculation, it was possible to review which would be the best Machine Learning (ML) model to predict the incidence in capital cities of Colombia. The results showed that the *ExtraTreesRegressor* model obtained the best values in the metrics used; therefore, it is considered an adequate model to achieve the objective.

# Resumen

La COVID-19 es una enfermedad infecciosa causada por un virus denominado SARS-CoV-2, la cual fue reportada por primera vez el 31 de diciembre de 2019, al advertirse de un grupo de casos de neumonía vírica en Wuhan, República Popular de China. La epidemia de la COVID-19 fue catalogada por la Organización Mundial de la Salud (OMS) como una Emergencia de Salud Pública de Importancia Internacional (ESPII) el 30 de enero de 2020. En Colombia, el primer caso fue confirmado el 6 de marzo del 2020. Posteriormente, el 11 de marzo de 2020 la OMS declaró que la COVID-19 podía caracterizarse como una pandemia. En consecuencia, el DANE publicó un índice de vulnerabilidad ante la COVID-19 denominado “Índice de vulnerabilidad por manzana con el uso de variables demográficas y comorbilidades”, el cual permite conocer la vulnerabilidad por manzana que tienen las personas en caso de ser contagiadas de COVID-19.

Los datos empleados fueron los obtenidos en el Censo Nacional de Población y Vivienda 2018 (CNPV) y el Registro Individual de Prestaciones de Salud (RIPS). Para identificar las manzanas con altos niveles de vulnerabilidad fue realizado un análisis de clúster K-means, y los niveles de vulnerabilidad se clasificaron como: baja, media-baja, media, media-alta, y alta. Desafortunadamente, para el cálculo del índice propuesto, el DANE no consideró múltiples factores de riesgo (que podrían aumentar el riesgo de la COVID-19) encontrados en la literatura (como, por ejemplo: factores ambientales tales como temperatura y precipitación). El objetivo de este trabajo es identificar los factores de riesgo más relevantes en el contagio por la COVID-19, y proponer un índice de vulnerabilidad que considere dichos factores.

Este trabajo propone dos índices de vulnerabilidad, el primero es un índice que considera la información del CNPV 2018 (la cual no incluye información de comorbilidades por privacidad de los pacientes). El segundo índice se desarrolló considerando que los resultados del primer índice tuvieron resultados bajos en las métricas seleccionadas. El segundo índice es de tipo multidimensional, ya que considera los valores del índice ya calculado por el DANE y datos de otras variables encontradas como factores de riesgo (como por ejemplo: tasa de desempleo, PIB, variables de movilidad, entre otros). Para la realización de ambos índices se consideraron las siguientes fases: recolección de los datos, preparación de los datos y, modelado y evaluación considerando algoritmos de clasificación (para el primer índice) y de regresión (para el segundo índice). Los resultados de las fases realizadas permitieron observar que al considerar diferentes tipos de factores de riesgo se presenta una mayor correlación entre las variables y la incidencia al COVID-19. Con el cálculo del segundo índice de vulnerabilidad, se logró revisar cuál sería el mejor modelo de *Machine Learning (ML)* para predecir la incidencia en las ciudades capitales de Colombia. Los resultados demostraron

que el modelo *ExtraTreesRegressor* obtuvo los mejores valores en las métricas utilizadas, por lo que, se considera un modelo adecuado para lograr el objetivo.

## Tabla de contenido

1. Introducción .....	1
1.1. Planteamiento del problema .....	1
1.2. Pregunta de investigación e hipótesis .....	6
1.3. Motivación .....	6
1.4. Objetivos .....	7
1.4.1. Objetivo general .....	7
1.4.2. Objetivos específicos .....	7
1.5. Metodología.....	8
1.6. Contenido de la monografía.....	9
2. Identificación de variables .....	10
2.1. Revisión de literatura.....	10
2.2. Índices de vulnerabilidad .....	13
2.3. Resultados capítulo 2 .....	15
3. Índice de vulnerabilidad inicial .....	18
3.1. Comprensión del negocio .....	18
3.1.1. Objetivos del negocio.....	18
3.1.2. Evaluación de la situación .....	19
3.1.3. Determinar los objetivos de la minería de datos .....	21
3.1.4. Plan del proyecto .....	21
3.2. Comprensión de los datos .....	22
3.2.1. Información del DANE .....	22
3.2.2. Datos del índice de vulnerabilidad del DANE .....	23
3.3. Preparación de los datos .....	24
3.3.1. Limpieza y pre-procesamiento de datos .....	24
3.3.2. Integración de los <i>datasets</i> .....	30
3.4. Análisis Exploratorio de Datos (EDA).....	30
3.5. Modelado y evaluación.....	37
3.5.1. Modelado.....	37
3.5.2. Evaluación.....	39
4. Índice de vulnerabilidad multidimensional .....	43
4.1. Comprensión del negocio .....	43
4.1.1. Objetivos del negocio.....	43
4.1.2. Evaluación de la situación .....	44
4.1.3. Determinar los objetivos de la minería de datos .....	45
4.1.4. Plan del proyecto .....	45

4.2.	Comprensión de los datos .....	45
4.2.1.	Lectura de datos .....	45
4.2.2.	Información de fuentes adicionales .....	46
4.3.	Preparación de los datos .....	49
4.3.1.	Limpieza y pre-procesamiento de datos .....	49
4.3.2.	Integración de los <i>datasets</i> .....	52
4.4.	Análisis Exploratorio de Datos (EDA).....	52
4.5.	Modelado y evaluación.....	63
4.5.1.	Modelado.....	63
4.5.2.	Evaluación.....	67
4.6.	Optimización del modelo.....	78
4.6.1.	<i>Decision Tree Regressor</i> .....	78
4.6.2.	<i>Random Forest Regressor</i> .....	83
4.6.3.	<i>Gradient Boosting Regressor</i> .....	86
4.6.4.	<i>Extra Trees Regressor</i> .....	91
4.6.5.	<i>AdaBoost Regressor</i> .....	93
5.	Análisis de resultados .....	98
6.	Conclusiones .....	104
	Referencias.....	107
	Anexo A. Revisión índices de vulnerabilidad.....	114
	Anexo B. Artículo generado con la revisión de literatura .....	115
	Anexo C. Repositorio de Github.....	116
	Anexo D. <i>Dataset</i> inicial.....	117
	Anexo E. <i>Dataset</i> del índice multidimensional.....	118
	Anexo F. Artículo para la revista JPM de MDPI.....	119



# Lista de Figuras

<b>Figura 1.</b> Gráfico de burbujas en el que se mapea y asocia el tipo de investigación con el contexto de la misma. Los porcentajes se calculan por cada eje. ....	12
<b>Figura 2.</b> Plan del proyecto.....	21
<b>Figura 3.</b> Densidad de los datos para cada categoría de la variable objetivo.....	31
<b>Figura 4.</b> Gráfico de cajas para los departamentos. ....	32
<b>Figura 5.</b> Gráfico de dispersión para CASA. ....	33
<b>Figura 6.</b> Gráfico de dispersión para Per_edad_20 a 24.....	33
<b>Figura 7.</b> Gráfico de dispersión para Indígena.....	34
<b>Figura 8.</b> Valores de correlación de Pearson entre las variables independientes y variable dependiente. ....	35
<b>Figura 9.</b> Valores de correlación de Spearman entre las variables independientes y variable dependiente.....	36
<b>Figura 10.</b> Densidad de los datos para la variable objetivo incidencia. ....	53
<b>Figura 11.</b> Gráfico de cajas para las 24 ciudades principales. ....	54
<b>Figura 12.</b> Gráfico de cajas para los la variable año respecto a la incidencia. ....	55
<b>Figura 13.</b> Gráfico de cajas para la variable valor de vulnerabilidad respecto a la incidencia.....	55
<b>Figura 14.</b> Gráfico de dispersión para trimestre.....	56
<b>Figura 15.</b> Gráfico de dispersión para porcentaje de desempleo.....	56
<b>Figura 16.</b> Gráfico de dispersión para temperatura.....	57
<b>Figura 17.</b> Gráfico de dispersión para precipitación.....	57
<b>Figura 18.</b> Gráfico de dispersión para PIB.....	57
<b>Figura 19.</b> Gráfico de dispersión para retail_and_recreation. ....	58
<b>Figura 20.</b> Gráfico de dispersión para grocery_and_pharmacy.....	58
<b>Figura 21.</b> Gráfico de dispersión para parks.....	58
<b>Figura 22.</b> Gráfico de dispersión para transit_stations.....	59
<b>Figura 23.</b> Gráfico de dispersión para workplaces.....	59
<b>Figura 24.</b> Gráfico de dispersión para residential.....	59
<b>Figura 25.</b> Gráfico de dispersión para porcentaje de vacunación. ....	60
<b>Figura 26.</b> Gráfico de dispersión para Vulnerabilidad_numero. ....	60
<b>Figura 27.</b> Matriz de confusión (Correlación de Pearson).....	61
<b>Figura 28.</b> Matriz de confusión (Correlación de Spearman).....	62
<b>Figura 29.</b> Árbol de decisión obtenido (Profundidad=7).....	64
<b>Figura 30.</b> Árbol de decisión obtenido – Parte 1.....	64
<b>Figura 31.</b> Árbol de decisión obtenido – Parte 2.....	64
<b>Figura 32.</b> Árbol de decisión obtenido – Parte 3.....	65
<b>Figura 33.</b> Árbol de decisión obtenido – Parte 4.....	65
<b>Figura 34.</b> Árbol de decisión obtenido – Parte 5.....	65
<b>Figura 35.</b> Árbol de decisión obtenido – Parte 6.....	66
<b>Figura 36.</b> Total impureza vs valor efectivo de alpha.....	79
<b>Figura 37.</b> Número de nodos vs alpha, y profundidad del árbol vs alpha.....	79
<b>Figura 38.</b> Puntaje vs alpha.....	80
<b>Figura 39.</b> Error de validación cruzada vs ccp_alpha.....	81
<b>Figura 40.</b> Árbol podado con el mejor valor de ccp_alpha.....	81

<b>Figura 41.</b> Mejores hiper-parámetros obtenidos para Decision Tree Regressor. ....	82
<b>Figura 42.</b> Modelo optimizado para Decision Tree Regressor. ....	82
<b>Figura 43.</b> Resultados de Decision Tree Regressor con GridSearchCV. ....	82
<b>Figura 44.</b> Valor óptimo mínimo para n_estimators en Random Forest. ....	83
<b>Figura 45.</b> Valor óptimo máximo para n_estimators en Random Forest. ....	84
<b>Figura 46.</b> Valor óptimo mínimo para max_features en Random Forest. ....	84
<b>Figura 47.</b> Valor óptimo máximo para max_features en Random Forest. ....	85
<b>Figura 48.</b> Mejores hiper-parámetros para Random Forest según out-of-bag. ....	85
<b>Figura 49.</b> Mejores hiper-parámetros para Random Forest implementando GridSearchCV. ....	86
<b>Figura 50.</b> Modelo optimizado para Random Forest Regressor. ....	86
<b>Figura 51.</b> Evolución del cv_error respecto al número de árboles para Gradient Boosting Regressor. ....	87
<b>Figura 52.</b> Evolución del train error y el cv-error respecto al número de árboles para 3 medidas de learning rate para Gradient Boosting Regressor. ....	88
<b>Figura 53.</b> Evolución del cv-error respecto a la profundidad de los árboles. ....	88
<b>Figura 54.</b> Mejores hiper-parámetros obtenidos con GridSearchCV para Gradient Boosting Regressor. ....	89
<b>Figura 55.</b> Modelo optimizado para Gradient Boosting Regressor. ....	89
<b>Figura 56.</b> Mejores hiper-parámetros obtenidos con Hist Gradient Boosting. ....	90
<b>Figura 57.</b> Modelo optimizado para Hist Gradient Boosting. ....	90
<b>Figura 58.</b> Evolución del cv_error respecto al número de árboles para Extra Trees Regressor. ....	91
<b>Figura 59.</b> Evolución del cv-error respecto a la profundidad de los árboles para Extra Trees Regressor. ....	92
<b>Figura 60.</b> Mejores hiper-parámetros obtenidos con GridSearchCV para Extra Trees Regressor. ....	92
<b>Figura 61.</b> Modelo optimizado para Extra Trees Regressor. ....	92
<b>Figura 62.</b> Modelo implementado para AdaBoost Regressor. ....	93
<b>Figura 63.</b> Mejores hiper-parámetros obtenidos con GridSearchCV para AdaBoost Regressor. ....	93
<b>Figura 64.</b> Modelo optimizado implementado para AdaBoost Regressor. ....	94
<b>Figura 65.</b> Tabla valores predichos vs valores reales – parte 1. ....	96
<b>Figura 66.</b> Tabla valores predichos vs valores reales – parte 2. ....	96
<b>Figura 67.</b> Gráfica de valores predichos vs valores reales. ....	96
<b>Figura 68.</b> Tabla de importancia para el algoritmo Decision Tree Regressor. ....	101
<b>Figura 69.</b> Tabla de importancia para el algoritmo <i>Random Forest Regressor</i> . ....	101
<b>Figura 70.</b> Tabla de importancia para el algoritmo <i>Extra Trees Regressor</i> . ....	101
<b>Figura 71.</b> Tabla de importancia para el algoritmo <i>Gradient Boosting Regressor</i> . ....	101
<b>Figura 72.</b> Tabla de importancia para el algoritmo <i>AdaBoosting Regressor</i> . ....	102

# Lista de Tablas

<b>Tabla 1.</b> Factores de riesgo identificados de acuerdo al tipo de factor de riesgo .....	15
<b>Tabla 2.</b> Resultados obtenidos en la evaluación 80%-20% del índice inicial.....	40
<b>Tabla 3.</b> Resultados obtenidos en la evaluación 70%-30% para el índice inicial.....	41
<b>Tabla 4.</b> Resultados obtenidos en la evaluación 80%-20% para el índice multidimensional .....	68
<b>Tabla 5.</b> Resultados obtenidos en la evaluación 70%-30% para el modelo multidimensional .....	70
<b>Tabla 6.</b> Resultados obtenidos en la evaluación 80%-20% para el índice del DANE. ....	73
<b>Tabla 7.</b> Resultados obtenidos en la evaluación 70%-30% para el índice del DANE. ....	75
<b>Tabla 8.</b> Resultados obtenidos en la optimización de los hiper-parámetros .....	95

# Lista de acrónimos

<b>CDC</b>	Centers for Disease Control and Prevention
<b>CNPV</b>	Censo Nacional de Población y Vivienda
<b>CO</b>	Carbon monoxide
<b>COICA</b>	Coordinadora de Organizaciones Indígenas de la Cuenca del Río Amazonas
<b>CRISP-DM</b>	Cross-Industry Standard Process for Data Mining
<b>CSV</b>	Comma Separated Values
<b>CV</b>	Cross Validation
<b>DANE</b>	Departamento Administrativo Nacional de Estadística
<b>DIVIPOLA</b>	División Político Administrativa
<b>DMQ</b>	Distrito Metropolitano de Quito
<b>DNP</b>	Departamento Nacional de Planeación
<b>ESPII</b>	Emergencia de Salud Pública de Importancia Internacional
<b>GEIH</b>	Gran Encuesta Integrada de Hogares
<b>GNB</b>	Gaussian Naive Bayes
<b>IETS</b>	Instituto de Evaluación Tecnológica en Salud
<b>INS</b>	Instituto Nacional de Salud
<b>KNN</b>	k-nearest neighbors
<b>LDA</b>	Linear Discriminant Analysis
<b>ML</b>	Machine Learning
<b>NO</b>	Nitric oxide
<b>NO<sub>2</sub></b>	Nitrogen Dioxide
<b>O<sub>3</sub></b>	Ozono
<b>OMS</b>	Organización Mundial de la Salud
<b>OPS</b>	Organización Panamericana de la Salud
<b>PIB</b>	Producto Interno Bruto
<b>PDF</b>	Portable Document Format
<b>PM<sub>2.5</sub></b>	Particulate Matter
<b>PMBOK</b>	Project Management Body of Knowledge
<b>PMI</b>	Project Management Institute
<b>QDA</b>	Quadratic Discriminant Analysis
<b>RIPS</b>	Registros Individuales de Prestación de Servicios de Salud
<b>RMSE</b>	Root Mean Square Error
<b>SAPSC</b>	San Andrés, Providencia y Santa Catalina
<b>SVM</b>	Support Vector Machine
<b>UCI</b>	Unidad de Cuidados Intensivos
<b>WP</b>	Work Page

# Capítulo 1.

## 1. Introducción

### 1.1. Planteamiento del problema

La COVID-19 es una enfermedad causada por un virus denominado SARS-CoV-2, reportada por primera vez el 31 de diciembre de 2019, al advertirse de un grupo de casos de neumonía vírica en Wuhan, República Popular de China [1]. La epidemia del COVID-19 fue catalogada por la Organización Mundial de la Salud (OMS) como una Emergencia de Salud Pública de Importancia Internacional (ESPII) el 30 de enero de 2020 [2]. En Colombia, el primer caso fue confirmado el 6 de marzo del 2020 [3]. Posteriormente, el 11 de marzo de 2020 la OMS declaró que la COVID-19 podía caracterizarse como una pandemia [4]; es decir, una enfermedad epidémica que ha sido propagada por diversos países del mundo de forma simultánea [5].

La mayoría de personas que han sido infectadas por el virus, presentan cuadros respiratorios que pueden ser leves, moderados o graves. Así mismo, las personas mayores y con afecciones médicas subyacentes, como diabetes, hipertensión, obesidad, enfermedades cardiovasculares, cáncer o enfermedades crónicas tienen más probabilidades de presentar cuadros graves [6].

Las tasas de pruebas confirmadas por la COVID-19 brindan información fundamental para entender el impacto total de la pandemia e identificar maneras de disminuir la morbilidad y mortalidad. A su vez, las hospitalizaciones han sido afectadas por factores socioeconómicos como fondos insuficientes para la compra de medicamentos o conocimientos limitados acerca de la COVID-19. Estos factores impiden detectar su prevalencia, especialmente en países de ingresos bajos/medios donde muchas personas se quedan en casa, sin diagnóstico, ni pruebas. Las tasas de hospitalización ofrecen un medio indirecto para seguir los brotes locales y son un indicador de la mortalidad debida a la COVID-19 [7].

En Colombia hasta el 5 de noviembre de 2021 se habían registrado 5.009.007 casos positivos confirmados de COVID-19, los cuales representan el 9,72% de la población

total, de los cuales la mayoría de casos confirmados se encuentran en personas con un rango de edad entre 30-39 años con un total de 1.122.354 (22,40% del total de casos confirmados). Bogotá D.C es la ciudad con mayor número de casos positivos confirmados con un total 1.455.736, con un 29,06% del total de casos [3].

En consecuencia, el gobierno Nacional ha desarrollado un portal web en el cual es posible visualizar la situación del país, teniendo en cuenta variables como los casos confirmados, activos, recuperados y fallecidos, distribución de casos por departamento, municipio, sexo, rango de edad, etnia y comorbilidades [8].

Debido al alto grado de contagios de la COVID-19, se busca conocer la exposición que tiene la población al virus que produce la enfermedad, para ello se usa el índice de vulnerabilidad, el cual mide la exposición de una población a un riesgo específico. La vulnerabilidad es una condición dinámica, que está en función de interacciones de una variedad de factores de riesgo socioeconómicos, ambientales, personales, entre otros. La vulnerabilidad ayuda a comprender crisis humanitarias, no como fenómenos inevitables, sino como resultado de causas estructurales que pueden ser modificables por la acción humana [9]. El índice es un agregado de indicadores, ya sean cuantitativos o cualitativos, que al ser multidimensional permite medir la correlación entre variables que representan los factores o dimensiones más relevantes de un concepto [10]. Adicional a ello, un grupo de expertos mediante un estudio definieron la vulnerabilidad como “un estado dinámico que refleja con convergencia de efectos de un conjunto de factores personales y ambientales que interactúan y se amplifican entre sí, los cuales en conjunto incrementan la susceptibilidad de un individuo a enfermarse dificultando el proceso de recuperación una vez que la enfermedad ocurre” [11].

El Departamento Administrativo Nacional de Estadística (DANE), es la entidad responsable de la planeación, levantamiento, procesamiento, análisis y difusión de las estadísticas oficiales de Colombia, la cual de la mano con el Departamento Nacional de Planeación (DNP) y el Instituto de Evaluación Tecnológica en Salud (IETS), están colaborando con información estadística de la COVID-19 al Ministerio de Salud y Protección Social y al Instituto Nacional de Salud (INS). Lo anterior, con el objetivo de construir herramientas que faciliten la toma de decisiones, con mayor certeza, al alto gobierno en cuanto a la emergencia sanitaria ocasionada por la

COVID-19. El DANE publicó un índice de vulnerabilidad ante la COVID-19 denominado “Índice de vulnerabilidad por manzana con el uso de variables demográficas y comorbilidades”. Los datos fueron obtenidos en el Censo Nacional de Población y Vivienda 2018 (CNPV) y el Registro Individual de Prestaciones de Salud (RIPS). Las patologías identificadas como factores de riesgo a la COVID-19 fueron hipertensión, diabetes, cardiopatía isquémica, pulmonar crónica y cáncer. Respecto a las variables sociodemográficas fueron consideradas el hogar donde viven los individuos mayores de 60 años, si es unipersonal o familiar, la densidad poblacional y ciertas cabeceras municipales con cierto rango de habitantes. El nivel de granularidad del análisis fue por manzanas, las cuales en total fueron 407.277. Para la identificación de vulnerabilidad en las manzanas fue planteada la realización de un análisis de clúster K-means, y los niveles de vulnerabilidad se clasificaron como: baja, media-baja, media, media-alta, alta [12-13].

La naturaleza infecciosa y la velocidad de propagación de la COVID-19 ha dado paso a la investigación de factores que puedan potenciar o disminuir su transmisión. Diversos estudios han analizado el impacto de algunos parámetros meteorológicos y de calidad del medio ambiente considerando la ubicación geográfica de cada país, con el objetivo de evidenciar la posible relación entre factores ambientales y la morbilidad y mortalidad por COVID-19. Estos estudios han establecido un vínculo entre la infección por SARS-CoV-2 y diversas variables meteorológicas como temperatura, precipitación, niveles de materia particulada de 2.5 micrómetros de diámetro (PM<sub>2.5</sub>), Ozono (O<sub>3</sub>), óxido nítrico (NO), dióxido de nitrógeno (NO<sub>2</sub>), monóxido de carbono (CO), radiación solar, velocidad del viento, disposición de aguas residuales, y eliminación del hábitat de animales silvestres [14]. A nivel mundial, existe gran cantidad de evidencia al respecto, sin embargo, a nivel nacional es muy escasa. Además, es importante recalcar que el índice de vulnerabilidad realizado por el DANE no consideró este tipo de variables ambientales, las cuales podrían influir en el cálculo realizado y permitirían que se obtengan resultados diferentes los cuales tienen en cuenta diversos tipos de factores de riesgo.

La Organización Panamericana de la Salud (OPS), junto con la Coordinadora de Organizaciones Indígenas de la Cuenca del Río Amazonas (COICA) quien trabaja con organizaciones indígenas de Perú, Bolivia, Brasil, Colombia y Ecuador, han

pedido a los gobiernos nacionales que se garantice a las comunidades indígenas la atención de salud que se requiere en la lucha contra la pandemia ocasionada por la COVID-19, debido a que el acceso a los servicios de salud es un desafío para estos pueblos, ya que su gran mayoría viven en asentamientos remotos a áreas urbanas. Para ciertos grupos étnicos la afectación de la COVID-19 puede llegar a ser más alta, lo cual es importante considerarlo en la determinación de un índice de vulnerabilidad de esta enfermedad [15].

Algunos estudios han evidenciado que las vacunas contra la COVID-19 proporcionan una protección eficaz contra las comorbilidades asociadas a esta enfermedad, incluso la mortalidad [16]. El viceministro de Salud Pública y Prestación de Servicio de Colombia, Luis Alexander Moscoso (quien estuvo en el cargo desde marzo de 2020 hasta Diciembre de 2021), resaltó que la vacunación logró reducir de manera sustancial los ingresos a las Unidades de Cuidados Intensivos (UCI) y la mortalidad en la población con mayor riesgo y consiguó la transmisión del contagio [17]. Desde que comenzó la inmunización a nivel mundial, muchos países evaluaron las tasas de hospitalización y muerte por COVID-19 entre personas vacunadas y no vacunadas, con el fin de calcular la efectividad de los esquemas de vacunación. En Colombia, la Dirección de Epidemiología y Demografía y la Dirección de Medicamentos y Tecnologías del Ministerio de Salud y Protección Social actualizan constantemente el estudio *“Efectividad de las vacunas contra el covid-19 -Cohorte Esperanza”*, el cual muestra que el riesgo de hospitalización y muerte se incrementan con la edad, aumentando el riesgo para las personas no vacunadas, y puede variar dependiendo del momento de pico (situación epidemiológica) y el tiempo que haya transcurrido desde la última aplicación de la vacuna, especialmente cuando no se ha recibido un refuerzo [18]. En ese sentido, los datos de vacunación pueden ser una fuente útil para determinar un índice de vulnerabilidad, sin embargo, el DANE no consideró esa información, debido a que cuando realizaron su índice de vulnerabilidad no había suficientes datos de vacunación en Colombia, por lo tanto, no era claro establecer la influencia en el cálculo del índice.

En 2020, a raíz de la crisis sanitaria, América Latina y el Caribe experimentaron el peor desempeño económico de los últimos 120 años. La región presentaba problemas de bajo crecimiento incluso desde antes de la pandemia, se estimaba un



crecimiento del 5,9% para 2021, pero no era suficiente para recuperar los niveles del Producto Interno Bruto (PIB) de 2019. En 2020 la pobreza habría crecido en 22 millones de personas respecto al año anterior. La pérdida de ingresos en el trabajo a causa del desempleo ocasionado por la COVID-19 aumentando también las tasas de pobreza y desigualdad de ingresos en personas en condiciones de vulnerabilidad como trabajadores informales, mujeres y jóvenes indígenas, afrodescendientes, personas con discapacidad. La inequidad y heterogeneidad refleja un aumento en el porcentaje de servicios esenciales de salud interrumpidos a medida que disminuye el nivel de ingreso de los países. Las personas en condiciones de vulnerabilidad socioeconómica presentan mayor riesgo de contagio y muerte por COVID-19, debido que las desigualdades están directamente relacionadas con la capacidad de protección al contagio. También una incidencia más alta de comorbilidades se asocia a una mayor gravedad de la enfermedad e incluso con la muerte. Existe una correlación entre el exceso de mortalidad por COVID-19 y la vulnerabilidad socioeconómica, donde por ejemplo en la ciudad de Sao Paulo, se observó que las áreas de bajos ingresos fueron las más afectadas, y la mayoría de las muertes ocurrieron en hospitales públicos y entre población afrodescendiente, asiática e indígena, el porcentaje de defunciones por COVID-19 de personas en lista de espera por camas UCI fue de 19,1 % en hospitales públicos comparado con el 1% correspondiente a hospitales privados. El porcentaje de población en hacinamiento juega también un papel importante ya que áreas con mayor proporción de personas hacinadas se ven mayormente afectadas por la COVID-19 [19].

Variables como factores meteorológicos y ambientales, el pertenecer a una comunidad indígena, los ingresos económicos, entre otros; son características que no se tuvieron en cuenta en la creación del índice de vulnerabilidad por parte del DANE, debido a que existen gran cantidad de variables que pueden generar mayor fiabilidad en el desarrollo del mismo y su inclusión en un nuevo índice, podría influir en su nivel de certeza.

## **1.2. Pregunta de investigación e hipótesis**

Teniendo en cuenta la problemática mencionada anteriormente y la literatura analizada en el estado del arte, surgió la siguiente pregunta de investigación **¿Cómo determinar un índice de vulnerabilidad de la COVID-19 para Colombia, que considere los datos de casos de COVID-19 publicados diariamente por el Instituto Nacional de la Salud y factores de riesgo humanos, socioeconómicos, sociodemográficos y ambientales, adicionales a los propuestos por el DANE?**

La hipótesis del trabajo planteada sostenía que, mediante la implementación de un modelo de aprendizaje automático, se puede determinar un índice de vulnerabilidad a la COVID-19 en ciudades capitales de Colombia, teniendo en cuenta la base de datos de casos históricos de la COVID-19 y los factores de riesgo más relevantes, con el fin de ayudar en la toma de decisiones en los programas de salud pública. Las fuentes de datos adicionales a las del índice de vulnerabilidad realizado por el DANE que se tuvieron en cuenta fueron: los datos de COVID del Ministerio de salud como fuente principal, información de desempleo y PIB del DANE, datos de vacunación del ministerio de salud, información de movilidad, climatológica y espacial de satélites.

## **1.3. Motivación**

La COVID-19 ha tomado alta relevancia a nivel mundial, debido a que ha afectado a gran parte de la población. Es por ello que se decidió realizar un estudio que considerara diferentes tipos de factores, con el objetivo de calcular la vulnerabilidad de las personas en el contagio de la COVID-19. Actualmente, en Colombia existe un índice de vulnerabilidad desarrollado por el DANE, el cual no considera factores de riesgo de varios tipos, excluyendo el impacto que pueden generar otro tipo de variables encontradas en la literatura como factores de riesgo. Adicional a ello, en la revisión de la literatura se encontraron diversos índices de vulnerabilidad que no tienen en cuenta todos los tipos de factores de riesgo, considerando en los cálculos realizados solo algunas variables y exceptuando información que puede aportar al valor obtenido. Estos índices se pueden revisar en detalle en el Anexo A denominado “Revisión de índices de vulnerabilidad”.

Este trabajo consideró factores de riesgo socioeconómicos, sociodemográficos, ambientales y humanos para generar un índice multidimensional que abarcara mayor cantidad de variables a las comprendidas en el estudio que se ha realizado en Colombia, y así evaluar si se obtenía un índice con mejores métricas que pudiera apoyar a la toma de decisiones de los organismos de salud permitiendo identificar la vulnerabilidad al COVID-19 en las principales ciudades del país. Lo anterior requirió la evaluación de varios algoritmos de aprendizaje de máquina (en inglés, *Machine Learning, ML*), con el propósito de identificar cuál se comportaba de mejor manera para el índice propuesto. Esta es otra diferencia relevante respecto a otros índices propuestos previamente en trabajos similares, los cuales presentan resultados de solamente un algoritmo para el cálculo del índice, como es el caso del índice del DANE, o realizan los cálculos mediante la determinación y aplicación de ecuaciones.

## **1.4. Objetivos**

### **1.4.1. Objetivo general**

Proponer un modelo de aprendizaje automático, que permita calcular un índice de vulnerabilidad de la COVID-19 que considere de manera integrada factores de riesgo humanos, ambientales, sociodemográficos y socioeconómicos.

### **1.4.2. Objetivos específicos**

- Caracterizar los factores de riesgo humanos, ambientales, sociodemográficos y socioeconómicos que explican la vulnerabilidad al contagio por COVID-19.
- Construir un modelo de aprendizaje automático que permita calcular un índice de vulnerabilidad de la COVID-19, con las variables que se determinen como las más relevantes.
- Evaluar el modelo creado para calcular un índice de vulnerabilidad de la COVID-19 obtenido, realizando una comparación entre el índice calculado y un índice de referencia.

## 1.5. Metodología

El desarrollo de los objetivos propuestos fue realizado a través de una serie de actividades, según las directrices detalladas en el Cuerpo de Conocimientos de Gestión de Proyectos (PMBOK), instrumento desarrollado por el Instituto de Gestión de Proyectos (Project Management Institute o PMI) [20]. El proyecto fue segmentado, inicialmente, en los siguientes 5 paquetes de trabajo (WP, por su nombre en inglés Work Package, los cuales están relacionados con las fases de la metodología para el proceso de análisis de datos (*Cross-Industry Standard Process for Data Mining*, CRISP-DM).

- WP1 Identificación y selección de variables: Este WP realiza una revisión de la literatura, se entiende el problema y se establecen los pasos a seguir para proponer una solución. Para este proyecto, la fase consiste en revisar la literatura, haciendo énfasis en factores de riesgo de COVID-19, así como fuentes de datos abiertas de COVID-19 y métodos que permitan plantear un índice de vulnerabilidad de este, además, se escogen las variables más importantes para el desarrollo de este proyecto.
- WP2 Recolección de datos: En este WP se hace una recolección de las diferentes fuentes de datos, se analiza el contenido de cada una y se prevé la posibilidad de uso en las siguientes fases. Para el presente proyecto, como resultado del WP se analizan los datos encontrados con el propósito de prepararlos y generar un *dataset* que se ajuste al objetivo del proyecto. Las fuentes de datos adicionales a las del DANE a tener en cuenta son: los datos de COVID del ministerio de salud como fuente principal, información de desempleo y PIB del DANE, datos de vacunación del ministerio de salud, información de movilidad, climatológica y espacial de satélites.
- WP3 Preparación de datos: Al manejar datos de distintas fuentes, cada tipo de dato tiene una resolución espacial y temporal, además de su estructura, por lo que, es necesario hacer una limpieza de datos y un pre-procesamiento para generar el *dataset* final.

- WP4 Creación y evaluación del modelo multidimensional: En este WP se toma el *dataset* creado en el WP3, con el propósito de construir el modelo para así poder calcular el índice de vulnerabilidad de COVID-19. El modelado se hace a nivel de ciudades capitales debido a las limitaciones de acceso a datos con un alcance menor (manzanas). Además, se verifica el correcto funcionamiento del modelo que calcula el índice de vulnerabilidad. Para la validación del modelo se espera determinar (durante el desarrollo del paquete de este paquete de trabajo) un índice de referencia (a nivel nacional o internacional) para compararlo con el índice que sea calculado con el modelo.
- WP5 Preparación de documentación y otros entregables: En este WP se realiza la preparación de los capítulos de la monografía, así como la presentación de los resultados.

Durante el desarrollo del proyecto, se identificó relevante realizar un índice de vulnerabilidad inicial, el cual tendría en cuenta la mayor parte de las características utilizadas por el índice realizado por DANE, las cuales fuera posible obtener, y otras variables adicionales brindadas por el CNPV 2018. Una vez analizados los resultados de dicho índice inicial, se determinaría la necesidad de generar un índice multidimensional que tuviese en cuenta una mayor cantidad de posibles factores, lo cual sugiere la obtención de otras fuentes diferentes a las del DANE.

Considerando lo anterior, se determinó realizar los 5 paquetes de trabajo mencionados previamente, para cada uno de los dos índices a desarrollar en el proyecto.

## **1.6. Contenido de la monografía**

La redacción de la monografía está basada en las directrices del PMBOK, así como en los objetivos específicos del proyecto. En el capítulo 2 se presentan los resultados de la identificación de variables generales del proyecto. En el capítulo 3 se reporta el proceso realizado con el índice de vulnerabilidad inicial. En el capítulo 4, se presenta el desarrollo del índice multidimensional. Finalmente, en el capítulo 5 están las conclusiones y trabajo futuro de este trabajo de grado.

# Capítulo 2.

## 2. Identificación de variables

Este proceso incluyó inicialmente una revisión de literatura y posteriormente una revisión de índices de vulnerabilidad (a nivel nacional e internacional).

### 2.1. Revisión de literatura

La revisión de literatura se realizó en tres etapas: un mapeo sistemático, un análisis de resultados, y la identificación de factores relevantes, etapas que se encuentran consignadas en el Anexo B denominado “Artículo generado con la revisión de literatura”. El artículo se tituló “*Risk Factors for COVID-19: A Systematic Mapping Study*”, el cual es un artículo de revisión, que presenta en detalle el proceso realizado para la identificación de variables relevantes.

A continuación, se presenta la información más relevante de las tres etapas de revisión de literatura.

El mapeo sistemático fue realizado de acuerdo con el documento “*Guidelines for conducting systematic mapping studies in software engineering: An update*” [21]. Para ello, se realizó la búsqueda de artículos de revisión en *Scopus*, base de datos que tiene amplia cobertura de investigación científica y que abarca gran contenido en ciencias de la salud y aprendizaje automático [22], en ella se identificaron y revisaron 1786 estudios relacionados de los cuales solo 564 cumplieron con los criterios de inclusión.

A medida que se iban revisando los documentos, se detectaron características similares entre ellos, como el tipo de revisión y los factores que abarcaban, debido a esto, fueron generadas dos clasificaciones en el mapeo. La primera clasificación se denominó “tipo de investigación” y hacía referencia al tipo de revisión que se había realizado. Hubo revisiones que realizaban sus propios estudios a las que se les llamó “Revisión y experimentación”; también estaban las revisiones que no realizaban

estudios, pero si cumplían el objetivo de una revisión a las que se les tituló “Revisión”; y por último, se encontraban revisiones que en el resumen (*abstract*) no indican si tuvieron en cuenta bases de datos ni el número de documentos examinados para llegar a los resultados y conclusiones, por lo que, se les nombró “Revisión no formal”.

La segunda clasificación se denominó “contexto de investigación” y hacía referencia a los tipos de factores encontrados, y las posibles combinaciones entre ellos. En esta clasificación se contó con las siguientes categorías:

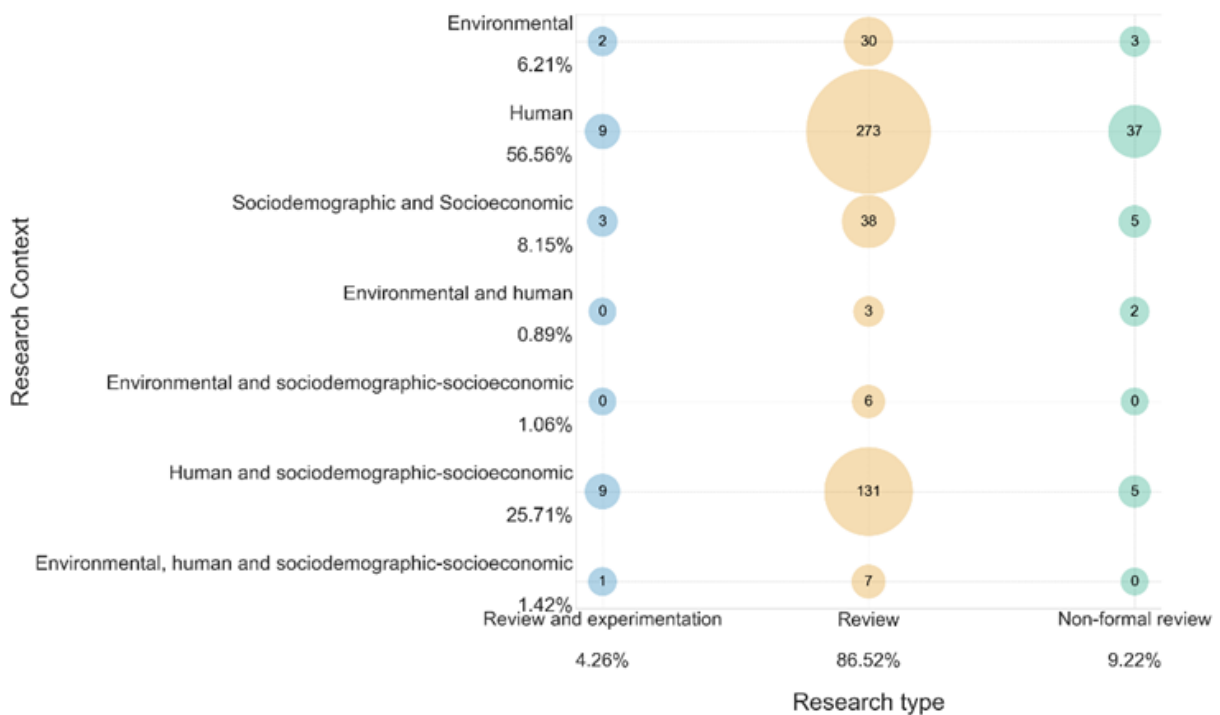
- Factores de riesgo humano: Hacen referencia a las condiciones de salud de las personas.
- Factores de riesgo sociodemográficos y socioeconómicos: Indican las características de la población.
- Factores ambientales: Estos factores incluyen variables ambientales.
- Factores de riesgo sociodemográficos y socioeconómicos, y humanos.
- Factores de riesgo sociodemográficos y socioeconómicos, y ambientales.
- Factores de riesgo sociodemográficos y socioeconómicos, ambientales y humanos.

Posteriormente, se realizó un análisis de resultados, en el cual se evidenciaron los factores de riesgo; datos principales para este estudio, los cuales fueron obtenidos mediante la revisión de cada uno de los documentos, realizada por los investigadores.

En ese sentido, el mapa de estudios obtenido en el mapeo sistemático, el cual se puede observar en la Figura 1, permitió identificar que el mayor número de documentos corresponde al tipo "revisión" y sólo aborda los factores de riesgo humanos. El segundo mayor número de documentos, siendo también de tipo "revisión", estudia la categoría de factores de riesgo sociodemográficos y socioeconómicos, y humanos. En el lado opuesto, los documentos de "revisión y experimentación" son los que menos predominan en el mapeo. Respecto a los factores de riesgo ambiental, son los que menos representación tienen en los documentos revisados. Además, entre todas las categorías de "tipo de investigación", se obtuvieron muy pocos resultados para los documentos que abordan la categoría de factores de riesgo ambientales y humanos. El tipo de

"revisión no formal" presenta baja dominancia en el mapeo sistemático, y en cuanto al contexto de la investigación, los factores de riesgo de humanos fueron los más prevalentes para esta categoría.

Por último, cabe destacar que los factores de riesgo con mayor presencia en los documentos fueron las comorbilidades como la diabetes, la hipertensión, la obesidad y las enfermedades cardiovasculares, así como factores como la edad y el sexo.



**Figura 1.** Gráfico de burbujas en el que se mapea y asocia el tipo de investigación con el contexto de la misma. Los porcentajes se calculan por cada eje.

Finalmente, se realizó una identificación de factores, es decir, después de extraer los datos de los estudios identificados, se utilizó un enfoque de clasificación, teniendo en cuenta los 3 tipos de factores de riesgo determinados en el contexto de la investigación; factores de riesgo humanos, factores de riesgo sociodemográficos y socioeconómicos, y factores de riesgo ambientales. Además, se realizó un conteo del número de veces que los factores encontrados se mencionaban en la totalidad de los estudios, asociando términos similares, para así identificar variables que influyen en el contagio por COVID-19. Como resultado de esta etapa, se obtuvieron los factores de riesgo relevantes, encontrando que, en los factores de riesgo humanos, comorbilidades como diabetes, hipertensión, obesidad y enfermedades cardiovasculares son las que mayor presencia tenían. En cuanto a los factores de riesgo ambientales, los que mayor dominancia tenían son calidad del aire y



contaminación. Respecto a los factores de riesgo sociodemográficos y socioeconómicos, la edad, el sexo y la raza son los factores predominantes. En las Tablas 3, 4 y 5 del Anexo B, se pueden observar todos los factores encontrados y la cantidad de veces que fueron mencionados en un documento como factor de riesgo.

Después de realizar este proceso, se consideró relevante revisar otros índices de vulnerabilidad existentes a nivel nacional e internacional, y analizar las variables que consideraban dichos índices; ese proceso es expuesto en la siguiente subsección.

## **2.2. Índices de vulnerabilidad**

Después de realizar el mapeo sistemático, se procedió a realizar una investigación de índices de vulnerabilidad para así, conocer qué se había realizado en otros estudios y las variables que habían sido consideradas. En esa investigación se encontraron varios índices de vulnerabilidad los cuales se detallan en el Anexo A denominado “Revisión índices de vulnerabilidad”. A continuación, se presenta la información más relevante de los índices encontrados que se consideraron más significativos.

- Índice de vulnerabilidad de México. Se calculó un índice de vulnerabilidad ante la COVID-19 para México en el año 2020, a una escala municipal que comprendía 2457 municipios, el cual tenía el objetivo de identificar la distribución de los diversos factores de riesgo que ocasionan una mayor susceptibilidad al daño o las consecuencias desfavorables que pueden tener las personas. Este índice manejó 3 dimensiones: demográfica, salud y socioeconómica. Cada una de las dimensiones constituye una serie de indicadores que están relacionados con el tipo de vulnerabilidad y exposición. El índice de vulnerabilidad maneja 4 grados de clasificación: medio, alto, muy alto y crítico.

Para obtener el valor de cada una de las dimensiones se cortaron en quintiles cada una de las variables que las componen y se asignó un valor entre 1 y 5 en cada caso (donde un valor de 1 representa el valor más bajo y 5 el más alto). Después, se procedió a calcular el promedio de los nuevos valores del conjunto de las variables de cada dimensión.

Finalmente, “el índice de vulnerabilidad es el resultado de la suma de los valores de cada dimensión, ponderadas por un factor de  $\frac{1}{3}$ ”, donde el indicador hallado es una “variable ordinal, por ende, su valor numérico no tiene una interpretación directa más allá de una relación de mayor que, menos que” [23].

- Índice de vulnerabilidad en el Distrito Metropolitano de Quito. En este trabajo se analizó la trayectoria espacial de los contagios de la COVID-19 en el Distrito Metropolitano de Quito (DMQ) durante el periodo de confinamiento social denominado “semáforo rojo” el cual tuvo lugar en Quito desde el 9 de abril al 3 de junio de 2020. Como principal objetivo estaba la identificación de las principales tendencias de expansión del contagio y la correlación con las dimensiones de la vida urbana.

El índice de vulnerabilidad desarrollado se basó en la investigación de 3 dimensiones: demográficas, socioeconómicas y del hábitat [24].

- Índice de vulnerabilidad C19VI. Este índice fue desarrollado en Estados Unidos por el Centro para el Control y la Prevención de Enfermedades (CDC por sus siglas en inglés), considerando las siguientes variables para calcular el índice: estado socioeconómico, composición del hogar y discapacidad, estado e idioma de las minorías, tipo de vivienda y transporte, factores epidemiológicos y factores del sistema de salud.

Como resultado se obtuvo un mapa de vulnerabilidad en el cual cada ciudad fue identificada con un color de acuerdo con el nivel de vulnerabilidad encontrado [25].

- Índice de vulnerabilidad de Colombia (elaborado por el DANE). El DANE realizó un índice de vulnerabilidad de COVID-19, el cual tiene un nivel de desagregación geográfica por manzanas.

El objetivo del estudio fue categorizar qué personas, según la manzana dónde viven, tienen mayor probabilidad de complicaciones en caso de contagiarse por COVID-19. Para lo anterior, se consideraron características demográficas y las condiciones de salud de las personas. A continuación, se presentan las variables contempladas para calcular el índice de vulnerabilidad del DANE:

- Comorbilidades: hipertensión, diabetes, cardiopatía isquémica, pulmonares crónicas y cáncer.

- Características demográficas: identificación de personas mayores de 60 años, hogares en hacinamiento en cuarto y dormitorios, hogares en riesgo intergeneracional alto y medio por manzana.

A partir de esas variables se llevaron a cabo una serie de pasos que permitieron consolidar una base de datos de 407.277 filas con las columnas mencionadas anteriormente. Posterior a ello, se aplicó el análisis de *cluster Kmeans*, el cual permitió agrupar las manzanas según las características demográficas y de comorbilidades.

Finalmente, el resultado que se obtuvo fue el mapa de vulnerabilidad de Colombia, en el cual se puede encontrar la vulnerabilidad al COVID-19 por manzana [12].

## 2.3. Resultados capítulo 2

Considerando el mapeo sistemático y la fase de identificación de factores realizada, a continuación, se presenta la Tabla 1, en la cual se pueden observar algunos de los factores encontrados en la literatura. Cabe resaltar, que en las Tablas 3, 4 y 5 del Anexo B se encuentran la totalidad de factores identificados.

Tabla 1. Factores de riesgo identificados de acuerdo al tipo de factor de riesgo

Factores de riesgo humanos	Número de veces*	Factores de riesgo sociodemográficos y socioeconómicos	Número de veces*	Factores de riesgo ambientales	Número de veces*
Diabetes	107	Edad	399	Ambiental	16
Obesidad	102	Sexo	81	Calidad del aire	14
Cardiovascular	90	Raza	27	Contaminación	11
Hipertensión	82	Etnia	16	Aerosoles	10
Enfermedad crónica	30	Economía	16	Temperatura	4
Cáncer	26	Trabajo	14	Humedad	3
Fumador	24	Ingresos	10		
Pulmonar	24				
Renal	24				
Mental	18				
Depresión	4				

Número de veces\*: Número de veces que fueron mencionados esos factores en los documentos.

En la Tabla 1 se pueden observar algunos de los factores de riesgo encontrados, donde los factores con mayor presencia en los documentos son los de tipo factores de riesgo humanos, en general, para este tipo de factores se pueden determinar 2 grandes grupos; factores relacionados a comorbilidades y factores relacionados a la salud mental de las personas. Por otra parte, los factores de riesgo ambientales fueron los menos estudiados y se encontraron factores como ambiental, calidad del aire y contaminación. Finalmente, respecto a los factores sociodemográficos y socioeconómicos, se identificaron factores como: edad, sexo, raza, etnia.

Respecto a la investigación de índices de vulnerabilidad, se encontró que el índice de vulnerabilidad realizado por el CDC era el más adecuado para este trabajo debido a que, considera variables a las cuales se tiene acceso en Colombia y que pertenecen a diferentes tipos de factores de riesgo.

En consecuencia, a lo mencionado anteriormente, se determinó que para este trabajo de grado se iba a desarrollar inicialmente un índice de vulnerabilidad que considerara algunas variables semejantes a las contempladas por el CDC, pero con la restricción de abordar tipos de factores de riesgo considerados por el índice del DANE. Sin embargo, es importante resaltar que tanto el CDC como el DANE tuvieron en cuenta en sus índices de vulnerabilidad factores de riesgo humanos como comorbilidades; datos que en Colombia son restringidos a causa de la confidencialidad que se debe tener con esa información por lo que, no pudieron ser incluidos en los índices realizados. Lo anterior, presenta una gran limitante dado que, las comorbilidades habían sido identificadas como factores de riesgo y podrían ser muy representativas en los índices a desarrollar.

Después de definir qué se iba a realizar un índice inicial, se encontró que los *datasets* del CNPV 2018 [26] contenían datos de variables semejantes a las consideradas en el índice del CDC; los *datasets* del CNPV 2018 que se encontraron fueron: vivienda, hogares, fallecidos, personas y marco de georreferenciación [27]. El CNPV 2018 caracterizó las personas, así como las viviendas y hogares colombianos, por ende, comprende información relevante y que se ajusta a lo propuesto en el índice.

Con base en lo anterior, para el primer índice (del cual se presenta información detallada en el capítulo 3) fueron consideradas las siguientes variables: tipo de vivienda, número de dormitorios por hogar, número de fallecidos por hogar, total de personas en el hogar, sexo de la persona fallecida, edad de la persona fallecida, sexo de la persona, edad de la persona en grupos quinquenales, reconocimiento étnico, habla la lengua nativa de su pueblo, habla otra(s) lengua(s) nativa(s), calidad de la prestación del servicio de salud, sabe leer y escribir, nivel educativo más alto alcanzado, qué hizo durante la semana pasada (variable enfocada a preguntar si la persona trabajó y recibió ingresos por el trabajo).

Posteriormente, se llevó a cabo el modelado y evaluación para comparar qué tanto se acercaba el índice de vulnerabilidad desarrollado con los datos de vulnerabilidad publicados por el DANE. Como resultado se obtuvo que el índice de vulnerabilidad inicial no se aproximaba al índice realizado por el DANE, por ende, las variables consideradas (las cuales se presentan en el capítulo 3) no eran lo suficientemente relevantes.

Debido a los resultados del primer índice, y resaltando que no se incluyeron las comorbilidades; las cuales son variables comprendidas en el índice de vulnerabilidad desarrollado por el DANE y el CDC, se decidió utilizar los valores de vulnerabilidad hallados en el índice elaborado por el DANE y añadir nuevas variables, las cuales hacían referencia a los otros tipos de factores de riesgo. Las variables incluidas fueron: temperatura, precipitación, porcentaje de vacunación de COVID-19, PIB, desempleo y datos de movilidad. Lo anterior con el fin de elaborar un índice de vulnerabilidad multidimensional, que considerara otras variables diferentes a las incluidas por el DANE y que pueden ser relevantes.

Con los resultados de este capítulo se cumple el primer objetivo específico de este trabajo: “Caracterizar los factores de riesgo humanos, ambientales, sociodemográficos y socioeconómicos que explican la vulnerabilidad al contagio por COVID-19.”

# Capítulo 3.

## 3. Índice de vulnerabilidad inicial

En este capítulo se presentan la comprensión del negocio, comprensión de los datos, preparación de los datos, y la creación y evaluación del modelo del índice de vulnerabilidad inicial realizado; las secciones de este capítulo consideraron las fases de la metodología CRISP-DM. Por otra parte, es importante mencionar que el objetivo de este capítulo es conocer si las variables consideradas; las cuales son de tipo sociodemográfico y socioeconómico, son suficientes para predecir el valor de vulnerabilidad ya calculado por el índice del DANE para los municipios de Colombia. Lo anterior, con el propósito de revisar si con un solo tipo de factores de riesgo se puede realizar una predicción óptima de la vulnerabilidad de la COVID-19.

### 3.1. Comprensión del negocio

La comprensión del negocio es la primera fase en el proceso de minería de datos, esta fase permite determinar los objetivos y requisitos del proyecto desde un punto de vista de negocio, para así posteriormente transformar los objetivos a una perspectiva técnica y desarrollar un plan de proyecto [28].

#### 3.1.1. Objetivos del negocio

En las siguientes subsecciones se presenta la determinación de los objetivos del negocio.

##### 3.1.1.1. Contexto

En referencia a la situación del negocio, en Colombia el DANE publicó un índice de vulnerabilidad ante la COVID-19 denominado “Índice de vulnerabilidad por manzana con el uso de variables demográficas y comorbilidades”. Los datos empleados fueron los obtenidos en el Censo Nacional de Población y Vivienda 2018 (CNPV) y el Registro Individual de Prestaciones de Salud (RIPS). Se identificaron diferentes comorbilidades y variables sociodemográficas como factores de riesgo. El nivel de granularidad del análisis fue por manzanas (grupo de edificaciones rodeadas por 4

calles). Se realizó un análisis de clúster K-means para determinar los niveles de vulnerabilidad clasificados como baja, media-baja, media, media-alta, alta [12-13].

#### **3.1.1.2. Objetivos del negocio**

El objetivo del negocio fue realizar predicciones de la vulnerabilidad lo más fiables posibles partiendo de información obtenida gracias a las fuentes abiertas disponibles para Colombia. De esta forma, construir herramientas que faciliten la toma de decisiones, al Gobierno Nacional en cuanto a la emergencia sanitaria ocasionada por la COVID-19.

#### **3.1.1.3. Criterios de éxito del negocio**

Como criterios de éxito desde el punto de vista del negocio se identificaron los siguientes: la posibilidad de realizar predicciones sobre los niveles de clasificación de la vulnerabilidad de las personas a contagiarse de la COVID-19; y en función de ello, que el gobierno tome medidas adecuadas para evitar su propagación.

### **3.1.2. Evaluación de la situación**

Se identificaron los *datasets* creados por el DANE a través del CNPV 2018, donde se caracterizó a las personas, así como a las viviendas, personas fallecidas, hogares colombianos y un marco de georreferenciación.

#### **3.1.2.1. Inventario de recursos**

En cuanto a recursos software, para el desarrollo se usó *Jupyter Notebook*, el cual es un entorno de desarrollo interactivo que muestra la ejecución de código mediante un navegador web, permite visualizar gráficos, fórmulas y escribir comandos [29]. Soporta distintos lenguajes de programación, en este caso se usó *Python*. *Python* es un lenguaje de programación de alto nivel, *open source*, orientado a objetos, el cual cuenta con una semántica dinámica integrada y es usado especialmente para el desarrollo de aplicaciones web y aplicaciones informáticas [30]. *Python* es un lenguaje de propósito general que permite estructurar, limpiar, manipular, buscar, analizar y visualizar datos gracias a librerías como *Pandas* y *Scikit-learn*. *Pandas* es una herramienta que permite llevar a cabo todo el flujo de trabajo en cuanto a análisis de datos como distribución de variables y modelos de regresión [31]. *Scikit-Learn* es una librería gratuita de *Python*, cuenta con distintos algoritmos de clasificación, regresión,

*clustering* y reducción de dimensionalidad; la gran variedad de algoritmos convierte a *Scikit-learn* en una herramienta básica para programar y estructurar datos en los cuales se pretenda realizar sistemas de análisis de datos y modelado estadístico [32]. En cuanto a recursos hardware, se dispuso de dos computadores portátiles con las siguientes características:

- Computador portátil 1
  - Marca: ASUS
  - Modelo: VivoBook 14 X412D
  - Procesador: AMD Ryzen 5 3500U CPU @ 2.10 GHz
  - Memoria RAM: 12 GB
  - Capacidad de almacenamiento: SSD 512 GB
  - Tarjeta gráfica: Radeon Vega Mobile
  - Sistema operativo: Windows 10 Home Single Language
- Computador portátil 2
  - Marca: Lenovo
  - Modelo: IdeaPad L340-15IRH Gaming
  - Procesador: Intel Core i5-9300H CPU @ 2.40GHz
  - Memoria RAM: 8GB
  - Capacidad de almacenamiento: SSD 128 GB
  - Tarjeta gráfica: nVIDIA GEFORCE GTX 1560
  - Sistema operativo: Windows 10

La fuente de datos corresponde al CNPV realizado en Colombia en el año 2018, el cual contiene 5 *datasets* (personas, vivienda, hogares, fallecidos y marco de georreferenciación)

### **3.1.2.2. Requisitos, supuestos y restricciones**

El acceso a la información de salud como comorbilidades es restringido debido a la política de privacidad y confidencialidad del Ministerio de Salud y Protección Social de Colombia.

### **3.1.2.3. Costos y beneficios**

Los datos implementados en este proyecto no suponen ningún costo adicional a la Universidad, debido a que la información recolectada se obtuvo de fuentes de datos abiertas, las cuales son gratuitas y de acceso libre al público.



Respecto a beneficios, para la Universidad no se genera un beneficio económico, pero si intelectual debido al artículo publicado a partir de este trabajo en IOS Press Ebooks [33].

### 3.1.3. Determinar los objetivos de la minería de datos

Los objetivos en términos de minería de datos fueron:

- Identificar en el CNPV 2018 aquellas variables que se asemejen a las variables que tuvo en cuenta el índice del CDC.
- Predecir el valor de vulnerabilidad de los municipios de Colombia a partir de la información obtenida del CNPV 2018.

#### 3.1.3.1. Criterios de éxito de minería de datos

Se estableció como criterio de éxito desde el punto de vista de la minería de datos la posibilidad de realizar predicciones con un elevado porcentaje de fiabilidad.

### 3.1.4. Plan del proyecto

El plan del proyecto que se muestra en la Figura 2 incluye el tiempo de elaboración del índice inicial (Capítulo 3) y el índice multidimensional (Capítulo 4).

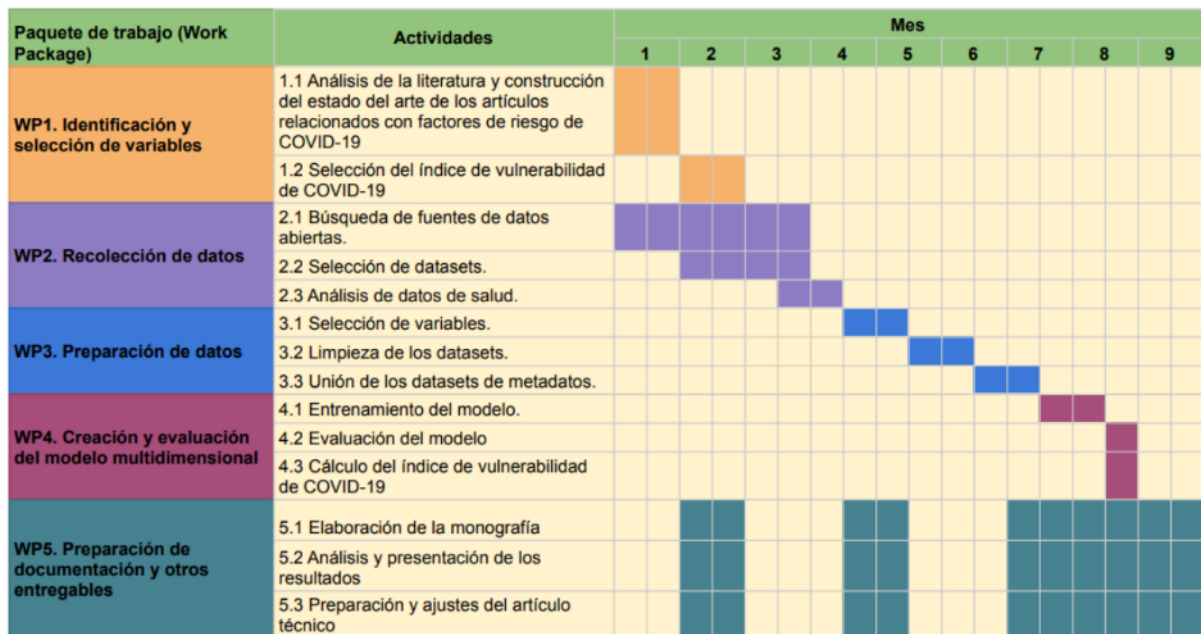


Figura 2. Plan del proyecto

## 3.2. Comprensión de los datos

En esta fase se realizó la captura inicial de los datos, de esta manera se logró conocer las características generales de los mismos, como su formato, cantidad de registros, entre otras. Además, mediante una exploración de los datos, observar la distribución y relaciones entre distintos conjuntos de estos, e identificar problemas de calidad en los mismos

### 3.2.1. Información del DANE

La recolección de datos para el índice de vulnerabilidad inicial se hizo a partir de los datos abiertos que proporciona el DANE respecto al último censo realizado, el CNPV 2018 donde caracterizó las personas residentes en Colombia, al igual que las viviendas y los hogares del territorio nacional [26]. Estos datos están en formato de Valores Separados por Comas (*Comma Separated Values* o *CSV*), y clasificados por departamento, además cuentan con información sociodemográfica dividida en 5 *datasets* de los cuales, a partir de las variables seleccionadas se consideraron 4 que proveían información relevante; los *datasets* usados fueron: vivienda, hogares, fallecidos y personas [27]. En la respectiva fuente de datos se encuentra el diccionario de datos, para cada *dataset*, en el cual se explica cada variable y las posibles opciones de respuesta de la misma. Estos datos pudieron ser leídos mediante la librería Pandas de *Python*. En el Anexo C denominado “Repositorio de Github” en la carpeta “*Initial Index*” se encuentra el diccionario de datos para cada uno de los *datasets* que se obtuvieron como resultado de los *datasets* publicados por el DANE.

Considerando las variables semejantes a las del índice realizado por el CDC se obtuvieron los siguientes *datasets*.

#### 3.2.1.1. *Dataset* “Vivienda”

El *dataset* “vivienda” cuenta con una dimensionalidad de 16.080.499 filas y 3 columnas, donde las variables U\_DPTO y U\_MPIO manejan como tipo de dato *int64* y no presentan datos nulos, por el contrario, la variable V\_TIPO\_VIV es de tipo *float64* y tiene 9.606 datos nulos.

### **3.2.1.2. Dataset “Hogares”**

El *dataset* “hogares” cuenta con una dimensionalidad de 14.252.829 filas y 5 columnas, donde las variables U\_DPTO y U\_MPIO son de tipo de dato *int64* y no presentan datos nulos, por otro lado, las variables H\_NRO\_DORMIT y HA\_TOT\_PER tiene 9.606 datos nulos y tienen como tipo de dato *float64* al igual que la variable HA\_NRO\_FALL que tiene 14.025.567 datos nulos.

### **3.2.1.3. Dataset “Fallecidos”**

El *dataset* “fallecidos” cuenta con una dimensionalidad de 242.744 filas y 4 columnas, donde todas sus variables, U\_DPTO, U\_MPIO, FA2\_SEXO\_FALL y FA3\_EDAD\_FALL, son de tipo *int64* y no presentan datos nulos.

### **3.2.1.4. Dataset “Personas”**

El *dataset* “personas” cuenta con una dimensionalidad de 44.164.417 filas y 12 columnas, donde la variable U\_DPTO no presenta datos nulos al igual que, U\_MPIO, P\_SEXO, P\_EDADR y PA1\_GRP\_ETNIC, las anteriores variables tienen como tipo de dato *int64*, mientras que, las siguientes variables son de tipo *float64*, PA\_HABLA LENG tiene 42.253.080 datos nulos al igual que PB\_OTRAS LENG; PB1\_QOTRAS LENG tiene 43.973.903 datos nulos, PA1\_CALIDAD\_SERV contiene 40.872.069 datos nulos, P\_ALFABETA tiene 3.037.781 datos nulos, P\_NIVEL\_ANOSR tiene 3.037.781 datos nulos y P\_TRABAJO contiene 6.682.277 datos nulos.

## **3.2.2. Datos del índice de vulnerabilidad del DANE**

Los datos de vulnerabilidad fueron recolectados del geo-portal de vulnerabilidad realizado por el DANE el cual permite descargar los archivos en formato *Shapefile*. Este es un formato que permite almacenar la información geométrica y la información de atributos de las entidades geográficas. Estos datos son estáticos y se encuentran para la mayoría de los municipios en Colombia, puesto que, son el resultado del índice de vulnerabilidad que llevó a cabo el DANE el cual se menciona en la sección 2.2. de este documento [34]. Este *dataset* cuenta con una dimensionalidad de 1102 filas y 4 columnas las cuales no presentan datos nulos, las variables COD\_MPIO y

Vulnerabilidad\_numero están como tipo de dato *int64*, y las variables CATEGORIA y ipm son de tipo *float64*.

### **3.3. Preparación de los datos**

#### **3.3.1. Limpieza y pre-procesamiento de datos**

Es necesaria la limpieza de datos para cada *dataset*, debido a que estos no cuentan con todos sus datos, es decir, existen casillas vacías; y en otros casos las variables cuentan con múltiples opciones, por tal motivo se aplicó la técnica *One Hot Encoding*, que es un proceso de conversión de variables de datos categóricos, es decir que se convierte cada valor categórico en una nueva columna categórica y se asigna un valor binario de 1 o 0 a esas columnas [35], lo anterior se realizó para los *datasets* de vivienda, hogares, fallecidos y personas. En el Anexo C (denominado “Repositorio de Github”) en la carpeta “*Initial index*” también se encuentran los códigos realizados para la limpieza y pre-procesamiento de los datos. Además, en las siguientes subsecciones se menciona el proceso de limpieza para los *datasets* considerados en este índice.

##### **3.3.1.1. Limpieza del *dataset* “vivienda” para todos los departamentos**

De las 30 variables que tiene el archivo vivienda, se escogieron: departamento, municipio y tipo de vivienda; porque son las que van acorde al índice de vulnerabilidad del CDC, esto se realizó para cada uno de los departamentos. Luego, se unieron los *datasets* resultantes en uno llamado vivienda, es decir, el *dataset* de vivienda cuenta con la información de los 32 departamentos y el distrito capital Bogotá D.C. Este *dataset* está compuesto por 3 columnas y 16.080.499 filas, de las cuales la columna V\_TIPO\_VIV (tipo de vivienda) tiene 9.606 filas nulas. Las filas nulas fueron reemplazadas por el número 6 que hace referencia a otra clase de vivienda. Lo anterior, debido a que, el no tomar una decisión respecto a los datos nulos generaría un problema para continuar con el pre-procesamiento y si se eliminaban estas filas ocasionaba que se perdieran datos de otras columnas para esas 9.606 filas.

##### **3.3.1.2. Pre-procesamiento del *dataset* “vivienda”**

La variable V\_TIPO\_VIV puede tener valores enteros entre el 1 y el 6, por lo que, considerando la técnica de *One Hot Encoding* se creó una columna para cada opción

teniendo en cuenta que el orden es casa, apartamento, tipo cuarto, vivienda tradicional indígena, vivienda tradicional étnica u otra clase de vivienda respectivamente, y se añadió la columna de nombre departamento para así identificar fácilmente el Departamento al cual pertenecía cada municipio. Posterior a ello, se procedió a asignar el número uno o cero en cada una de las columnas creadas considerando el dato de la columna V\_TIPO\_VIV para cada fila, teniendo así por fila 5 columnas en cero y una columna en uno, la cual hacía referencia al tipo de vivienda que tenía relación con el valor de 1 a 6 del tipo de vivienda.

Finalmente, para las 6 columnas creadas se realizó una suma de todas las filas por municipio, con el objetivo de tener un dato para cada columna por municipio y en ese sentido se obtuvo un *dataset* de 9 columnas por 1122 filas.

### **3.3.1.3. Limpieza del *dataset* “hogares” para todos los departamentos**

De las 13 variables que tiene el archivo hogares, se consideraron las siguientes: departamento, municipio, número de cuartos para dormir, total fallecidos en el hogar (2017) y total de personas en el hogar. Se unieron los *datasets* resultantes en uno llamado hogares, es decir, el *dataset* de hogares cuenta con la información de los 32 departamentos y el distrito capital Bogotá D.C. Este *dataset* está compuesto por 5 columnas y 14.252.829 filas, de las cuales la columna H\_NRO\_DORMIT (número de cuartos para dormir) tiene 9.606 filas nulas, HA\_NRO\_FALL (total fallecidos en el hogar 2017) 14.025.567 y HA\_TOT\_PER (total personas en el hogar) 9.606.

Los datos nulos en H\_NRO\_DORMIT se reemplazaron por 1, indicando que hay al menos 1 dormitorio, para HA\_NRO\_FALL se reemplazaron por 0, indicando que no hay ningún fallecido, para HA\_TOT\_PER se reemplazaron por 1, indicando que al menos una persona vive en el hogar. Lo anterior, debido a que, para continuar con el pre-procesamiento se necesitaba que las filas nulas tuvieran un valor. Para el caso de la columna H\_NRO\_DORMIT y HA\_TOT\_PER se realizaron las suposiciones mencionadas puesto que, “un hogar es una persona o grupo de personas que ocupan la totalidad o parte de una vivienda” [36]. Considerando lo anterior, se reemplazaron los datos nulos por 1, debido a que, según la literatura en un hogar al menos habría una persona y en ese sentido su habitación haría referencia a 1 dormitorio. Respecto al valor de HA\_NRO\_FALL se reemplazaron los datos nulos por cero puesto que, no se tenía certeza si en el año inmediatamente anterior a la encuesta del CNPV 2018

había muerto al menos una persona por hogar, y en ese sentido estaba dentro de las posibilidades que ninguna persona hubiese muerto.

#### **3.3.1.4. Pre-procesamiento del *dataset* “hogares”**

La variable H\_NRO\_DORMIT puede tener valores enteros del 1 al 20, los cuales indican la cantidad de dormitorios y el valor 99 para la opción de no informa. Considerando la técnica de *One Hot Encoding* se crearon nuevas columnas del 1 al 4 para indicar el número de dormitorios en el hogar, se añadió una columna para más de 4 dormitorios y otra en caso de no informar número de dormitorios, cabe mencionar que se creó una columna para más de 4 dormitorios puesto que, crear 21 columnas para los datos de esta variable (incluyendo todas las posibilidades numéricas consideradas por el DANE) incrementaba considerablemente la dimensionalidad del *dataset*.

La variable HA\_NRO\_FALL puede tener valores enteros entre 0 y 20 para señalar el número de fallecidos. Se crearon nuevas columnas del 0 al 2 para indicar el número de fallecidos en el hogar, y se agregó una columna para más de 2 fallecidos, lo anterior nuevamente considerando la técnica de *One Hot Encoding* y la dimensionalidad del *dataset*.

La variable HA\_TOT\_PER puede tener valores enteros entre 1 y 40, por lo que, teniendo en cuenta la técnica de *One Hot Encoding* y la dimensionalidad del *dataset* se crearon nuevas columnas del 1 al 4 para indicar la cantidad total de personas en el hogar y se agregó una columna para más de 4 personas.

Además, fue agregada la columna de nombre de Departamento.

Después, se procedió a asignar el número uno o cero en cada una de las columnas creadas, considerando el dato que había en las 3 columnas de referencia HA\_NRO\_DORMIT, HA\_NRO\_FALL, HA\_TOT\_PER.

Finalmente, se sumaron todas las filas por municipio y se obtuvo un *dataset* de 18 columnas por 1122 filas.

#### **3.3.1.5. Limpieza del *dataset* “fallecidos” para todos los departamentos**

De las 11 variables que tiene el archivo fallecidos, se seleccionaron las siguientes variables: departamento, municipio, sexo fallecido y edad al morir. Se unieron los *datasets* resultantes en uno llamado fallecidos, es decir, el *dataset* de fallecidos

cuenta con la información de los 32 departamentos y el distrito capital Bogotá D.C. Este *dataset* está compuesto por 4 columnas y 242.744 filas, no contiene datos nulos.

#### **3.3.1.6. Pre-procesamiento del *dataset* “fallecidos”**

La variable FA2\_SEXO\_FALL (sexo del fallecido) puede ser hombre, mujer o no informa, por lo que se creó una columna para cada una de las opciones teniendo en cuenta la técnica *One Hot Encoding*.

La variable FA3\_EDAD\_FALL (edad al morir) puede ser un número entero del 0 al 121, indicando la edad o 999 para no informa. Considerando la cantidad de opciones para la variable se realizaron agrupaciones para así no aumentar significativamente la dimensionalidad del *dataset*. Se creó una columna para la primera infancia, esta contiene las edades de 0 a 5 años; la columna infancia contiene desde el 6 al 11; la columna juvenil-adolescente desde el 12 hasta 26; la columna adultez desde los 27 hasta los 59 y la columna persona mayor desde los 60 hasta los 121; se añade una columna para no informa y nombre departamento. La clasificación anterior se realizó teniendo en cuenta el ciclo de vida publicado por el ministerio de salud [37]. Luego, se procedió a asignar el número uno o cero en cada una de las columnas creadas, considerando el dato que había en las 2 columnas de referencia FA2\_SEXO\_FALL y FA3\_EDAD\_FALL.

Finalmente, se sumaron todas las filas por municipio y se obtuvo un *dataset* de 12 columnas por 1119 filas.

#### **3.3.1.7. Limpieza del *dataset* “personas” para todos los departamentos**

De las 49 variables que tiene el archivo personas, se trabajó con las siguientes: departamento, municipio, sexo, edad en grupos quinquenales, reconocimiento étnico, habla la lengua nativa de su pueblo, habla otra(s) lengua(s) nativa(s), calidad de la prestación del servicio de salud (Esta pregunta depende de la variable: algún problema de salud en los últimos 30 días, sin hospitalización; en caso de que la respuesta fuera no, no aplicaba la variable de calidad de la prestación del servicio de salud), sabe leer y escribir, nivel educativo más alto alcanzado y último año o grado aprobado en ese nivel (recodificado) y, qué hizo durante la última semana. Se unieron los *datasets* resultantes en uno llamado personas, es decir, el *dataset* de personas cuenta con la información de los 32 departamentos y el distrito capital Bogotá D.C. Este *dataset* está compuesto por 12 columnas y 44.164.417 filas, de las cuales

PA\_HABLA\_LENG (habla la lengua nativa de su pueblo) tiene 42.253.080 filas nulas, al igual que PB\_OTRAS\_LENG (habla otras lenguas nativas), PB1\_QOTRAS\_LENG (número de otras lenguas nativas) tiene 43.973.903, PA1\_CALIDAD\_SERV (calidad de la prestación del servicio de salud) 40.872.069, P\_ALFABETA (sabe leer y escribir) 3.037.781, P\_NIVEL\_ANOSR (nivel educativo más alto alcanzado y último año o grado aprobado en ese nivel) 3.037.781 y P\_TRABAJO (que hizo durante la semana pasada) 6.682.277. Los datos nulos fueron reemplazados en cada una de las respectivas columnas por el valor que hacía referencia a la opción “no aplica”, puesto que, era necesario tomar una decisión para poder continuar con el pre-procesamiento de los datos. Para esta sección se escogió la opción de “no aplica” ya que, en la variable “habla la lengua nativa de su pueblo”, se refiere a personas con un reconocimiento étnico que hablan una determinada lengua, entonces, al tener este dato nulo se procedió a reemplazarlo por “no aplica” lo cual indica que a esa persona no se le hizo esa pregunta. De igual manera se realizó con la variable de calidad de la prestación del servicio de salud, debido a que, esta pregunta aplicaba si la persona había asistido en los últimos 30 días al hospital, por lo que, si estaba nula la mejor opción era asignar la opción “no aplica”. En las otras variables se escogió no aplica para indicar que no se realizó esa pregunta por tanto el valor estaba nulo, y de esa manera no afectar sustancialmente un análisis al escoger otra opción.

#### **3.3.1.8. Pre-procesamiento del *dataset* “personas”**

La variable P\_SEXO (sexo) puede ser hombre o mujer, por lo tanto, se creó una columna para cada opción.

La variable P\_EDADR (edad en grupos quinquenales) puede ser un valor entero entre 1 y 21, se creó una columna para cada grupo quinquenal de la 1 hasta la 12, y una columna extra para más de 60 años, con el objetivo de no aumentar considerablemente la dimensionalidad del *dataset*.

La variable PA1\_GRP\_ETNIC (reconocimiento étnico) puede ser un valor entero del 1 al 6 o 9, se creó una columna para cada opción teniendo en cuenta que el orden es indígena, gitano o Rrom, raizal del archipiélago de San Andrés, Providencia y Santa Catalina (SAPSC), palenquero de San Basilio, negro-mulato-afrodescendiente-afrocolombiano, ningún grupo étnico o no informa.



La variable PA\_HABLA\_LENG (habla la lengua nativa de su pueblo) tiene 4 opciones, se creó una columna para cada una de ellas, las opciones son: si, no, no informa o no aplica.

Las variables PB\_OTRAS\_LENG y PB1\_QOTRAS\_LENG se eliminaron porque la gran mayoría de filas son nulas.

La variable PA1\_CALIDAD\_SERV tiene 6 opciones, por lo tanto, se creó una columna para cada opción siendo muy bueno, bueno, malo, muy malo, no informa y no aplica, respectivamente.

La variable P\_ALFABETA tiene 4 opciones, se creó una columna para cada opción las cuales son si sabe leer y escribir, no sabe leer y escribir, no informa y no aplica, respectivamente.

La variable P\_NIVEL\_ANOSR tiene opciones del 1 al 10, 99 y 999, se creó una columna para cada opción, estas son preescolar, primaria, secundaria, media académica, media técnica, normalista, técnica profesional, universitario, posgrado, ninguno, no informa escolaridad y no aplica, respectivamente.

La variable P\_TRABAJO tiene opciones del 0 al 9 y no aplica, se crearon 4 columnas recibe ingresos, no recibe ingresos, no informa ingresos, no aplica ingresos. En la columna recibe ingresos entraron las opciones 1, 3 y 5; en la columna no recibe ingresos entraron las opciones 2, 4, 6, 7, 8 y 9.

Es importante mencionar que el proceso de crear nuevas columnas para los valores de las variables mencionadas anteriormente, se realizó considerando la técnica *One Hot Encoding*. Seguidamente, se procedió a asignar el número uno o cero en cada una de las columnas creadas, considerando el dato que había en cada una de las columnas que brinda el DANE.

Finalmente, se sumaron todas las filas por municipio y se obtuvo un *dataset* de 71 columnas por 1122 filas.

### **3.3.1.9. Pre-procesamiento del *dataset* “Vulnerabilidad COVID-19”**

Para el *dataset* de vulnerabilidad de la COVID-19 se tomó el reporte general que se descarga del geo-portal de vulnerabilidad del DANE. Posteriormente, fue leído el *Shapefile* en *Jupyter*, se eliminaron las columnas que no eran de interés para este trabajo (vulnerabilidad embarazo adolescente, reactivación económica empleo joven, código DANE, código departamento y *geometry*) y se asignaron valores numéricos para los niveles de vulnerabilidad existentes. Después de ello, fue calculada la

mediana (la cual representa el valor de la mitad de todo el conjunto de datos organizándolos ascendentemente [38]). Lo anterior, se hizo considerando que el índice de vulnerabilidad fue calculado por manzanas, por lo cual, se podría estar tomando un valor promedio por municipio que no fuera el adecuado. En este proceso, fue muy importante borrar los datos nulos que existían por municipio dado que, al tener un valor de cero (valor por el que se reemplazaron los datos nulos), ocasionaban que para algunos municipios la mediana presentara un valor errado. Para lo anterior, es importante mencionar que los valores de vulnerabilidad estaban desde el valor 1 a 5, por lo que el valor de 0 no indicaba ningún nivel de vulnerabilidad, sin embargo, si se dejaban esos datos, la mediana podía en algunos casos tomar ese valor y no se conocería el nivel de vulnerabilidad a predecir de ese municipio. Finalmente, se descargó el archivo CSV en el cual se tiene la mediana del valor de vulnerabilidad por cada municipio. Es importante resaltar la relevancia de esta columna, ya que se trata de la variable dependiente.

### **3.3.2. Integración de los *datasets***

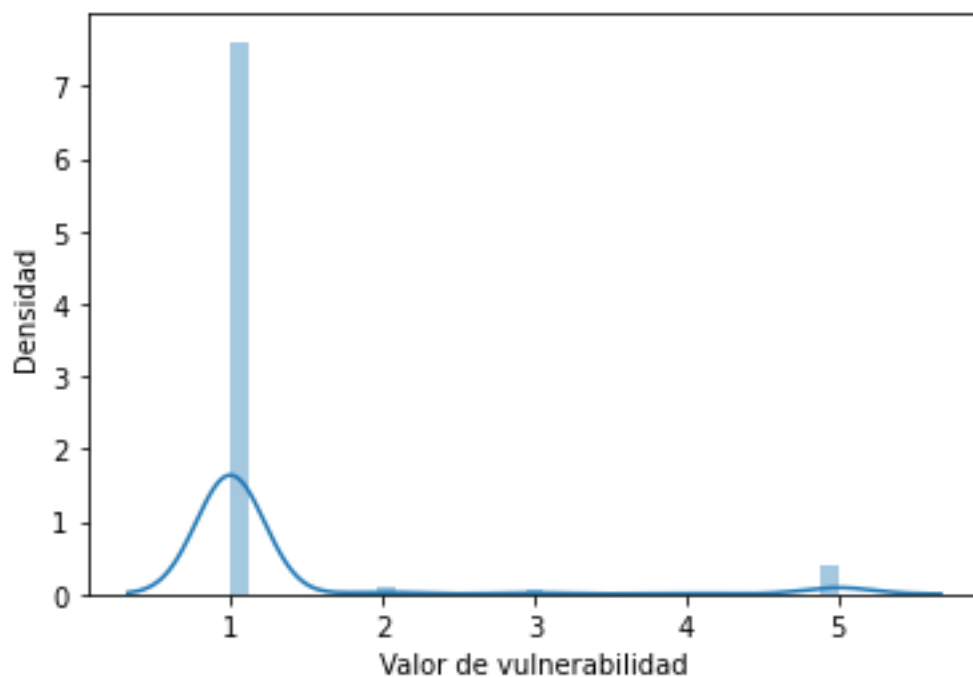
Con el objetivo de tener un único *dataset* que incluyera todas las variables se realizó la integración de datos, se tomó el *dataset* final de cada grupo, es decir el *dataset* de vivienda, hogares, fallecidos, personas y vulnerabilidad COVID-19 dando como resultado un *dataset* de 89 columnas con 1122 filas, en el cual se eliminaron las filas que no tenían un valor para vulnerabilidad, puesto que, esa era nuestra variable dependiente, arrojando así un *dataset* de 89 columnas por 1103 filas. En el Anexo D denominado “*Dataset* inicial”, está el *dataset* que se obtuvo como resultado de esta etapa.

## **3.4. Análisis Exploratorio de Datos (EDA)**

El análisis exploratorio de datos (EDA) [39] se realizó a partir del *dataset* inicial creado en la integración de datos, este *dataset* contiene una dimensionalidad 88 columnas y 1103 filas, donde cada fila contiene información de un municipio de Colombia. De las 89 columnas, 88 son tipo de dato numérico (10 de ellas son *float64* y 78 son *int64*) y una es *object* (Nom\_DPTO), la cual indica el nombre del Departamento. Este *dataset* no cuenta con datos nulos debido al pre-procesamiento hecho antes de la unificación de los *datasets*. La representación de datos mediante gráficas, permite un mejor

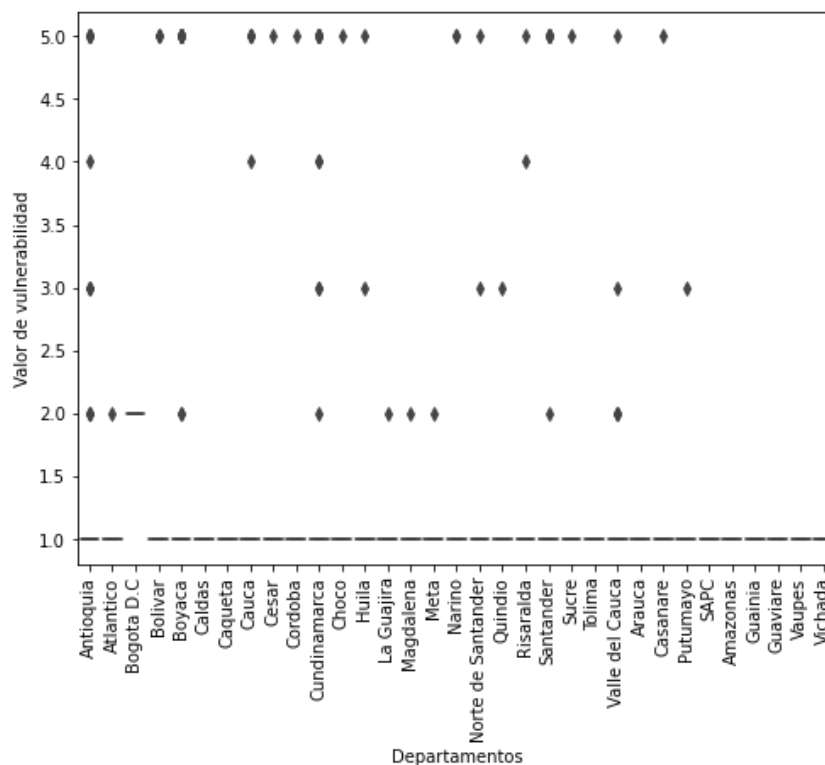
entendimiento para el análisis de datos. Para la elaboración de las gráficas se implementó la librería *Seaborn* de *Python*, la cual está basada en *Matplotlib*.

Realizando un análisis uni-variable para la variable objetivo “Vulnerabilidad\_numero” (con el fin de estudiarla con mayor detenimiento debido a que esta será la variable a predecir), se puede decir que consta de 1102 datos, los cuales se dividen en 5 categorías (valores enteros de vulnerabilidad entre 1 y 5). El promedio de los datos es 1,245009, y se presenta una desviación estándar de 0,911052. El histograma en la Figura 3 muestra que la mayoría de datos se encuentran en el valor 1, y muy pocos datos para las demás categorías indicando que se tienen datos desbalanceados, por lo tanto, no hay suficientes datos para entrenar adecuadamente las demás categorías.



**Figura 3.** Densidad de los datos para cada categoría de la variable objetivo.

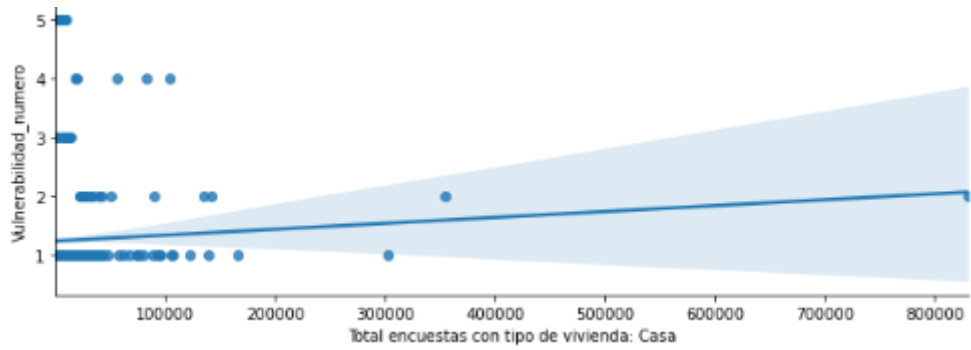
La relación de las variables independientes y la variable dependiente ayuda a conocer la distribución de los datos, en la Figura 4 mediante un gráfico de cajas se observa la distribución de los datos para cada Departamento. En las Figuras 5, 6, y 7 se presentan gráficos de dispersión donde se muestra la relación de algunas variables independientes (personas que viven en una casa, edad de las personas entre los 20 y 24 años, y personas identificadas como indígenas) y la variable *Vulnerabilidad\_numero*.



**Figura 4.** Gráfico de cajas para los departamentos.

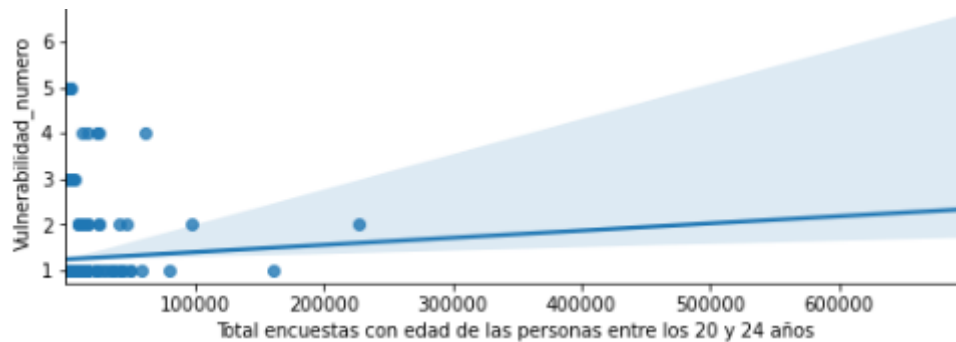
En la Figura 4 se observa que, en todos los Departamentos, no hay una adecuada distribución de datos, debido a que la gran cantidad de datos de valor 1. respecto a las demás categorías es muy grande, exceptuando Bogotá D.C. donde el valor de vulnerabilidad se concentra en 2. Por ello, al presentar datos en las otras opciones, la gráfica los representa como valores atípicos y no dentro del rango de los posibles valores.

Los gráficos de dispersión se utilizan para observar la correlación entre dos variables. Las Figuras 5, 6, y 7 muestran los gráficos de dispersión obtenidos para las variables independientes que representan tipo de vivienda Casa, Edad de las personas entre 20 y 24 años y reconocimiento étnico indígena. Estas variables se ubican en eje X, mientras que en el eje Y está el valor de vulnerabilidad que es la variable dependiente o variable objetivo.



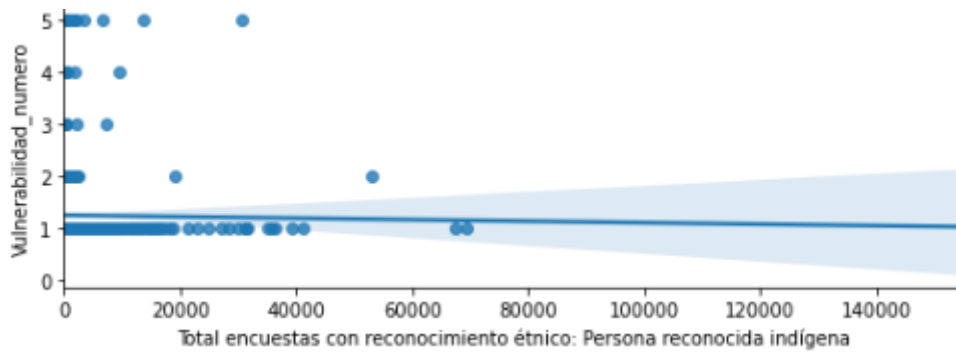
**Figura 5.** Gráfico de dispersión para CASA.

De la Figura 5 se puede observar que en la mayoría de los municipios se encuentran entre 0 y 100000 encuestas en las que se tiene un tipo de vivienda CASA. También se nota que la vulnerabilidad de 1 es la que tiene mayor dominancia en la gráfica y no se percibe ninguna relación entre la variable vulnerabilidad y la cantidad de encuestas. Por otra parte, también es posible visualizar que los valores de vulnerabilidad de 3 y 5 son los que menos presencia tienen en los municipios y esa vulnerabilidad se presenta en municipios donde se tienen entre 0 y aproximadamente 3000 encuestas con tipo de vivienda casa.



**Figura 6.** Gráfico de dispersión para Per\_edad\_20 a 24.

La Figura 6 permite observar que en la mayoría de municipios se presentan entre 0 y aproximadamente 80000 encuestas en donde las personas tienen entre 20 y 24 años de edad y hay un valor de vulnerabilidad de 1. También se nota que el valor de vulnerabilidad de 1 es el de mayor dominancia. Respecto a la relación entre las variables, la Figura 3 permite visualizar que no existe una relación lineal entre la cantidad de encuestas donde las personas tienen entre 20 y 24 años y el valor de vulnerabilidad.



**Figura 7.** Gráfico de dispersión para Indígena.

De la Figura 7 se visualiza que no existe una relación lineal entre la cantidad de encuestas donde las personas se reconocen como indígenas y el valor de vulnerabilidad. También, se puede observar que el valor de vulnerabilidad de 1 es el de mayor dominancia en esta gráfica.

En los gráficos de dispersión de las Figuras 5, 6, y 7 se observó que los puntos en su gran mayoría se encuentran cerca del origen del eje X, y no se observa una tendencia indicando que no hay relación lineal entre las variables.

Para realizar un análisis más objetivo se realizaron pruebas de correlación, aplicando una matriz de correlación general o también conocida como mapa de calor. El coeficiente de correlación mide la relación entre 2 variables, esto necesariamente no implica causalidad, pero puede ser útil para realizar predicciones. La correlación de Pearson evalúa la relación lineal entre dos variables cuantitativas [40], por lo observado en los gráficos de dispersión se puede decir que los valores de correlación son cercanos a cero, lo que indica que no hay relación lineal entre las variables.

El coeficiente de correlación de Pearson indica cuán asociadas están dos variables y es un valor entre -1 y 1. Una correlación negativa significa que las variables se relacionan inversamente, una correlación positiva, que las variables se relacionan directamente, y si el valor es cercano a cero significa que no existe una relación lineal entre las variables. En la Figura 8 se observa el valor de correlación de Pearson entre las variables independientes y la variable dependiente, donde el valor para todas las variables es bajo.

Vulnerabilidad_numero					
Vulnerabilidad_numero	1.000000	CSS malo	0.043910	Fall_Per_Mayor	0.040852
CSS muy malo	0.049236	Media_Academica	0.043138	4-PER	0.040836
No informa sexo fall	0.048800	3-PER	0.041957	Lee_y_Escribe	0.040734
Fall_Infancia	0.047953	Ningun grupo etnico	0.041926	2-PER	0.040728
CSS muy bueno	0.047838	Per_edad_30 a 34	0.041793	No aplica otra lengua	0.040412
No informa edad fall	0.047473	Per_edad_35 a 39	0.041708	No aplica lengua nativa	0.040412
4-DOR	0.046552	CSS bueno	0.041455	Per_edad_20 a 24	0.040406
Mas de 4 dor	0.045373	Recibe ingresos	0.041424	0-FALL	0.040366
3-DOR	0.045226	Per_edad_55 a 59	0.041092	Per_edad_45 a 49	0.040222
APTO	0.044964	Per_edad_50 a 54	0.041075	Fall_Adultez	0.040167
Tecnica_Profesional	0.044760	Normalista	0.041048	Fall_Hombre	0.040164
Mas de 2 fall	0.044655	Per_edad_25 a 29	0.040977	1-FALL	0.040131
		Secundaria	0.040915	Per_Mujer	0.040107
Fall_Mujer	0.040067	No_informa_Alfabeta	0.037713	Media_tecnica	0.035042
Per_edad_40 a 44	0.039986	Per_edad_10 a 14	0.037223	CASA	0.035041
Per_edad_Mas de 60	0.039963	No_inf_escolaridad	0.036820	Posgrado	0.034921
2-DOR	0.039894	No Aplica ingresos	0.036740	2-FALL	0.034304
Per_Hombre	0.039765	No Aplica P_ALFABETA	0.036693	TIP_CUARTO	0.032459
CSS No Aplica	0.039611	No Aplica P_NIVEL_ANOSR	0.036693	Fall_P_Infancia	0.030252
1-PER	0.039436	Per_edad_0 a 4	0.036693	OTRA_VIV	0.030076
Per_edad_15 a 19	0.038866	No informa ingresos	0.036583	VIV_TRAD_ETNICA	0.025508
No recibe ingresos	0.038651	Per_edad_5 a 9	0.036539	NoLee_y_Escribe	0.016252
Universitario	0.038396	Mas de 4 per	0.036446	Ninguno	0.014153
Primaria	0.038345	No informa grupo etnico	0.036289	Gitano o Rrom	0.014065
Fall_Juv-Adoles	0.038159	No informa Num_dormit	0.036066	Palenquero de San Basilio	-0.003234
Preescolar	0.038146	1-DOR	0.035139	Raizal del Archi SAPSC	-0.007449
		No habla lengua nativa	-0.007471		
		No habla otra lengua	-0.008209		
		Negro, mulato, afrodescendiente, afrocolombiano	-0.009754		
		Habla lengua Nativa	-0.009883		
		Indigena	-0.011360		
		U_MPIO	-0.012137		
		VIV_TRAD_INDIG	-0.012140		
		Habla otra lengua	-0.020387		
		No informa otra lengua	-0.031855		
		No informa lengua nativa	-0.035833		
		U_DPTO	-0.063122		

**Figura 8.** Valores de correlación de Pearson entre las variables independientes y variable dependiente.

Con el objetivo de conocer si se presenta una relación monótona entre las variables independientes y la dependiente, es decir, si las variables tiendan a cambiar en el mismo instante de tiempo, se calculó la correlación de Spearman, los resultados se presentan en la Figura 9.

Vulnerabilidad_numero					
Vulnerabilidad_numero	1.000000	No informa lengua nativa	0.038637	Ningun grupo etnico	0.029013
Mas de 2 fall	0.078096	Mas de 4 dor	0.037810	Per_edad_Mas de 60	0.028183
Raizal del Archi SAPSC	0.078002	CSS malo	0.037302	Fall_Hombre	0.027624
No informa edad fall	0.065671	Normalista	0.036219	No informa Alfabeta	0.024744
No informa sexo fall	0.065658	4-DOR	0.035594	Media_tecnica	0.024306
Gitano o Rrom	0.061300	Posgrado	0.033741	Tecnica_Profesional	0.024140
TIP_CUARTO	0.053688	1-PER	0.033589	Palenquero de San Basilio	0.023481
Fall_Infancia	0.052897	CSS bueno	0.032018	Recibe ingresos	0.022888
APTO	0.046694	Fall_Mujer	0.031892	Fall_Adultez	0.021335
Fall_Juv-Adoles	0.045024	CSS muy malo	0.030268	OTRA_VIV	0.021266
Fall_Per_Mayor	0.042890	3-DOR	0.029349	2-PER	0.021252
2-FALL	0.042814	1-FALL	0.029289	No informa otra lengua	0.021124
		CSS muy bueno	0.029065	No informa grupo etnico	0.021056
No aplica otra lengua	0.020772	Per_Hombre	0.016136	Preescolar	0.012246
No aplica lengua nativa	0.020772	0-FALL	0.016128	No recibe ingresos	0.012217
Per_edad_55 a 59	0.020438	Per_edad_20 a 24	0.016030	No informa ingresos	0.011732
Per_edad_50 a 54	0.019290	Per_edad_25 a 29	0.015461	VIV_TRAD_INDIG	0.011670
3-PER	0.018507	Per_edad_35 a 39	0.015005	Ninguno	0.009019
2-DOR	0.017716	Per_edad_40 a 44	0.014622	CASA	0.008705
Per_edad_45 a 49	0.017475	CSS No Aplica	0.014554	Per_edad_10 a 14	0.008288
No_inf_escolaridad	0.017229	Universitario	0.014268	Per_edad_5 a 9	0.007805
Per_edad_30 a 34	0.016822	Per_Mujer	0.013723	No Aplica ingresos	0.007507
4-PER	0.016743	Mas de 4 per	0.013364	Fall_P_Infancia	0.006670
Primaria	0.016481	Media_Academica	0.013130	NoLee_y_Escribe	0.006092
1-DOR	0.016462	Per_edad_15 a 19	0.012439	No informa Num_dormit	0.005714
Lee_y_Escribe	0.016235	Secundaria	0.012315	No Aplica P_ALFABETA	0.004423
		Per_edad_0 a 4	0.004423		
		No Aplica P_NIVEL_ANOSR	0.004423		
		Negro, mulato, afrodescendiente, afrocolombiano	-0.005541		
		No habla lengua nativa	-0.027573		
		VIV_TRAD_ETNICA	-0.029540		
		No habla otra lengua	-0.033648		
		Indigena	-0.036481		
		U_MPIO	-0.042272		
		Habla lengua Nativa	-0.043552		
		Habla otra lengua	-0.046172		
		U_DPTO	-0.061280		

**Figura 9.** Valores de correlación de Spearman entre las variables independientes y variable dependiente.

Al igual que la Figura 8, la Figura 9 permite visualizar que no hay correlación entre las variables.



## 3.5. Modelado y evaluación

### 3.5.1. Modelado

Los algoritmos de aprendizaje supervisado son en los que se conoce la variable objetivo, es decir que su aprendizaje está basado en datos de entrenamiento donde se conoce esta variable [41]. En ese sentido, estos algoritmos pueden clasificarse en regresión y clasificación. Cuando se aplica regresión, se predice la salida de valores continuos; para clasificación se predice salidas discretas o cualitativas [42]. Es por ello, que para el modelado y evaluación de este *dataset* se tienen en cuenta algoritmos de aprendizaje supervisado, puesto que, la variable objetivo para este *dataset* es la de vulnerabilidad. Adicionalmente, considerando que la variable de vulnerabilidad es multi-categorica, ya que tiene 5 posibles valores; 1, 2, 3, 4 y 5, los cuales son un conjunto de posibilidades finitas, se aplicaron algoritmos de clasificación.

En ese sentido, el objetivo del índice inicial es evaluar si las variables consideradas para el *dataset* inicial, las cuales fueron mencionadas en la sección 3.2. de este documento, realizan una buena predicción del valor de vulnerabilidad (El cual brinda el DANE en su índice) por municipio para así determinar qué tan adecuado es el índice inicial creado.

Inicialmente, se dividió el *dataset* en dos grupos y se probaron con porcentajes distintos, para el primer escenario se tomó 80% como entrenamiento y 20% como prueba, y para el segundo se tomó 70% como entrenamiento y 30% como prueba. Estos escenarios se consideraron con el objetivo de que no se presentara sobreajuste ni sub-ajuste, además, son escenarios comúnmente utilizados en aprendizaje automático [43-44].

Además, dado que, en las gráficas de dispersión, no se logró percibir ninguna tendencia, se usaron algoritmos diferentes para entrenar el *dataset*. Se utilizaron 6 modelos, cada uno con algoritmos diferentes, para así revisar qué modelo se comportaba mejor. Se incluyeron algoritmos de árboles de decisión, bayesianos y por instancia, para tener algoritmos de diferente naturaleza y verificar que tipo se comporta mejor.

El primer modelo implementado fue análisis del discriminante lineal (LDA), el cual, es un método de clasificación supervisado donde se construye un modelo predictivo para así predecir el grupo al cual pertenece. El segundo modelo fue análisis de discriminante cuadrático (QDA), este modelo es usado cuando el conjunto de variables predictoras que se desea clasificar tiene dos o más clases, se considera el equivalente al análisis del discriminante no lineal. El tercer modelo fue k-vecinos más cercanos (KNN), este es un algoritmo clasificador de aprendizaje supervisado, el cual se puede usar como algoritmo de regresión o clasificación. El cuarto clasificador fue el de árboles de decisión, el cual es un algoritmo de aprendizaje supervisado el cual es usado principalmente en problemas de clasificación, pero también se puede utilizar para variables de entrada y salida continuas. El quinto fue *Gaussian Naive Bayes* (GNB), es un algoritmo de *ML* probabilístico usado típicamente como clasificador. Y el sexto fue máquinas de vectores de soporte (SVM por su sigla en inglés), el cual se puede usar como clasificación, así como regresión.

#### **3.5.1.1. LDA, *Linear Discriminant Analysis* (Análisis del discriminante lineal)**

El análisis del discriminante lineal es un algoritmo que permite realizar un modelo predictivo, pronosticando las clases o grupos a los cuales pertenecen las observaciones, además, permite llevar a cabo una reducción de dimensionalidad supervisada puesto que, proyecta los datos de entrada en un subespacio lineal que se basa en las direcciones que maximizan la separación entre clases. La dimensionalidad requerida se puede configurar usando el parámetro *n\_components* el cual no influye en los métodos *fit* y *predict* [45].

#### **3.5.1.2. QDA, *Quadratic Discriminant Analysis* (Análisis del discriminante cuadrático)**

El análisis del discriminante cuadrático es una generalización de LDA, este asume que las observaciones para cada clase siguen una distribución normal multivariante. Este modelo llega a ser más flexible que LDA debido a que los límites de decisión que genera son curvos y no lineales [46].

#### **3.5.1.3. KNN, *K-Nearest-Neighbor* (K-vecinos)**

Este algoritmo de *ML* supervisado se basa en la idea de predicción de un dato teniendo en cuenta las características de los datos cercanos. Este algoritmo permite

escoger el número de vecinos (*n\_neighbors*) que puede tener en cuenta el modelo, dicho ajuste genera una mejor predicción o clasificación [47].

#### **3.5.1.4. *Decision Tree Classifier* (Clasificador de árboles de decisión)**

El algoritmo *Decision Tree Classifier* es un algoritmo de *ML* de aprendizaje supervisado. En este algoritmo se divide el *dataset* en dos o más conjuntos homogéneos, identificando la variable más significativa [48].

#### **3.5.1.5. *GNB, Gaussian Naive Bayes*.**

El algoritmo *Naive Bayes* se basa en el teorema de Bayes, funciona como clasificación probabilística en *ML*. Este teorema ofrece una fórmula que indica la probabilidad condicional de que ocurra cierto evento si ha ocurrido otro evento previamente. Este método supone que los predictores son independientes y contribuyen en la selección final de la misma manera [49].

#### **3.5.1.6. *SVM, Support vector machine* (Máquinas de vectores de soporte)**

El algoritmo de *SVM* puede ser aplicado tanto para clasificación, como para regresión. Se basa principalmente en dividir en varias clases los conjuntos de datos con el objetivo de encontrar un hiper-plano [50].

### **3.5.2. Evaluación**

Para evaluar el desempeño de los algoritmos se utilizaron las métricas *F1-score*, *precision*, *recall* y *accuracy*.

La métrica *F1-score* tiene en cuenta la cantidad de falsos positivos y falsos negativos, calculando un promedio ponderado entre precisión y sensibilidad. De esta manera se obtiene una puntuación única que representa a las dos variables; la métrica *precision* mide la precisión del clasificador al momento de predecir casos positivos, se calcula como la relación entre las predicciones correctas y el total de predicciones correctas previstas; la métrica *recall* permite detectar las instancias positivas, también se conoce como sensibilidad, se halla como la relación entre las predicciones positivas correctas y el total de predicciones positivas; la métrica *accuracy* determina la exactitud de la clasificación, es decir, la relación entre predicciones correctas y el total de predicciones [51].

Las Tablas 2 y 3 muestran los resultados de la evaluación de cada modelo para cada escenario, teniendo en cuenta las métricas de evaluación (F1-score, precision, recall, accuracy)

En la Tabla 2 se muestran los resultados obtenidos con un entrenamiento del 80% y prueba de 20% de los datos, con los diferentes modelos.

**Tabla 2.** Resultados obtenidos en la evaluación 80%-20% del índice inicial

<b>Modelo</b>	<b>Parámetros por defecto</b>	<b>F1-score</b>	<b>Precision</b>	<b>Recall</b>	<b>Accuracy</b>
<i>LinearDiscriminantAnalysis</i>	<i>solver='svd' shrinkage=None priors=None n_components=None store_covariance=False tol=0.0001 covariance_estimator=None</i>	0.256	0.236	0.293	0.900
<i>QuadraticDiscriminantAnalysis</i>	<i>priors=None reg_param=0.0 store_covariance=False tol=0.0001</i>	0.192	0.185	0.2	0.927
<i>KNeighborsClassifier</i>	<i>n_neighbors=5 weights='uniform' algorithm='auto' leaf_size=30 p=2 metric='minkowski metric_params=None n_jobs=None</i>	0.192	0.185	0.2	0.927
<i>DecisionTreeClassifier</i>	<i>criterion='gini' splitter='best' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=None random_state=None max_leaf_nodes=None min_impurity_decrease=0.0 class_weight=None ccp_alpha=0.0</i>	0.183	0.185	0.182	0.846
<i>GaussianNaiveBayes</i>	<i>priors=None var_smoothing=1e-09</i>	0.165	0.245	0.493	0.190
<i>SupportVectorMachine</i>	<i>C=1.0 kernel='rbf' degree=3</i>	0.192	0.185	0.2	0.927

<pre> gamma='scale' coef0=0.0 shrinking=True probability=False tol=0.001 cache_size=200 class_weight=None verbose=False max_iter=-1 decision_function_shape='ovr' break_ties=False random_state=None </pre>				
---	--	--	--	--

En la Tabla 3 se muestran los resultados obtenidos con un entrenamiento del 70% y prueba de 30% de los datos, con los diferentes modelos.

**Tabla 3.** Resultados obtenidos en la evaluación 70%-30% para el índice inicial.

<b>Modelo</b>	<b>Parámetros por defecto</b>	<b>F1-score</b>	<b>Precision</b>	<b>Recall</b>	<b>Accuracy</b>
<i>LinearDiscriminantAnalysis</i>	<pre> solver='svd' shrinkage=None priors=None n_components=None store_covariance=False tol=0.0001 covariance_estimator=None </pre>	0.239	0.226	0.259	0.897
<i>QuadraticDiscriminantAnalysis</i>	<pre> priors=None reg_param=0.0 store_covariance=False tol=0.0001 </pre>	0.192	0.185	0.2	0.927
<i>KNeighborsClassifier</i>	<pre> n_neighbors=5 weights='uniform' algorithm='auto' leaf_size=30 p=2 metric='minkowski' metric_params=None n_jobs=None </pre>	0.192	0.185	0.2	0.927
<i>DecisionTreeClassifier</i>	<pre> criterion='gini' splitter='best' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=None random_state=None max_leaf_nodes=None min_impurity_decrease=0.0 class_weight=None ccp_alpha=0.0 </pre>	0.429	0.427	0.432	0.873

<i>GaussianNaiveBayes</i>	<i>priors=None</i> <i>var_smoothing=1e-09</i>	0.141	0.231	0.504	0.184
<i>SupportVectorMachine</i>	<i>C=1.0</i> <i>kernel='rbf'</i> <i>degree=3</i> <i>gamma='scale'</i> <i>coef0=0.0</i> <i>shrinking=True</i> <i>probability=False</i> <i>tol=0.001</i> <i>cache_size=200</i> <i>class_weight=None</i> <i>verbose=False</i> <i>max_iter=-1</i> <i>decision_function_shape='ovr'</i> <i>break_ties=False</i> <i>random_state=None</i>	0.192	0.185	0.2	0.927

Las Tablas 2 y 3 permiten observar que el algoritmo *DecisionTreeClassifier* generó el mejor valor para la métrica *F1-score*, este valor se obtuvo cuando se dividió el conjunto de datos en 70%-30% (entrenamiento y prueba respectivamente), esto indica que el índice de vulnerabilidad realizado inicialmente logra predecir correctamente el valor de vulnerabilidad en un 87.3% entre el total de predicciones. Sin embargo, los valores de *precision* y *recall* son bajos, alrededor de 43%, lo que indica que el número de predicciones correctas respecto a las predicciones correctas previstas es bajo. Debido a esto, la métrica *F1-score* en el mejor resultado tiene un valor de 42.9%. Es importante mencionar que para estos modelos la métrica de referencia fue *F1-score*, la cual indica la cantidad de falsos positivos y falsos negativos. Esa métrica fue escogida como referencia debido a que el valor de vulnerabilidad (variable objetivo) es una variable multi-categorica, lo que ocasiona que la métrica *accuracy* tienda a ser alta y no sea la ideal.

De acuerdo a la revisión de la literatura, en la cual se encontró que las comorbilidades son consideradas factores de riesgo de la COVID-19; y teniendo en cuenta que tanto el índice del CDC como el del DANE utilizan los datos de comorbilidades para predecir la vulnerabilidad, las bajas métricas obtenidas en este modelo pueden deberse a la falta de datos para esta variable.

# Capítulo 4.

## 4. Índice de vulnerabilidad multidimensional

Un índice de vulnerabilidad multidimensional es el tipo de índice en el que se integran diferentes factores de riesgo. En este capítulo se presentan las fases de comprensión del negocio, comprensión de los datos, preparación de los datos, y la creación y evaluación del modelo del índice de vulnerabilidad multidimensional realizado.

### 4.1. Comprensión del negocio

La comprensión del negocio es la primera fase en el proceso de minería de datos, esta fase permite determinar los objetivos y requisitos del proyecto desde un punto de vista de negocio, para así posteriormente transformar los objetivos a una perspectiva técnica y desarrollar un plan de proyecto [28].

#### 4.1.1. Objetivos del negocio

En las siguientes subsecciones se presenta la determinación de los objetivos del negocio.

##### 4.1.1.1. Contexto

En referencia a la situación del negocio, se obtuvo que el índice de vulnerabilidad inicial no predecía valores altos para las métricas de vulnerabilidad del índice del DANE, por ende, las variables consideradas (las cuales se presentan en el capítulo 3) no eran lo suficientemente relevantes.

Para el índice multidimensional, se consideró como referencia los resultados del índice inicial. Se partió de los datos de vulnerabilidad entregados por el DANE y a ellos se les agregaron datos de variables de diferentes tipos encontradas en el mapeo sistemático y en la revisión de la literatura como variables de interés. Por lo cual, puede afirmarse que el índice propuesto considera factores sociodemográficos, socioeconómicos, ambientales y humanos. El índice propuesto se diferencia del índice de vulnerabilidad del DANE al tener en cuenta un conjunto más amplio de variables.

#### **4.1.1.2. Objetivos del negocio**

El objetivo del negocio fue realizar predicciones de la incidencia de casos de la COVID-19 lo más fiables posibles partiendo de información obtenida gracias a las fuentes abiertas disponibles para Colombia. De esta forma construir herramientas que faciliten la toma de decisiones, al Gobierno Nacional en cuanto a la emergencia sanitaria ocasionada por la COVID-19.

#### **4.1.1.3. Criterios de éxito del negocio**

Como criterios de éxito desde el punto de vista del negocio se identificaron: la posibilidad de realizar predicciones de la incidencia a contagiarse de la COVID-19 en las 24 ciudades principales consideradas; y en función de ello, que el gobierno tome medidas adecuadas para evitar su propagación.

#### **4.1.2. Evaluación de la situación**

Se identificaron las fuentes adicionales de datos, las cuales son, PIB, datos climatológicos, información de vacunación, porcentaje de desempleo, datos de movilidad, datos de la COVID-19 y dato de vulnerabilidad de la COVID-19.

##### **4.1.2.1. Inventario de recursos**

Para el inventario de recursos, en este índice se utilizaron los mismos recursos software y hardware nombrados en la sección 3.1.2.1.

Las fuentes de datos utilizadas fueron: PIB, datos climatológicos, información de vacunación, porcentaje de desempleo, datos de movilidad, datos de la COVID-19 y dato de vulnerabilidad de la COVID-19.

##### **4.1.2.2. Requisitos, supuestos y restricciones**

El acceso a la información de salud como comorbilidades es restringido debido a la política de privacidad y confidencialidad del Ministerio de Salud y Protección Social de Colombia.

##### **4.1.2.3. Costos y beneficios**

Los datos implementados en este proyecto no suponen ningún costo adicional a la Universidad, debido a que la información recolectada se obtuvo de fuentes de datos abiertas las cuales son gratuitas y de acceso libre al público.



Respecto a beneficios, para la Universidad no se genera un beneficio económico, pero si intelectual debido al artículo publicado a partir de este trabajo en IOS Press Ebooks [33].

#### **4.1.3. Determinar los objetivos de la minería de datos**

Los objetivos en términos de minería de datos fueron:

- Identificar las variables de interés para la creación de un *dataset* multidimensional a partir del mapeo sistemático y revisión de la literatura.
- Analizar la relación de las variables independientes consideradas para este índice y la variable dependiente (Incidencia).
- Predecir el valor de incidencia de las 24 ciudades principales del país para los trimestres que se encuentren dentro del conjunto de datos de prueba.

##### **4.1.3.1. Criterios de éxito de minería de datos**

Se estableció como criterio de éxito desde el punto de vista de la minería de datos la posibilidad de realizar predicciones de la incidencia de la COVID-19 en las 24 ciudades principales de Colombia.

#### **4.1.4. Plan del proyecto**

El plan del proyecto incluye el tiempo de elaboración del índice inicial (Capítulo 3) y el índice multidimensional (Capítulo 4), la distribución temporal se encuentra en la sección 3.1.4. Figura 2.

## **4.2. Comprensión de los datos**

Al igual que en índice inicial, en esta fase se realizó la captura inicial, descripción y exploración de los datos.

### **4.2.1. Lectura de datos**

Los datos abiertos recolectados a partir de las diferentes fuentes de datos estaban en formato CSV. Sin embargo, para los datos de vacunación, PIB y desempleo fue necesario crear los *datasets* de forma manual. Lo anterior, debido a que no se encontraron *datasets* que brindaran la información requerida, para posteriormente ser leídos mediante la librería *Pandas* de *Python*.

#### 4.2.2. Información de fuentes adicionales

Para localizar las diferentes fuentes de datos adicionales, se partió de las variables seleccionadas, las cuales se mencionan en la sección de resultados del Capítulo 2 de este documento. Posterior a ello, fueron consultados datos de acceso público, que permitieran obtenerlos. En la consulta realizada se encontró que la mayoría de las fuentes de datos para las variables seleccionadas (PIB, desempleo, porcentaje de vacunación, datos de movilidad, temperatura, precipitación, vulnerabilidad de la COVID-19) presentaban información solo de las ciudades principales y áreas metropolitanas en periodos trimestrales o anuales, por lo que, con base a esa restricción, se decidió trabajar con los datos de esas ciudades en particular, con un periodo trimestral, para tener un *dataset* multidimensional. Las ciudades capitales que se consideraron para la realización del índice fueron: Quibdó, Cali, Cúcuta, Armenia, Popayán, Ibagué, Neiva, Florencia, Valledupar, Tunja, Riohacha, Bogotá, Villavicencio, Pereira, Manizales, Medellín, Santa Marta, Sincelejo, Montería, Pasto, Bucaramanga, Barranquilla, Cartagena, San Andrés.

Los datos localizados fueron:

- Producto Interno Bruto (PIB): El PIB es la “Medida estándar del valor agregado creado mediante la producción de bienes y servicios en un país durante un periodo determinado” [52]. Estos datos dan una visibilidad más amplia de lo que aporta cada uno de los Departamentos al país cada año. Considerando el periodo de tiempo relevante de la pandemia de Covid-19, se tuvo en cuenta los datos de los años 2020 y 2021. Para el presente trabajo se utilizó el PIB a precios constantes, buscando una periodicidad trimestral.

Para el *dataset* de PIB fue necesario crear un archivo CSV utilizando la herramienta *Jupyter Notebook*, debido a que la información se encontraba en un archivo EXCEL, el cual estaba compuesto por las columnas: codDepartamento, codCiudad, departamento, ciudad, año, trimestre y PIB. En cada una de las filas se colocaron los datos respectivos para cada ciudad. Sin embargo, es importante mencionar que antes de colocar el valor de cada ciudad fue necesario realizar un cálculo de lo que aportaba cada una de estas. Lo anterior, debido a que los datos del PIB se encontraron por Departamento, por lo que, se debió considerar el número total de habitantes por Departamento, el número total de habitantes por ciudad, y teniendo en cuenta

el valor del PIB departamental, se implementó una regla de 3 (Regla que permite resolver problemas a partir de 3 valores conocidos y una incógnita [53]) para así hallar el valor que aportaba al PIB cada ciudad.

- Datos climatológicos. Los datos climatológicos han sido identificados como factores que incrementan el riesgo de la COVID-19. Por lo anterior, se buscaron datos de temperatura y precipitación. Para estas variables se encontraron inicialmente, datos recogidos por sensores que han instalado diferentes organismos en el país, pero lamentablemente los datos pertenecían solamente a algunas ciudades. Por lo anterior, se procedió a buscar alternativas como *Google Earth Engine*, la cual permite acceder a los datos de temperatura y precipitación de todos los municipios principales en Colombia [54].

*Google Earth Engine* es un sitio web que tiene información de diversas variables, como temperatura, o precipitación a través del mundo, según el satélite que se escoja. Esta información se encuentra con periodicidad diaria y se puede extraer para los municipios que se deseen. Considerando lo anterior, se procedió a obtener los datos de temperatura y precipitación para las 24 ciudades mencionadas, promediando los datos obtenidos para llegar a una periodicidad trimestral. La decisión de usar el promedio para obtener un valor por trimestre para las variables climatológicas, fue tomada considerando que para los datos de temperatura y precipitación mensuales o anuales de las ciudades siempre se analiza la media o el promedio [55 - 56].

El *dataset* de precipitación tiene una dimensionalidad de 1.096 filas y 3 columnas, la variable precipitación es de tipo de dato *float64*, mientras que, las variables *system\_time\_start* (fecha) y Ciudad son de tipo *object*, este *dataset* no presenta datos nulos.

El *dataset* de temperatura tiene una dimensionalidad de 1.096 filas y 3 columnas, la variable *LST\_Day\_1km*, correspondiente al valor de la temperatura, es de tipo de dato *float64* y contiene 702 valores nulos, por otro lado, las variables *system\_time\_start* (fecha) y Ciudad son de tipo *object* y no tiene valores nulos.

- Información vacunación: Para los datos de vacunación se tuvo en cuenta la información proporcionada por el Ministerio de Salud, en la cual se presenta un informe realizado en *Power BI* [57] , en el cual se puede visualizar la curva

de porcentaje de vacunación para cada uno de los municipios de Colombia con una periodicidad mensual. Para el presente trabajo se tomaron los datos de las 24 ciudades capitales con las que se determinó trabajar, buscando una periodicidad trimestral.

Para el *dataset* de vacunación fue necesario crear un archivo CSV utilizando la herramienta *Jupyter Notebook*, el cual estaba compuesto por las columnas: `codDepartamento`, `codCiudad`, `departamento`, `ciudad`, `año`, `trimestre` y `porcentaje de vacunación`. En cada una de las filas se colocaban los datos correspondientes para cada ciudad en base a los datos proporcionados por el Ministerio de Salud.

- Porcentaje de desempleo: Estos datos fueron recolectados a partir de la información que es publicada por la Gran Encuesta Integrada de Hogares (GEIH) realizada por el DANE, la cual brinda información relacionada con mercado laboral, ingresos y pobreza monetaria, y de las características sociodemográficas de la población residente en Colombia [58]. En esta información se puede identificar la tasa de desempleo de las 24 ciudades capitales para las cuales se hizo el índice de vulnerabilidad multidimensional con una periodicidad trimestral.

Para el *dataset* de desempleo fue necesario crear un archivo CSV utilizando la herramienta *Jupyter Notebook* debido a la que información se encontraba en un archivo *Portable Document Format* (PDF), el cual estaba compuesto por las columnas: `codDepartamento`, `codCiudad`, `departamento`, `ciudad`, `año`, `trimestre` y `porcentaje desempleo`. En cada una de las filas se colocaron los datos correspondientes para cada ciudad en base a los datos proporcionados en los informes del DANE.

- Datos de movilidad: Los datos de movilidad fueron recolectados a partir de los informes de movilidad comunitaria publicados por *google*. Estos informes permiten trazar la tendencia de movimiento a lo largo del tiempo en diversas categorías de lugares como: tiendas, parques, lugar de trabajo, etc, tomando como referencia la movilidad de las 5 semanas comprendidas entre el 3 de enero y el 6 de febrero de 2020, es decir que a partir de los datos de movilidad de esos días es que se calculan los porcentajes de incremento o decremento de movilidad en un área en específico [59]. Este *dataset* contenía información a nivel mundial en la cual se podía encontrar la información de Colombia, en

cada uno de sus departamentos y municipios principales con una periodicidad diaria. Los datos fueron ajustados en cuanto a su periodicidad (trimestral) para que se alineara a las demás fuentes de datos.

Este *dataset* cuenta con una dimensionalidad 9.951.244 filas y 15 columnas.

- Vulnerabilidad COVID-19: Para este índice nuevamente se consideraron los datos de vulnerabilidad de la COVID-19 brindados por el DANE, la recolección de datos se menciona en el punto 3.2.2. de este documento.

Este *dataset* cuenta con una dimensionalidad de 504.996 filas y 9 columnas, las cuales variables como *COD\_DPTO*, *COD\_MPIO*, *COD\_DANE* no presentan datos nulos y son de tipo de dato *object* al igual que *LABEL* que tiene 97.719 datos nulos, *embarazo\_a* con 101.080 datos nulos y *reactivaci* con 133.946; las variables *CATEGORIA* e *ipm* son de tipo *float64* y no contienen datos nulos, y la variable *geometry* es de tipo *geometry* y no presenta datos nulos.

- Datos de COVID-19: Los datos de COVID-19 se obtuvieron a partir de la información publicada por el Ministerio de Salud, en la cual se pueden encontrar los datos diarios reportados en cada uno de los municipios de Colombia en formato CSV [8].

Este *dataset* cuenta con una dimensionalidad de 6.314.769 filas y 23 columnas,

## 4.3. Preparación de los datos

### 4.3.1. Limpieza y pre-procesamiento de datos

La limpieza de datos es fundamental para cualquier *dataset* puesto que estos comúnmente, no cuentan con los datos expresados de la manera en que se requiere. En el Anexo C (denominado “Repositorio de GitHub” en la carpeta multidimensional *index*) se encuentra la preparación de los datos llevada a cabo para el índice propuesto.

#### 4.3.1.1. Limpieza y pre-procesamiento del *dataset* “temperatura”

Este *dataset* estaba compuesto por dos filas: valor de temperatura promedio diaria y el día al cual correspondía el valor, estos datos se encontraban dentro de un archivo en formato CSV, para cada una de las ciudades de interés. Fue necesario inicialmente realizar un promedio para calcular el valor de la temperatura por trimestre. Posterior

a ello, crear la columna trimestre con el objetivo de relacionar el valor calculado con el trimestre al que correspondía; para este caso el trimestre podía tener el valor 1, haciendo referencia a los datos de los meses de enero a marzo, el valor 2 para los datos de abril a junio, el valor 3 indicando los datos de julio a septiembre, o el valor 4 para los datos de octubre a diciembre. Por último, se crearon dos columnas para relacionar el valor de la temperatura y el trimestre con el municipio al cual correspondían esos datos, considerando que se debía asociar con el código de municipio que maneja la divipola (“estándar nacional que codifica y lista las entidades territoriales” [60]), y el nombre del municipio, obteniendo así un *dataset* de 4 columnas por 288 filas.

#### **4.3.1.2. Limpieza y pre-procesamiento del *dataset* “precipitación”**

Este *dataset* estaba compuesto por dos filas: valor de precipitación diaria y el día al cual correspondía el valor, estos datos se encontraban dentro de un archivo en formato CSV, para cada uno de los municipios considerados. Fue necesario inicialmente realizar un promedio para calcular el valor de la precipitación por trimestre. Posterior a ello, crear la columna trimestre con el objetivo de relacionar el valor calculado con el trimestre al que correspondía (tal como se explica en el punto 4.2.2). Por último, se crearon dos columnas para relacionar el valor de la precipitación y el trimestre con el municipio al cual correspondían esos datos, considerando que se debía asociar con el código de municipio que maneja el divipola y el nombre del municipio, obteniendo así un *dataset* de 4 columnas por 276 filas; esto debido a que para San Andrés no había datos para esta variable.

#### **4.3.1.3. Pre-procesamiento del *dataset* “vacunación”**

Debido a que este *dataset* tuvo que ser creado manualmente, lo único que se realizó fue la eliminación de los datos del año 2022, puesto que, no todas las variables tenían datos para ese año, lo que ocasionaba que se tuvieran valores nulos en el *dataset*. Este *dataset* contiene información del 2021 y los 3 primeros trimestres del 2022, para el año 2020 todos los porcentajes fueron 0 debido a que en ese momento aún no se iniciaba la vacunación en Colombia. Finalmente se obtuvo un *dataset* de 7 columnas por 264 filas.

#### **4.3.1.4. Pre-procesamiento del *dataset* “PIB”**

Debido a que este *dataset* tuvo que ser creado manualmente no fue necesario ningún pre-procesamiento. Este *dataset* contiene información de 2020 y 2021. Finalmente se obtuvo un *dataset* de 7 columnas por 192 filas.

#### **4.3.1.5. Pre-procesamiento del *dataset* “Desempleo”**

Debido a que este *dataset* tuvo que ser creado manualmente no fue necesario ningún pre-procesamiento. Este *dataset* contiene información de 2019, 2020 y 2021. Finalmente se obtuvo un *dataset* de 7 columnas por 288 filas.

#### **4.3.1.6. Pre-procesamiento del *dataset* “Movilidad”**

Para el *dataset* de movilidad se tomó el reporte global que brinda la página de *Google mobility* y se procedió a filtrar por el país (Colombia). Después de ello, se filtraron las ciudades de interés para así tener los datos deseados y luego, se realizó el promedio para tener los valores en una periodicidad trimestral. Para este caso las columnas del *dataset* fueron: *country\_region\_code*, *country\_region*, *sub\_region\_1*, año, trimestre, *codDepartamento*, *codCiudad*, *parks\_percent\_change\_from\_baseline*, *transit\_stations\_percent\_change\_from\_baseline*, *workplaces\_percent\_change\_from\_baseline*, *residential\_percent\_change\_from\_baseline*.

Finalmente, se obtuvo un *dataset* de 11 columnas por 192 filas.

#### **4.3.1.7. Pre-procesamiento del *dataset* “Vulnerabilidad COVID-19”**

Para el *dataset* de vulnerabilidad de COVID-19 se tuvo en cuenta el proceso que se menciona en la sección 3.3.1.9. de este documento, pero antes de descargar el CSV se procedió a filtrar solo por las ciudades para los cuales se iba a realizar el índice multidimensional. Es importante recalcar que para este índice, la vulnerabilidad al COVID-19 es una variable independiente.

#### **4.3.1.8. Pre-procesamiento del *dataset* “Datos de la COVID-19”**

Para el pre-procesamiento de este *dataset* se procedió a filtrar por las 24 ciudades de interés y se realizó una suma de los casos que se presentaban en cada trimestre, esto debido, a que el valor de casos de COVID-19 en Colombia es un valor creciente con el tiempo. Posteriormente, considerando que los datos obtenidos en las diferentes ciudades no son comparables como consecuencia de la diferente cantidad de

población de cada una, se procedió a obtener la incidencia de casos de COVID-19. Para el cálculo de la incidencia, fue necesario tener en cuenta la cantidad de población por ciudad, para así hacer una relación entre casos de COVID-19 y la cantidad población.

#### **4.3.2. Integración de los *datasets***

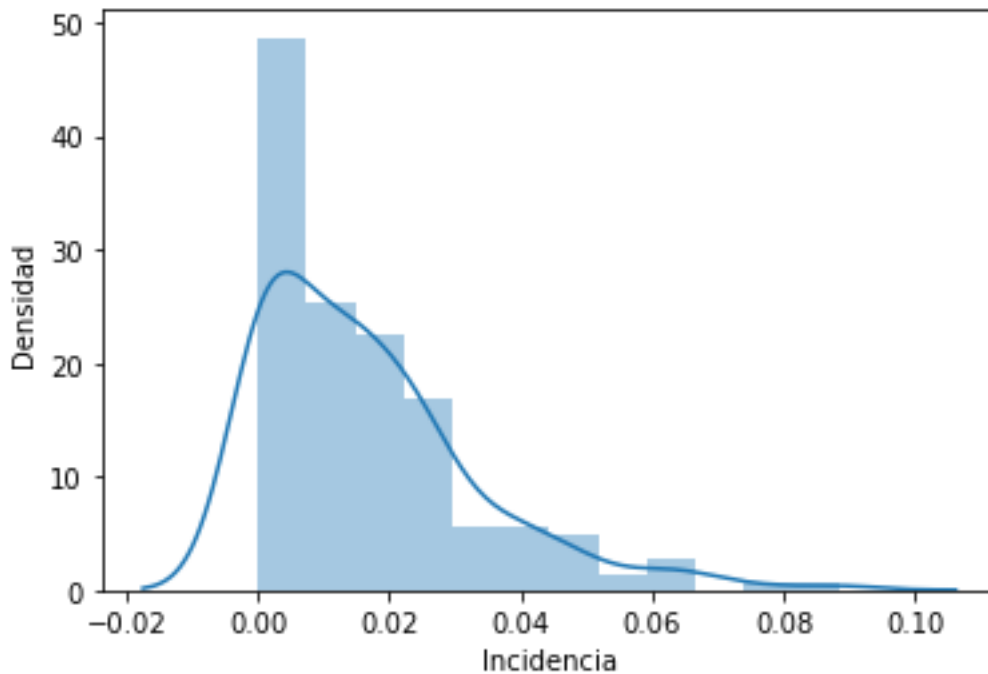
Después de realizar el pre-procesamiento de cada uno de los *datasets* se procedió a unirlos considerando el código divipola para así, tener diferentes variables para cada uno de los municipios. Es importante recalcar que los datos de 2022 finalmente no fueron considerados, ya que, para algunas variables aún no habían sido publicados los valores de ese año, como ocurrió con la variable de desempleo y PIB. Además, también fueron descartados los datos de 2019 debido a que el *dataset* de la variable dependiente (incidencia de la COVID-19) contenía información a partir del 2020. En el Anexo E denominado “*Dataset* índice multidimensional” se encuentra el *dataset* que se obtuvo después del pre-procesamiento realizado.

### **4.4. Análisis Exploratorio de Datos (EDA)**

En cuanto al análisis exploratorio de datos (EDA), se realizó a partir del *dataset* multidimensional creado en la integración de datos. Este *dataset* contiene una dimensionalidad 16 columnas y 192 filas, donde cada ciudad principal de Colombia contiene información trimestral para los años 2020 y 2021. Todas sus columnas son tipo de dato numérico (13 de ellas son *float64* y 3 son *int64*). Este *dataset* no cuenta con datos nulos debido al pre-procesamiento hecho antes de la unificación de los *datasets*.

Realizando un análisis uni-variable para la variable objetivo “Incidencia” (con el fin de estudiarla con mayor detenimiento debido a que esta será la variable a predecir), se puede decir que consta de 192 datos cuyos valores son porcentuales decimales entre 0 y 1. El promedio de los datos es 0.016642, presentan una desviación estándar de 0.016670, y el rango de los valores está entre 0.000000 y 0.088639. En la Figura 10 se muestra el Histograma generado de estos datos.

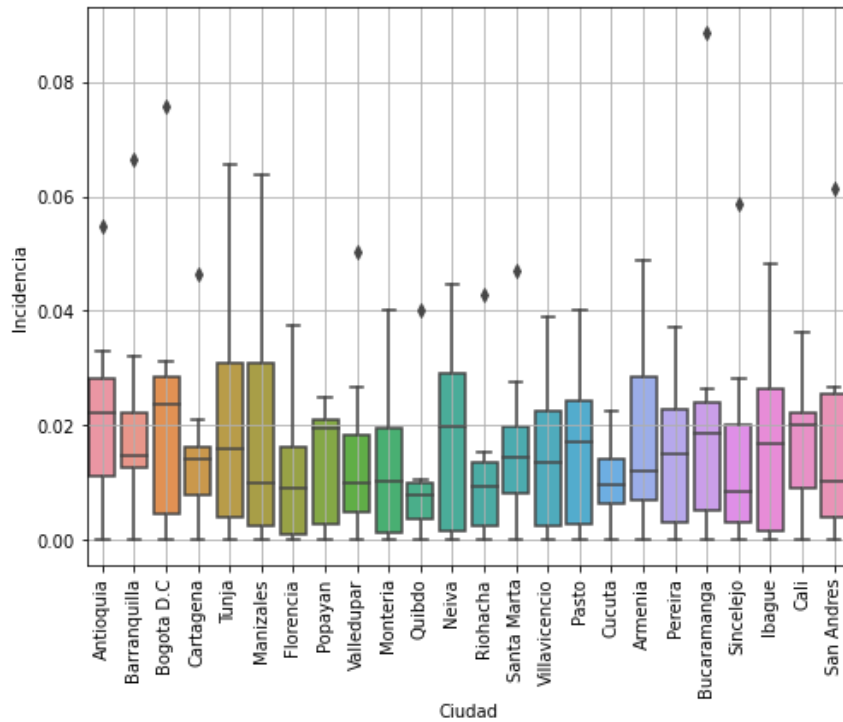




**Figura 10.** Densidad de los datos para la variable objetivo incidencia.

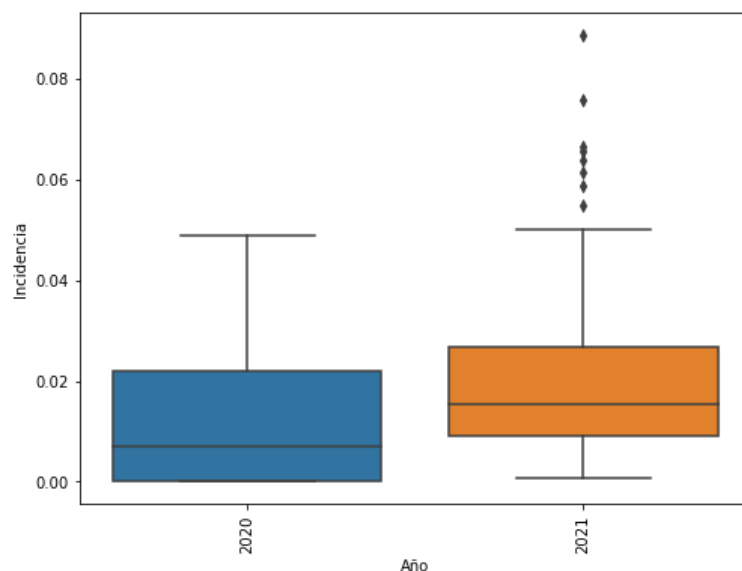
De la Figura 10 se observa que se presenta una desviación con respecto a la distribución normal. Calculando la asimetría se obtuvo un valor de 1.435701, indicando que la distribución se encuentra sesgada hacia la izquierda. Además, el valor de la curtosis fue 2.43768, al ser un valor positivo indica que maneja una distribución leptocúrtica debido a que los datos se encuentran concentrados hacia la media [61].

La relación de las variables independientes y la variable dependiente ayuda a conocer la distribución de los datos, en la Figura 11 mediante un gráfico de cajas se observa la distribución de los datos para las 24 ciudades principales. La Figura 12 presenta la distribución de los datos para los años 2020 y 2021 respecto a la incidencia, mientras que la Figura 13 presenta el valor de vulnerabilidad respecto a la incidencia.



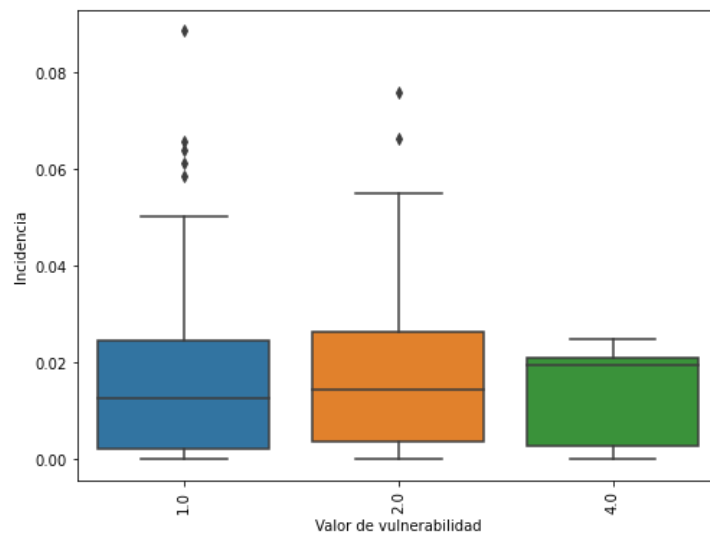
**Figura 11.** Gráfico de cajas para las 24 ciudades principales.

En la Figura 11 se observa que para las ciudades como Tunja, Florencia y Santa Marta la mediana se encuentra en el centro o muy cerca del centro de la caja indicando que la media, mediana y moda coinciden o son muy parecidas. Sin embargo, para las demás ciudades se presenta una asimetría ya sea positiva (si la parte más larga de la caja es la parte superior), caso de ciudades como Manizales o Armenia; o negativa (si la parte inferior de la caja es más larga que la superior, lo que significa que los datos se concentran en la parte superior de la distribución, caso) de ciudades como Bogotá D.C. y Neiva.



**Figura 12.** Gráfico de cajas para los la variable año respecto a la incidencia.

Respecto al gráfico de la Figura 12, se observa que para el año 2021 la caja se encuentra más arriba que para el año 2020, indicando que los rangos de valores para esta variable aumentaron debido a una mayor incidencia en los casos de COVID-19 confirmados. Sin embargo, presenta una asimetría positiva indicando que los datos para este año se centran en la parte inferior de la distribución al igual que el año 2020. Por otra parte, para el año 2021 se presentan una gran cantidad de datos atípicos.



**Figura 13.** Gráfico de cajas para la variable valor de vulnerabilidad respecto a la incidencia.

En la Figura 13 se observa que la variable Vulnerabilidad\_numero para las 24 ciudades principales del país presenta solo 3 de las 5 categorías (baja, baja-media y media-alta), debido a que para estas ciudades no se presentaron datos del valor de la vulnerabilidad de 3 y 5. Además, para los valores de vulnerabilidad 1 y 2 se presentan rangos de datos similares, aunque para el valor 1 se presentan más valores atípicos. Para los valores de vulnerabilidad 1 y 2 se observa una distribución simétrica dado que la mediana se encuentra justo en el centro. Sin embargo, para el valor de vulnerabilidad 4 se presenta una asimetría negativa o sesgada a la izquierda debido a que los datos se concentran en la parte superior de la distribución.

Continuando con el EDA, los gráficos de dispersión se usan para investigar sobre la intensidad de la relación entre dos variables, sobre el eje X se ubica la variable independiente y sobre el eje Y la variable dependiente (incidencia de la COVID-19). Las Figuras 14 – 26 presentan los gráficos de dispersión obtenidos para cada una de

las variables; trimestre, porcentaje desempleo, temperatura, precipitación, PIB, *retail and recreation, grocery and pharmacy, parks, transit stations, workplaces, residential*, porcentaje de vacunación y vulnerabilidad.

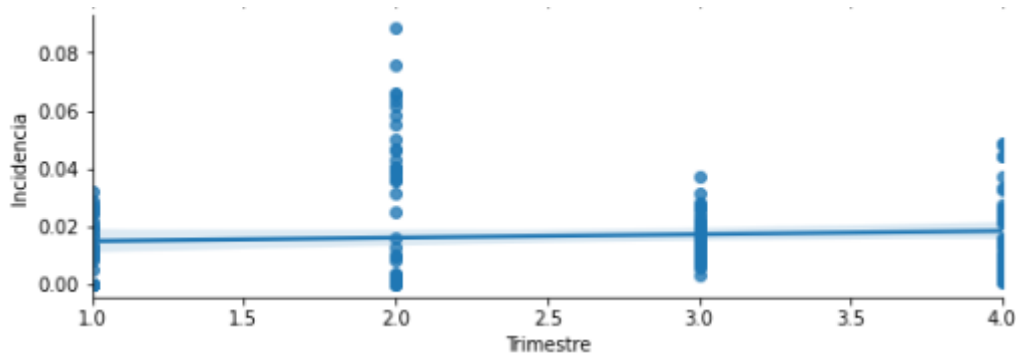


Figura 14. Gráfico de dispersión para trimestre.

Del gráfico de la Figura 14 se puede observar que en el trimestre 2 existe un aumento en la incidencia de la COVID-19. Por otra parte, en el trimestre 1 se puede visualizar que se tiene la incidencia más baja. Lo anterior permite notar que los datos de cada trimestre y la periodicidad de estos, son un factor importante para predecir la incidencia. Cabe mencionar que este comportamiento no refleja explícitamente la realidad de los casos diarios de COVID-19, lo cual se debe a que se utilizó una periodicidad trimestral. Dicha periodicidad fue considerada la más adecuada, debido a la periodicidad de los datos para todas las variables que componen el *dataset* multidimensional.

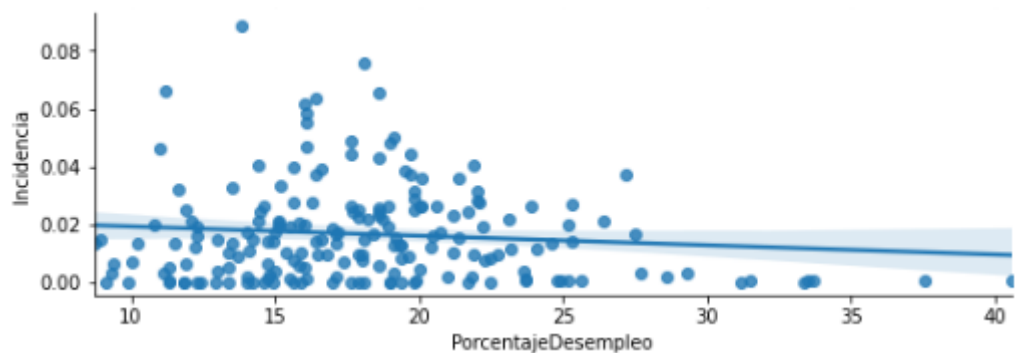
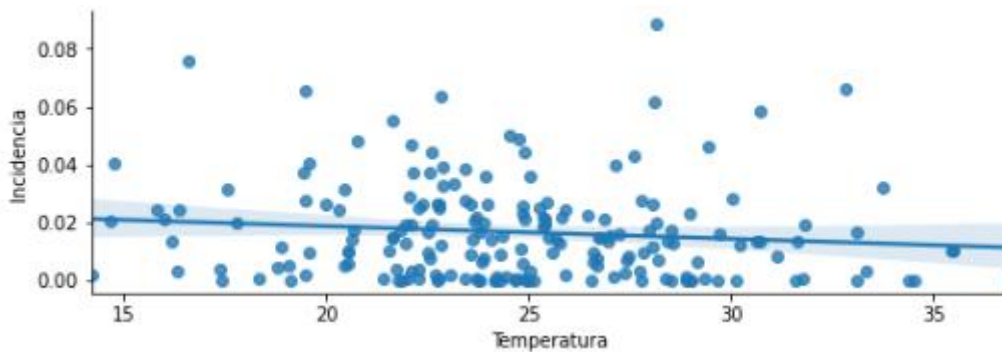


Figura 15. Gráfico de dispersión para porcentaje de desempleo.

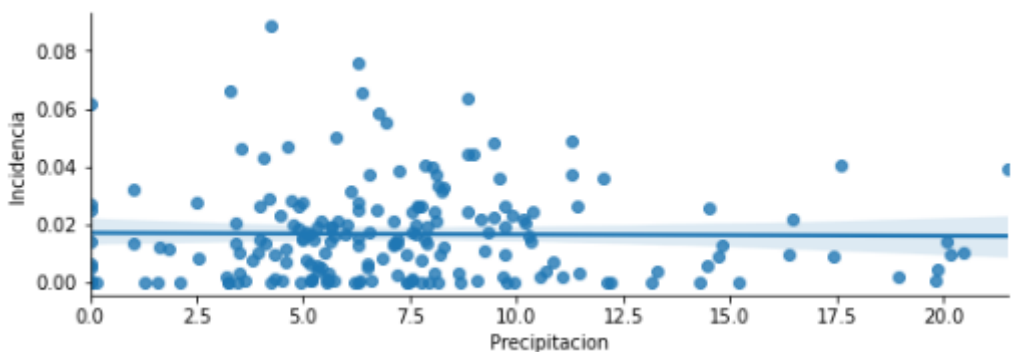
En la Figura 15 se observa que para las ciudades donde el porcentaje de desempleo se encuentra entre el 15% y 20% la incidencia presenta valores entre 0.00 y aproximadamente 0.08, alcanzando en ese rango, valores altos de incidencia. Por otro lado, es posible notar que a partir de aproximadamente el 28% de tasa de desempleo las ciudades tienen una incidencia baja. Sin embargo, hay muchos otros

puntos que se encuentran por debajo de 0.02 de incidencia, lo cual no permite observar una tendencia clara.



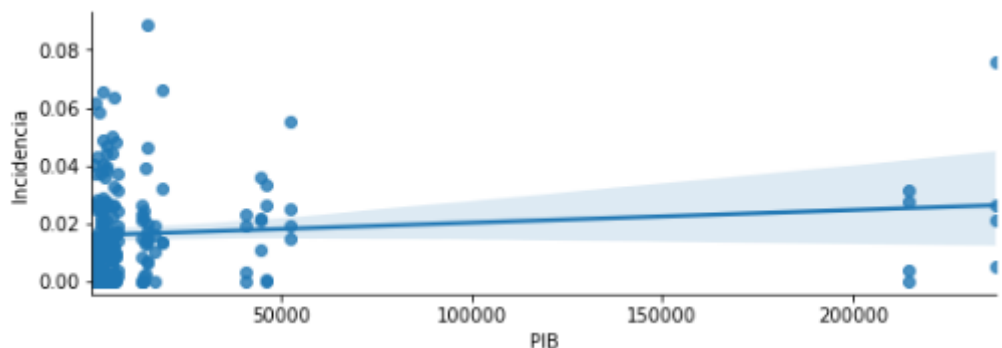
**Figura 16.** Gráfico de dispersión para temperatura.

El gráfico de dispersión de la Figura 16, realmente no permite visualizar una correlación clara entre la variable temperatura y la incidencia, dado que, los puntos no siguen ninguna tendencia, presentando así un gráfico en el cual no hay una relación lineal.



**Figura 17.** Gráfico de dispersión para precipitación.

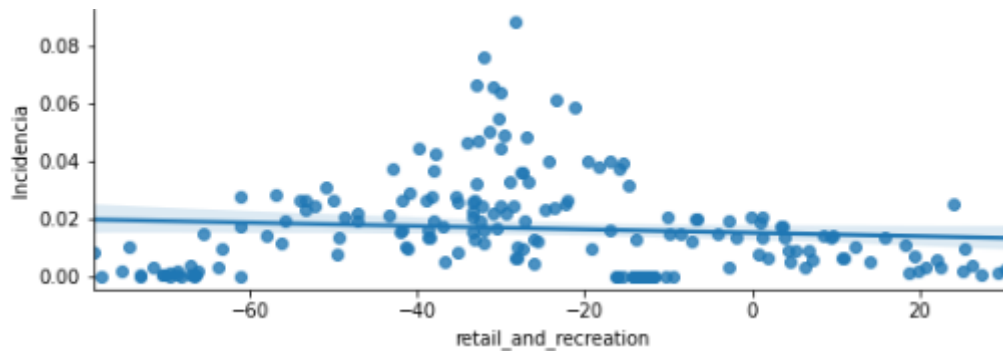
La Figura 17 al igual que la Figura 16, no presenta una tendencia, dificultando determinar la correlación entre la precipitación y la incidencia.



**Figura 18.** Gráfico de dispersión para PIB.

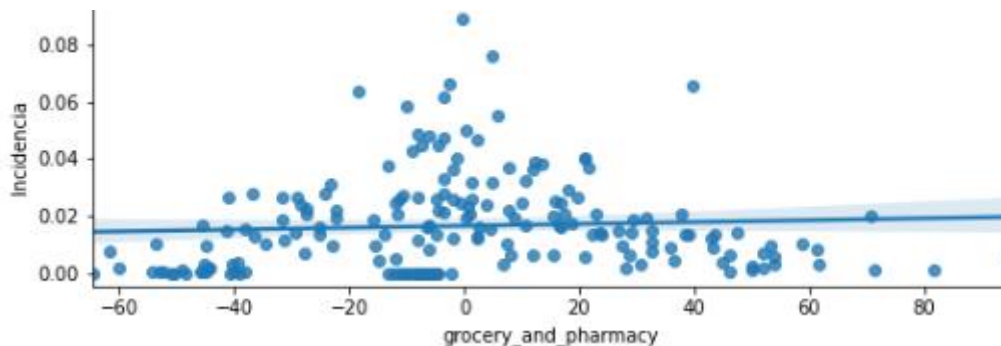
La Figura 18 permite notar que no hay una tendencia entre el PIB y la incidencia, puesto que la mayoría de los puntos se encuentran cerca al origen, pero tienen una incidencia diferente. Aunque, es posible observar que para la mayoría de los puntos

la incidencia es menor a 0.04, lo que indica que para la mayoría de ciudades la incidencia no se ve afectada por el PIB de la ciudad.



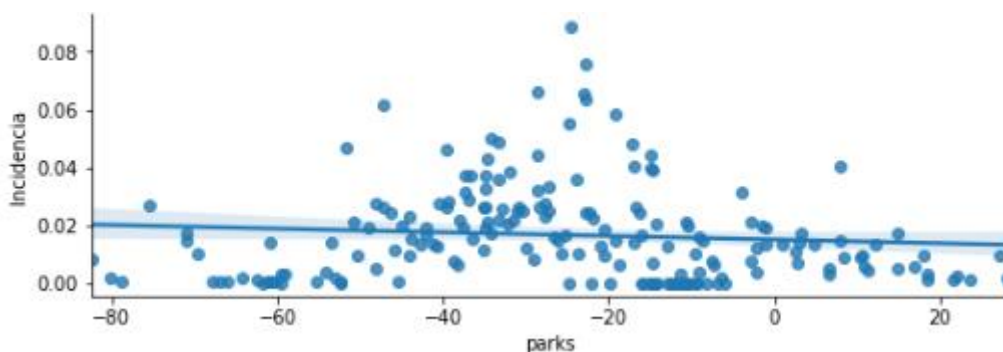
**Figura 19.** Gráfico de dispersión para *retail\_and\_recreation*.

En el gráfico de dispersión de la Figura 19 se visualiza que para los lugares de comercio y ocio cuando el porcentaje de movilidad cayó a valores entre -20% y -40%, la incidencia tomó valores entre 0.01 y 0.09 aproximadamente, obteniendo un máximo de incidencia entre esos porcentajes. Además, se observa una tendencia un poco curvilínea.



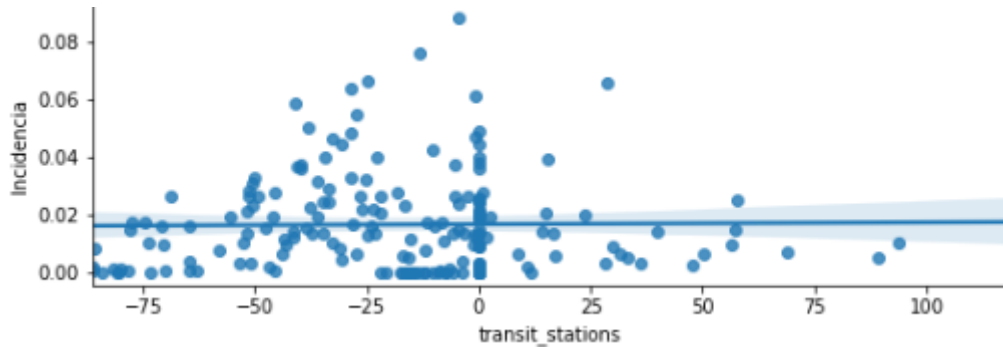
**Figura 20.** Gráfico de dispersión para *grocery\_and\_pharmacy*.

De la Figura 20 se puede notar nuevamente una tendencia algo curvilínea, con algunos puntos fuera de la tendencia. Para los porcentajes de movilidad de entre -20% y 20% se observa que se obtiene un valor mayor de incidencia, el cual es 0.09 aproximadamente.



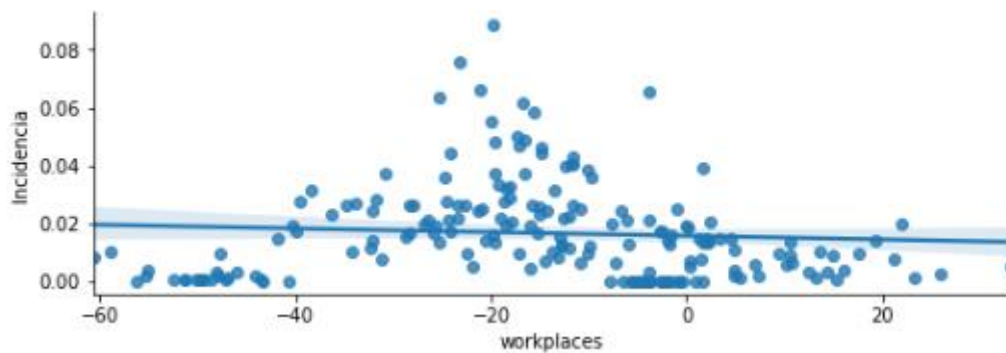
**Figura 21.** Gráfico de dispersión para *parks*.

En la Figura 21, se observa el gráfico de dispersión de movilidad en parques, en el cual se analiza una tendencia un poco curvilínea obteniendo valores altos de incidencia entre -30% y -20% aproximadamente. Sin embargo, hay algunos puntos fuera de la tendencia.



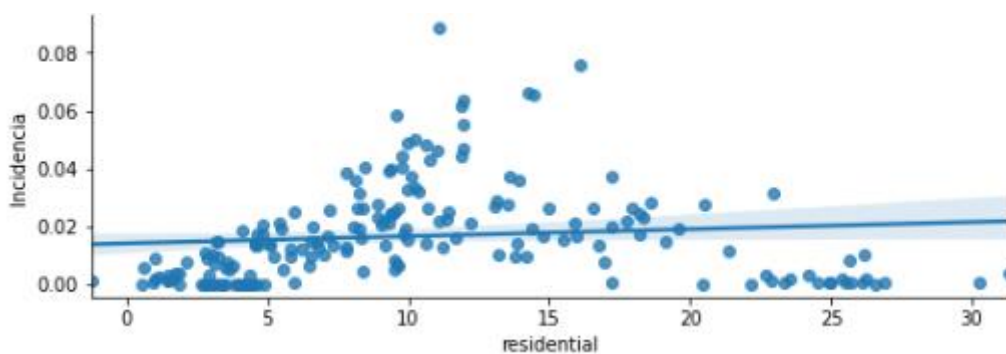
**Figura 22.** Gráfico de dispersión para *transit\_stations*.

En la Figura 22 se nota que para los porcentajes decrecientes de movilidad es donde hay un mayor aumento en la incidencia, obteniendo picos en los valores cercanos a movilidad cero. Esto indica que cuando el porcentaje de movilidad disminuyó, respecto a los datos de “porcentaje movilidad base” en las estaciones de tránsito, hubo un aumento de casos de la COVID-19.



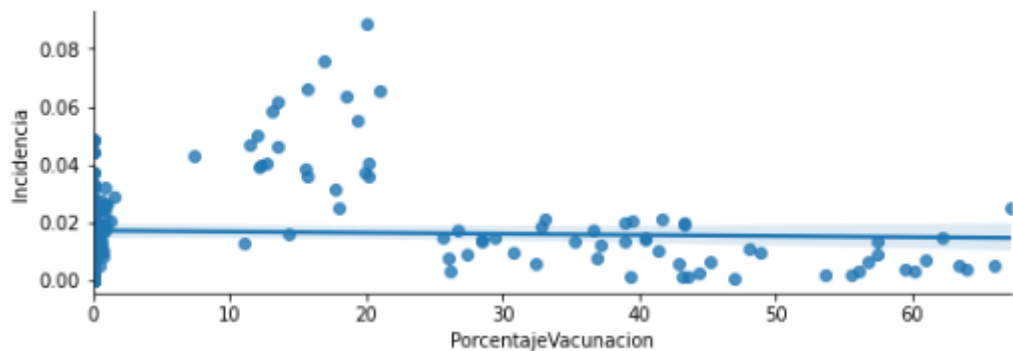
**Figura 23.** Gráfico de dispersión para *workplaces*.

El gráfico de dispersión de la Figura 23 permite visualizar que hay una tendencia curvilínea donde se alcanzan mayores valores de incidencia entre los porcentajes de movilidad de -40% y 0%.



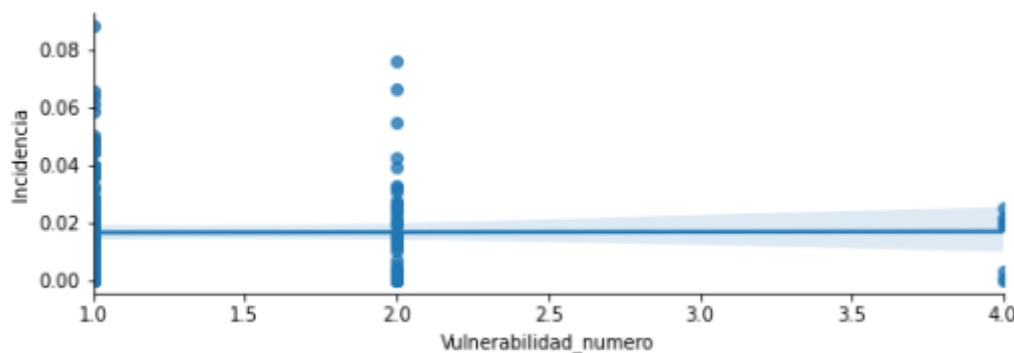
**Figura 24.** Gráfico de dispersión para *residencial*.

En la Figura 24 se observa que inicialmente se presenta una tendencia creciente, sin embargo, hay algunos puntos que no siguen dicha tendencia.



**Figura 25.** Gráfico de dispersión para porcentaje de vacunación.

De la Figura 25, se nota que no hay una correlación lineal entre el porcentaje de vacunación y la incidencia, diferente a la tendencia decreciente que se esperaba, puesto que, a mayor porcentaje de vacunación se esperaría una menor incidencia. Sin embargo, es posible que estén afectando muchos otros factores en cada una de las ciudades que hacen que no se siga una tendencia lineal.



**Figura 26.** Gráfico de dispersión para Vulnerabilidad\_numero.

En la Figura 26, se observa que para las opciones de vulnerabilidad\_numero 1 y 2 se presentaba ya fuera baja o alta incidencia, casos en los que se esperaba que la incidencia fuera baja dado que hay un menor nivel de vulnerabilidad.

En los anteriores gráficos de dispersión del modelo multidimensional, se observaron tendencias bastante débiles pero un poco más claras, a diferencia de los gráficos de dispersión del modelo inicial, en los cuales no se observaba ninguna tendencia.

Para realizar un análisis más objetivo, al igual que en el *dataset* inicial se realizaron matrices de confusión midiendo el coeficiente de correlación de *Pearson* para evaluar la relación lineal entre dos variables cuantitativas.



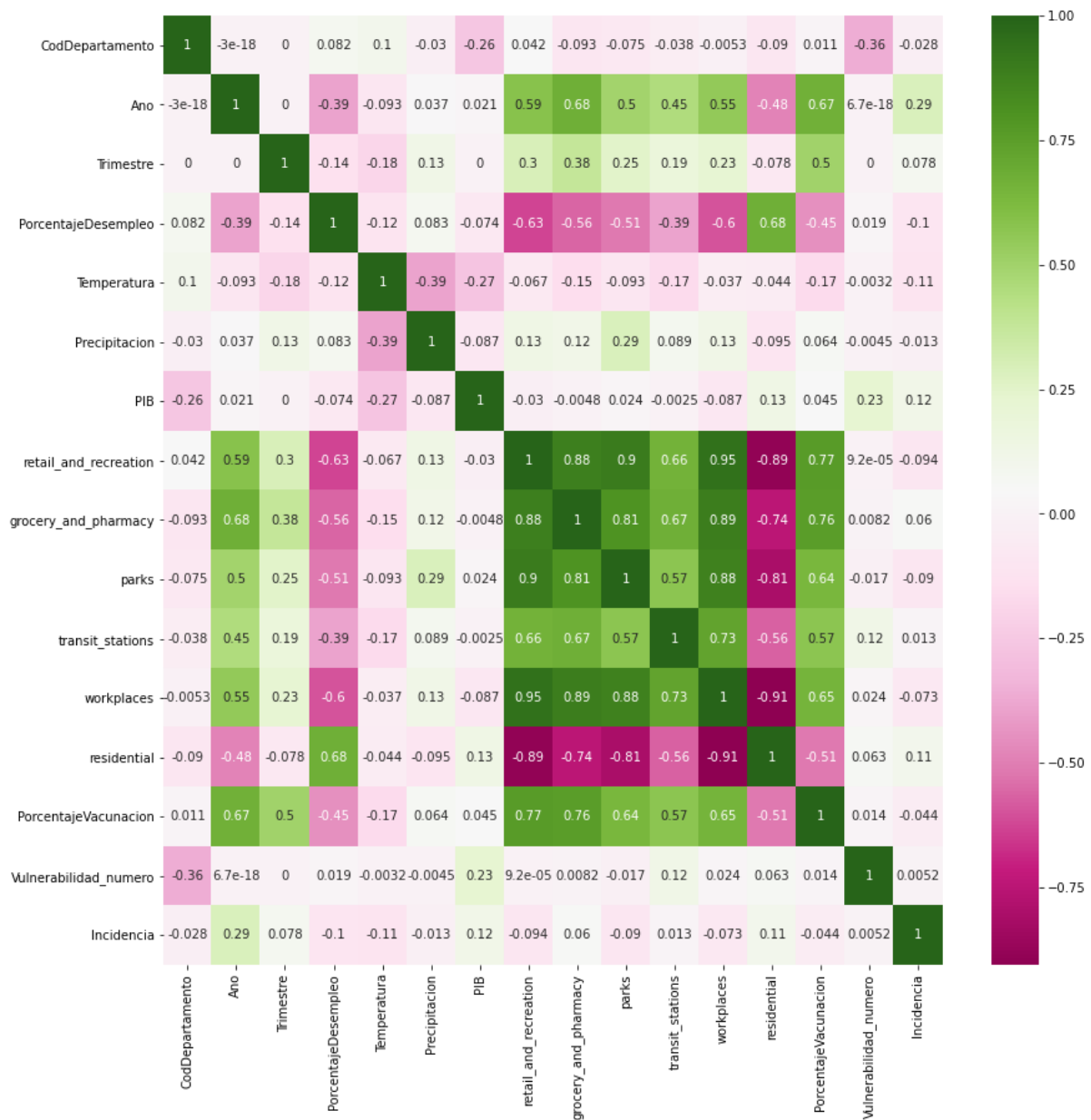


Figura 27. Matriz de confusión (Correlación de Pearson).

En la Figura 27, la columna llamada incidencia es la variable dependiente, por lo que es la variable de interés. Un color verde indica alta correlación positiva, un color fucsia indica alta correlación negativa, y un color blanco indica una correlación baja. La Figura 27 permite observar que variables como PIB y “residential” (de movilidad) tienen una correlación positiva débil; la variable año tiene una correlación positiva moderada, mientras que la variable temperatura presenta una correlación negativa débil.

Con el objetivo de conocer si se presenta una relación monótona entre las variables independientes y la dependiente, es decir, que las variables tiendan a cambiar en el

mismo instante de tiempo, se calculó la correlación de *Spearman*, los resultados se presentan en la Figura 28.

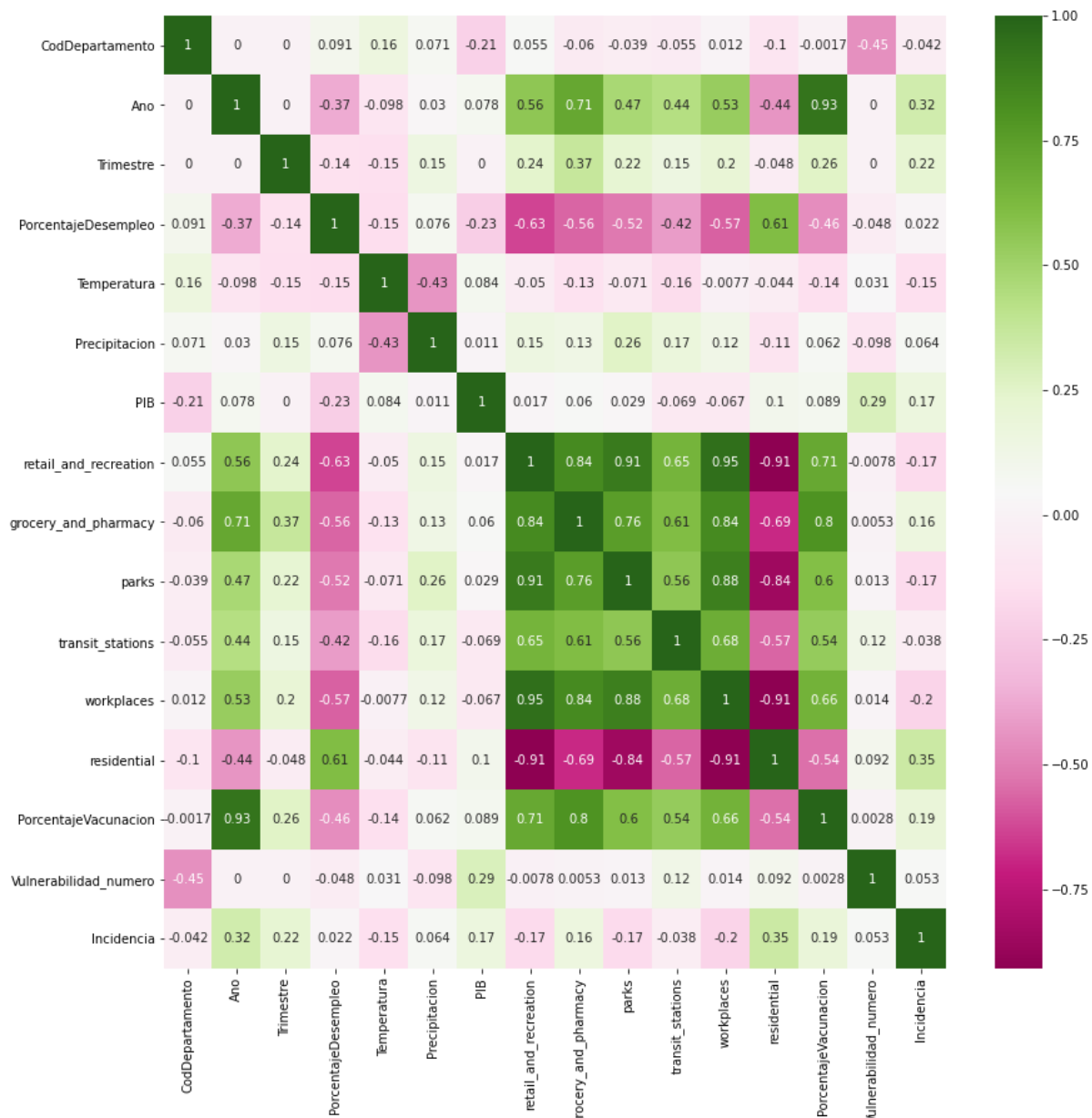


Figura 28. Matriz de confusión (Correlación de Spearman).

Al igual que en la Figura 27, en la Figura 28 la última columna muestra algunos tonos verdes y otros fucsias, esto indica que con esas variables se presenta correlación, ya sea positiva si el color es verde, o negativa si el color es fucsia, a mayor intensidad del color, mayor correlación. Considerando lo anterior, se puede observar que variables como el trimestre, PIB, *grocery and pharmacy* y porcentaje vacunación tienen una correlación débil positiva con la variable incidencia, valores entre 0.16 y 0.22. En cambio, las variables año con valor 0.32 y *residential* con valor 0.35 presentan una correlación moderada positiva respecto a la incidencia. Además, las

variables temperatura, *retail\_and\_recreation*, *parks* y *workplaces* tienen una correlación débil negativa. Las demás variables muestran una correlación muy cercana a cero indicando que son datos sin correlación.

## 4.5. Modelado y evaluación

### 4.5.1. Modelado

Para el modelado se aplicaron diversos algoritmos de aprendizaje supervisado de regresión, debido a que la variable dependiente, es decir incidencia es un valor continuo. Esta variable tiene valores que se encuentran entre el rango de 0 y 1, por lo que, se tiene un conjunto de infinitos valores. Los algoritmos utilizados fueron: regresión lineal, árboles de decisiones (*Decision Trees*), un algoritmo basado en instancias (*KNN*, *k-Nearest Neighbor*), máquinas de vectores de soporte (*SVM*, *Support Vector Machine*), bosques aleatorios (*Random Forest*), y potenciación del gradiente (*Gradient Boosting*). También se trabajó con dos meta-estimadores llamados *Extra Trees Regression* y *AdaBoost Regressor*. Se tuvieron en cuenta dos escenarios, en el primero se dividió el *dataset* en 80% para entrenamiento y 20% para prueba, y en el segundo escenario se dividió 70% para entrenamiento y 30% para prueba, Los anteriores valores de los escenarios fueron elegidos con el propósito de no generar sobreajuste, o sub-ajuste, además que, son dos tipos de escenarios comúnmente utilizados [43 - 44].

#### 4.5.1.1.1. Regresión lineal

El algoritmo de regresión lineal es un algoritmo que permite realizar predicciones para variables continuas, a partir de la relación lineal entre la variable dependiente y las variables independientes [62]. Los resultados obtenidos para la prueba realizada con este algoritmo se presentan en las Tablas 4 y 5.

#### 4.5.1.1.2. *Decision Tree Regressor* (Regresor del árbol de decisión)

Para la creación del modelo con este algoritmo es importante considerar el número de nodos terminales que se generan a partir del nodo raíz, este concepto es conocido como profundidad del árbol (*max\_depth*). Para este caso, se probaron diferentes valores de profundidad con el objetivo de obtener el mejor resultado. Se obtuvo el mejor resultado con una profundidad de 7, generando un total de 70 nodos terminales. En la Figura 29 se muestra el árbol obtenido, el cual, inicialmente toma la decisión por

la variable *residencial*, y luego, se basa en porcentaje de vacunación, de donde se empiezan a desprender el resto de decisiones. Al ser un árbol con tantos nodos se presenta el mismo árbol dividido en varias figuras (Figuras 30-35), las cuales permiten observar mejor el árbol resultante.

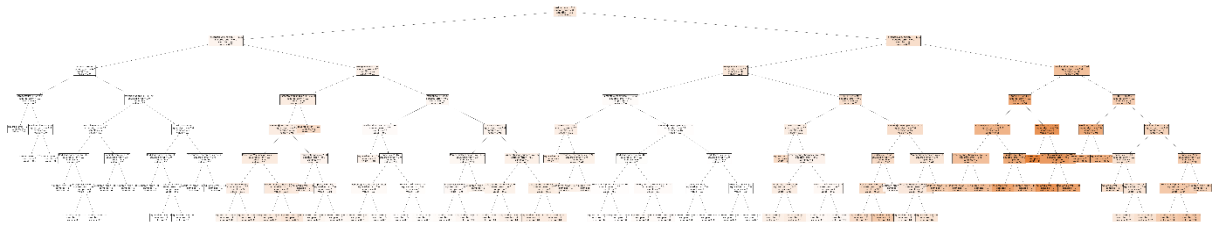


Figura 29. Árbol de decisión obtenido (Profundidad=7).

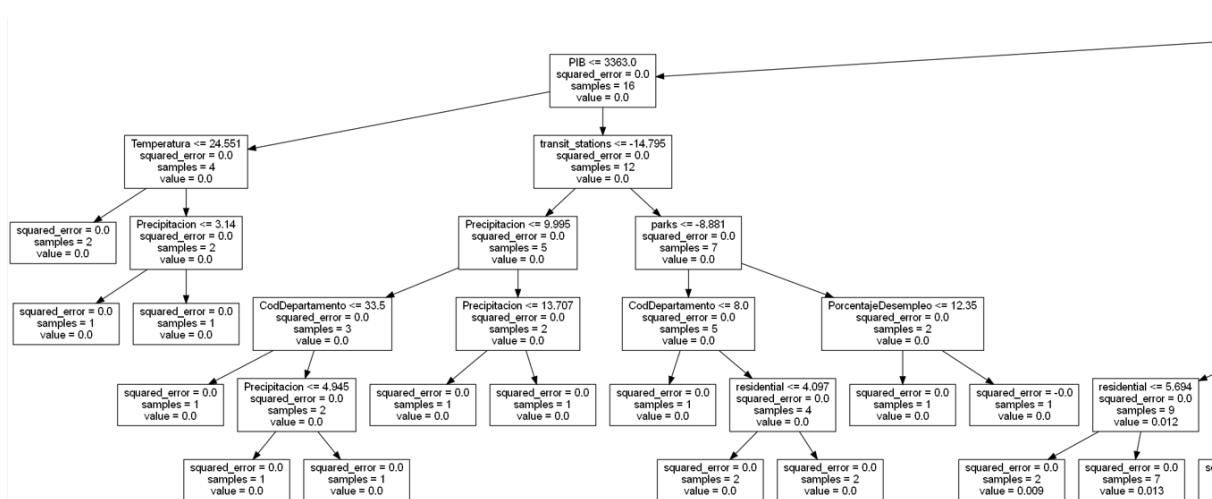


Figura 30. Árbol de decisión obtenido – Parte 1.

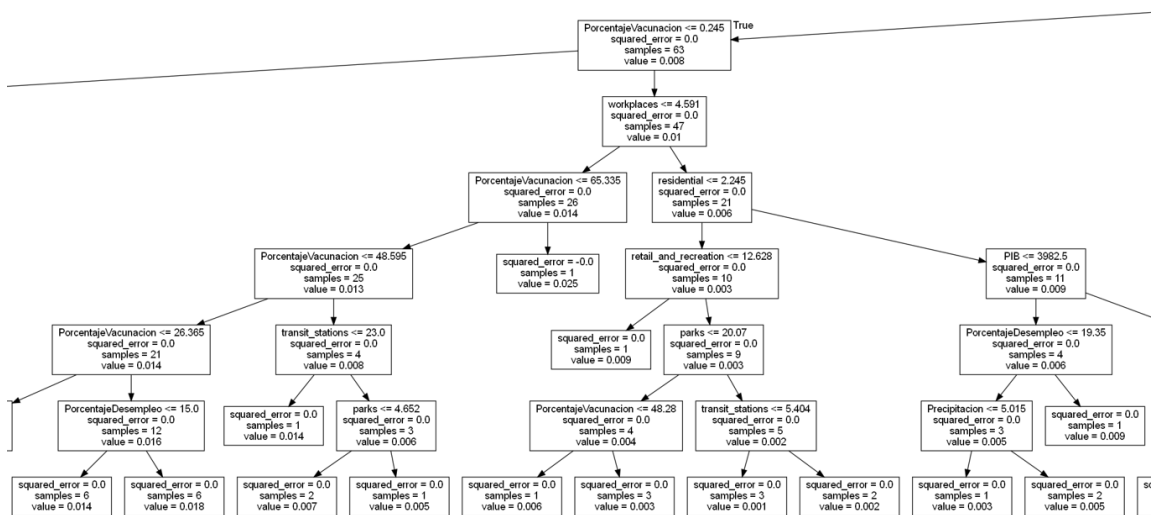


Figura 31. Árbol de decisión obtenido – Parte 2.



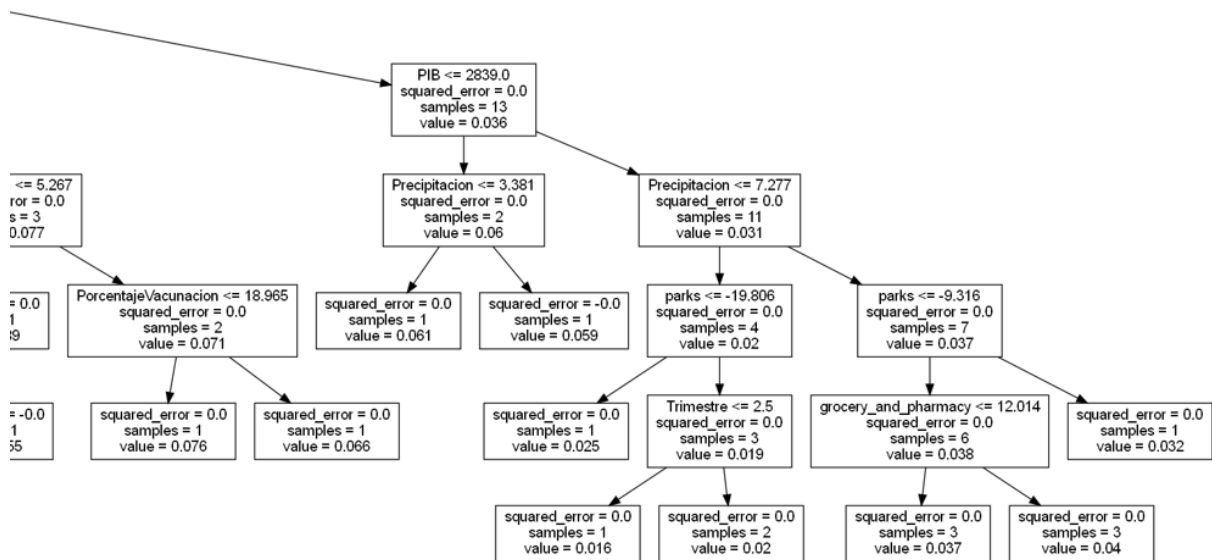


Figura 35. Árbol de decisión obtenido – Parte 6.

Los resultados obtenidos se muestran en las Tablas 4 y 5.

#### 4.5.1.1.3. *K-Nearest Neighbor (K- Vecinos más Cercanos)*

El algoritmo KNN, es un algoritmo de aprendizaje supervisado que puede ser usado para problemas de clasificación o regresión. Para regresión se toma el promedio de los k vecinos más próximos para realizar la predicción [63].

#### 4.5.1.1.4. *SVM, Support Vector Machine (Máquinas de vectores de soporte)*

Para este modelo también fue utilizado el algoritmo SVM, pero utilizándolo para regresión. En las Tablas 4 y 5 se observan los resultados obtenidos.

#### 4.5.1.1.5. *Random Forest (Bosque aleatorio)*

*Random forest* es un algoritmo de aprendizaje automático que une la salida de varios árboles de decisión para llegar a un solo resultado [64].

#### **4.5.1.1.6. Gradient Boosting Regressor (Regresor por potenciación del gradiente)**

El modelo *Gradient Boosting* aplicado para regresión se conforma por varios árboles de decisión individuales, los cuales son entrenados de forma consecutiva, donde cada uno de estos trata de mejorar los errores del árbol anterior [65]. Este modelo utiliza el parámetro *n-estimators* que indica la cantidad de árboles incluidos.

#### **4.5.1.1.7. Extra Trees Regressor**

Este algoritmo ejecuta un meta-estimador el cual se ajusta a una secuencia de árboles de decisión aleatorios en diferentes sub-muestras del grupo de datos, y usa el promedio para mejorar la precisión predictiva y controlar el sobreajuste [66].

#### **4.5.1.1.8. AdaBoost Regressor**

*AdaBoost Regressor* es un meta-regresor el cual entrena un modelo y calcula qué tan buena es la predicción. Posteriormente, ajusta copias adicionales del regresor en el mismo conjunto de datos y nuevamente va calculando qué tan buena es la predicción, pero considerando que los pesos de cada modelo dependen del modelo anterior [67].

### **4.5.2. Evaluación**

Para evaluar la aptitud de los algoritmos se utilizaron dos métricas: la raíz cuadrada del error cuadrático medio (*RMSE*) y R-cuadrado. La métrica *RMSE* indica qué tan cerca están los puntos de datos observados de los valores predichos. También se puede interpretar como la desviación estándar de la varianza inexplicada. Un valor bajo de *RMSE* indica un mejor ajuste. La otra métrica utilizada como referencia principal fue R-cuadrado, la cual indica la aptitud del modelo, es decir, que tan apto o que tan adecuado es el modelo para cierto fin, también es conocida como el coeficiente de determinación [68], [69]. Esta métrica toma valores entre 0 y 1, donde 0 representa que el modelo propuesto no mejora la predicción sobre el modelo medio y 1 indica una predicción perfecta.

Para realizar la evaluación del modelo se tuvo como variable objetivo la incidencia, por lo cual, se evaluó qué algoritmo presentaba mejor desempeño en la predicción de esta variable. En las Tablas 4 y 5 se observan los resultados obtenidos y los parámetros por defecto. Adicionalmente, se realizó la misma evaluación para el índice

de vulnerabilidad propuesto por el DANE considerando el campo de vulnerabilidad\_numero respecto a la incidencia. Lo anterior, es de gran importancia para el objetivo de este trabajo, puesto que, permite analizar si el índice propuesto por el DANE predice adecuadamente la incidencia a la COVID-19 presentada en algunas ciudades de Colombia y, en ese sentido, revisar si el índice propuesto al considerar diferentes tipos de riesgo (índice multidimensional) presenta mejores valores en sus métricas.

#### 4.5.2.1. Evaluación del *dataset* multidimensional

Para este *dataset* se tuvo en cuenta la totalidad de las columnas numéricas para realizar la evaluación.

En la Tabla 4, se muestran los resultados obtenidos con un entrenamiento del 80% y prueba de 20%, con los modelos utilizados.

**Tabla 4.** Resultados obtenidos en la evaluación 80%-20% para el índice multidimensional

<b>Modelo</b>	<b>Parámetros por defecto</b>	<b>RMSE</b>	<b>R-cuadrado</b>
<i>Linear Regression</i>	<i>fit_intercept=True</i> <i>copy_X=True</i> <i>n_jobs=None</i> <i>positive=False</i>	0.013	0.358
<i>Decision Tree Regressor</i>	<i>criterion='squared_error'</i> <i>splitter='best'</i> <i>max_depth=7*</i> <i>min_samples_split=2</i> <i>min_samples_leaf=1</i> <i>min_weight_fraction_leaf=0.0</i> <i>max_features=None</i> <i>random_state=None</i> <i>max_leaf_nodes=None</i> <i>min_impurity_decrease=0.0</i> <i>ccp_alpha=0.0</i>	0.010	0.611
<i>K-Nearest Neighbor</i>	<i>n_neighbors=5</i> <i>weights='uniform'</i> <i>algorithm='auto'</i> <i>leaf_size=30</i> <i>p=2</i> <i>metric='minkowski'</i> <i>metric_params=None</i> <i>n_jobs=None</i>	0.019	-0.207



<i>Support Vector Machine</i>	<i>kernel='rbf' degree=3 gamma='scale' coef0=0.0, tol=0.001 C=1.0, epsilon=0.1 shrinking=True cache_size=200 verbose=False max_iter=-1</i>	0.033	-2.697
<i>Random Forest Regressor</i>	<i>n_estimators=100 criterion='squared_error' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=1.0 max_leaf_nodes=None min_impurity_decrease=0.0 bootstrap=True oob_score=False n_jobs=None random_state=None verbose=0 warm_start=False ccp_alpha=0.0 max_samples=None</i>	0.007	0.790
<i>Gradient Boosting Regressor</i>	<i>loss='squared_error' learning_rate=0.1 n_estimators=100 subsample=1.0 criterion='friedman_mse' min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_depth=3 min_impurity_decrease=0.0 init=None random_state=None max_features=None alpha=0.9 verbose=0 max_leaf_nodes=None warm_start=False validation_fraction=0.1 n_iter_no_change=None tol=0.0001 ccp_alpha=0.0</i>	0.008	0.758

<i>Extra Trees Regressor</i>	<i>n_estimators=100 criterion='squared_error' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=1.0 max_leaf_nodes=None min_impurity_decrease=0.0 bootstrap=False oob_score=False n_jobs=None random_state=None verbose=0 warm_start=False ccp_alpha=0.0 max_samples=None</i>	0.007	0.828
<i>AdaBoost Regressor</i>	<i>estimator=DecisionTreeRegressor(max_depth=3) n_estimators=50 learning_rate=1.0 loss='linear' random_state=None base_estimator='deprecated'</i>	0.008	0.761

En la Tabla 5, se muestran los resultados obtenidos con un entrenamiento del 70% y prueba de 30%, con los modelos utilizados por cada algoritmo.

**Tabla 5.** Resultados obtenidos en la evaluación 70%-30% para el modelo multidimensional

<b>Modelo</b>	<b>Parámetros por defecto</b>	<b>RMSE</b>	<b>R-cuadrado</b>
<i>Linear Regression</i>	<i>fit_intercept=True copy_X=True n_jobs=None positive=False</i>	0.013	0.319
<i>Decision Tree Regressor</i>	<i>criterion='squared_error' splitter='best' max_depth=7* min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=None random_state=None max_leaf_nodes=None min_impurity_decrease=0.0 ccp_alpha=0.0</i>	0.012	0.469

<i>K-Nearest Neighbor</i>	<i>n_neighbors=5 weights='uniform' algorithm='auto' leaf_size=30 p=2 metric='minkowski' metric_params=None n_jobs=None</i>	0.019	-0.335
<i>Support Vector Machine</i>	<i>kernel='rbf' degree=3 gamma='scale' coef0=0.0, tol=0.001 C=1.0, epsilon=0.1 shrinking=True cache_size=200 verbose=False max_iter=-1</i>	0.030	-2.496
<i>Random Forest Regressor</i>	<i>n_estimators=100 criterion='squared_error' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=1.0 max_leaf_nodes=None min_impurity_decrease=0.0 bootstrap=True oob_score=False n_jobs=None random_state=None verbose=0 warm_start=False ccp_alpha=0.0 max_samples=None</i>	0.008	0.720
<i>Gradient Boosting Regressor</i>	<i>loss='squared_error' learning_rate=0.1 n_estimators=100 subsample=1.0 criterion='friedman_mse' min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_depth=3 min_impurity_decrease=0.0 init=None random_state=None max_features=None alpha=0.9 verbose=0</i>	0.009	0.637

	<pre> max_leaf_nodes=None warm_start=False validation_fraction=0.1 n_iter_no_change=None tol=0.0001 ccp_alpha=0.0 </pre>		
<i>Extra Trees Regressor</i>	<pre> n_estimators=100 criterion='squared_error' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=1.0 max_leaf_nodes=None min_impurity_decrease=0.0 bootstrap=False oob_score=False n_jobs=None random_state=None verbose=0 warm_start=False ccp_alpha=0.0 max_samples=None </pre>	0.009	0.700
<i>AdaBoost Regressor</i>	<pre> estimator=DecisionTreeRegressor(max_depth=3) n_estimators=50 learning_rate=1.0 loss='linear' random_state=None base_estimator='deprecated' </pre>	0.009	0.683

Para este modelo en el que se tuvieron en cuenta variables de varios tipos, se observó que el mejor resultado para las métricas R-cuadrado y *RMSE* fue para el modelo *Extra Trees Regressor*. Este resultado se obtuvo en la división 80%-20% (entrenamiento y prueba, respectivamente), el valor obtenido fue 0.828 para R-cuadrado, lo cual indica que la aptitud del modelo para las predicciones realizadas es cercana a 1, es decir que realiza la mayoría de sus predicciones adecuadamente. Respecto a *RMSE* se obtuvo un valor de 0.007, el cual es cercano a 0 (valor óptimo). Sin embargo, para estos modelos la variable de referencia fue R-cuadrado, debido a que, los valores de incidencia eran muy bajos, por lo cual el *RMSE* tendía a ser bajo también.

#### 4.5.2.2. Evaluación del *dataset* del DANE

Para evaluar el modelo del DANE, se tuvo en cuenta la columna que indica el número de vulnerabilidad, y se evaluó respecto a la incidencia de casos de la COVID-19, para

así definir qué modelo puede predecir con mayor eficiencia la incidencia de casos. Se utilizaron los mismos algoritmos que en el modelo propuesto, es decir, regresión lineal, árboles de decisiones (*Decision Trees*), basado en instancias (KNN, *k-Nearest Neighbor*), máquinas de vectores de soporte (SVM, *Support Vector Machine*), bosques aleatorios (*Random Forest*), y potenciación del gradiente (*Gradient Boosting*), también se trabajó con los dos meta-estimadores (*Extra Trees Regressor* y *AdaBoost Regressor*). La evaluación se realizó con los mismos escenarios de entrenamiento y prueba; escenario 1 80% entrenamiento y 20% prueba, escenario 2 70% entrenamiento y 30% prueba.

En la Tabla 6, se muestran los resultados obtenidos con un entrenamiento del 80% y prueba de 20%, con los modelos anteriormente mencionados.

**Tabla 6.** Resultados obtenidos en la evaluación 80%-20% para el índice del DANE.

<b>Modelo</b>	<b>Parámetros por defecto</b>	<b>RMSE</b>	<b>R-cuadrado</b>
<i>Linear Regression</i>	<i>fit_intercept=True</i> <i>copy_X=True</i> <i>n_jobs=None</i> <i>positive=False</i>	0.016	0.090
<i>Decision Tree Regressor</i>	<i>criterion='squared_error'</i> <i>splitter='best'</i> <i>max_depth=7*</i> <i>min_samples_split=2</i> <i>min_samples_leaf=1</i> <i>min_weight_fraction_leaf=0.0</i> <i>max_features=None</i> <i>random_state=None</i> <i>max_leaf_nodes=None</i> <i>min_impurity_decrease=0.0</i> <i>ccp_alpha=0.0</i>	0.012	0.517
<i>K-Nearest Neighbor</i>	<i>n_neighbors=5</i> <i>weights='uniform'</i> <i>algorithm='auto'</i> <i>leaf_size=30</i> <i>p=2</i> <i>metric='minkowski'</i> <i>metric_params=None</i> <i>n_jobs=None</i>	0.019	-0.287

<i>Support Vector Machine</i>	<i>kernel='rbf' degree=3 gamma='scale' coef0=0.0, tol=0.001 C=1.0, epsilon=0.1 shrinking=True cache_size=200 verbose=False max_iter=-1</i>	0.033	-2.697
<i>Random Forest Regressor</i>	<i>n_estimators=100 criterion='squared_error' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=1.0 max_leaf_nodes=None min_impurity_decrease=0.0 bootstrap=True oob_score=False n_jobs=None random_state=None verbose=0 warm_start=False ccp_alpha=0.0 max_samples=None</i>	0.010	0.608
<i>Gradient Boosting Regressor</i>	<i>loss='squared_error' learning_rate=0.1 n_estimators=100 subsample=1.0 criterion='friedman_mse' min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_depth=3 min_impurity_decrease=0.0 init=None random_state=None max_features=None alpha=0.9 verbose=0 max_leaf_nodes=None warm_start=False validation_fraction=0.1 n_iter_no_change=None tol=0.0001 ccp_alpha=0.0</i>	0.011	0.546

<i>Extra Trees Regressor</i>	<i>n_estimators=100 criterion='squared_error' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=1.0 max_leaf_nodes=None min_impurity_decrease=0.0 bootstrap=False oob_score=False n_jobs=None random_state=None verbose=0 warm_start=False ccp_alpha=0.0 max_samples=None</i>	0.011	0.561
<i>AdaBoost Regressor</i>	<i>estimator=DecisionTreeRegressor(max_depth=3) n_estimators=50 learning_rate=1.0 loss='linear' random_state=None base_estimator='deprecated'</i>	0.013	0.395

En la Tabla 7, se muestran los resultados obtenidos con un entrenamiento del 70% y prueba de 30%, con los diferentes modelos.

**Tabla 7.** Resultados obtenidos en la evaluación 70%-30% para el índice del DANE.

<b>Modelo</b>	<b>Parámetros por defecto</b>	<b>RMSE</b>	<b>R-cuadrado</b>
<i>Linear Regression</i>	<i>fit_intercept=True copy_X=True n_jobs=None positive=False</i>	0.016	0.033
<i>Decision Tree Regressor</i>	<i>criterion='squared_error' splitter='best' max_depth=7* min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=None random_state=None max_leaf_nodes=None min_impurity_decrease=0.0 ccp_alpha=0.0</i>	0.011	0.474

<i>K-Nearest Neighbor</i>	<i>n_neighbors=5 weights='uniform' algorithm='auto' leaf_size=30 p=2 metric='minkowski' metric_params=None n_jobs=None</i>	0.018	-0.220
<i>Support Vector Machine</i>	<i>kernel='rbf' degree=3 gamma='scale' coef0=0.0, tol=0.001 C=1.0, epsilon=0.1 shrinking=True cache_size=200 verbose=False max_iter=-1</i>	0.030	-2.496
<i>Random Forest Regressor</i>	<i>n_estimators=100 criterion='squared_error' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=1.0 max_leaf_nodes=None min_impurity_decrease=0.0 bootstrap=True oob_score=False n_jobs=None random_state=None verbose=0 warm_start=False ccp_alpha=0.0 max_samples=None</i>	0.010	0.596
<i>Gradient Boosting Regressor</i>	<i>loss='squared_error' learning_rate=0.1 n_estimators=100 subsample=1.0 criterion='friedman_mse' min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_depth=3 min_impurity_decrease=0.0 init=None random_state=None max_features=None alpha=0.9 verbose=0</i>	0.011	0.480



	<pre> max_leaf_nodes=None warm_start=False validation_fraction=0.1 n_iter_no_change=None tol=0.0001 ccp_alpha=0.0 </pre>		
Extra Trees Regressor	<pre> n_estimators=100 criterion='squared_error' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=1.0 max_leaf_nodes=None min_impurity_decrease=0.0 bootstrap=False oob_score=False n_jobs=None random_state=None verbose=0 warm_start=False ccp_alpha=0.0 max_samples=None </pre>	0.011	0.535
AdaBoost Regressor	<pre> estimator=DecisionTreeRegressor(max_depth=3) n_estimators=50 learning_rate=1.0 loss='linear' random_state=None base_estimator='deprecated' </pre>	0.010	0.610

Para este modelo en el que se tuvo en cuenta la variable de vulnerabilidad correspondiente al modelo del DANE, se observó que el mejor resultado para las métricas R-cuadrado y *RMSE* fue para el modelo *AdaBoost Regressor* cuando se hizo la división de 70%-30% (entrenamiento y prueba, respectivamente) obteniendo un valor de 0.610 para R-cuadrado y 0.010 para *RMSE*.

Respecto a las métricas obtenidas por los dos mejores modelos para cada *dataset*, se evidenció que el modelo multidimensional propuesto presenta un mejor desempeño, al analizar los valores obtenidos en las métricas. Las métricas R-cuadrado y *RMSE* con el modelo propuesto obtuvieron un valor máximo de 0.828 y 0.007, respectivamente, en comparación con el modelo del DANE que presentó un valor de 0.610 y 0.010, respectivamente. Lo anterior permite establecer, que el

modelo multidimensional realiza una mejor predicción de los valores de incidencia de la COVID-19.

## 4.6. Optimización del modelo

Para la optimización de los modelos se tomaron los algoritmos con mejores resultados en sus métricas (*Decision Tree Regressor*, *Random Forest*, *Gradient Boosting*, *Extra Trees Regressor* y *AdaBoost Regressor*) y se evaluaron algunas variaciones en los hiper-parámetros con el objetivo de mejorar los modelos ya implementados. Cabe resaltar que, dependiendo del modelo que se va a optimizar, cambian los hiper-parámetros, por lo que, en las siguientes subsecciones se indica qué hiper-parámetros fueron considerados para la optimización de cada modelo.

### 4.6.1. *Decision Tree Regressor*

Para este modelo se optó por utilizar una división de los datos de 80% para entrenamiento y 20% para prueba, debido a que en la Tabla 4 se observó que se obtenían mejores resultados. Cabe mencionar que los gráficos obtenidos fueron resultado de la evaluación del grupo de datos del 80% correspondiente al entrenamiento.

Se decidió utilizar la técnica de *pruning*, enfatizando en el parámetro de complejidad de costos (*ccp\_alpha*). Esta técnica se encarga de reducir (podar) la complejidad de un modelo de árbol de decisión y así mejorar el rendimiento, la poda puede ayudar a evitar el sobreajuste (que ocurre cuando un modelo aprende demasiado, debido a los datos de entrenamiento, ocasionando el no poder generalizar nuevos datos) [70]. Un valor alto de *ccp\_alpha* aumenta la cantidad de nodos podados, tratando de no afectar las métricas [71]. Por lo tanto, se desarrolló el código para obtener el mejor valor de este parámetro, con el fin de aumentar la métrica de desempeño R-cuadrado [72].

El *ccp\_alpha* halla los enlaces más débiles, los cuales cuentan con un valor de *alpha* efectivo, estos son eliminados primero. Se imprimieron todos los valores de *alpha* efectivos para este modelo, para obtener el valor de impureza, el cual está relacionado a la poda de la hoja. A medida que se incrementa el valor de *alpha*, hay mayor eliminación de las hojas, por lo tanto, aumentará la impureza del total de las hojas. En la Figura 36 se observa el resultado obtenido.

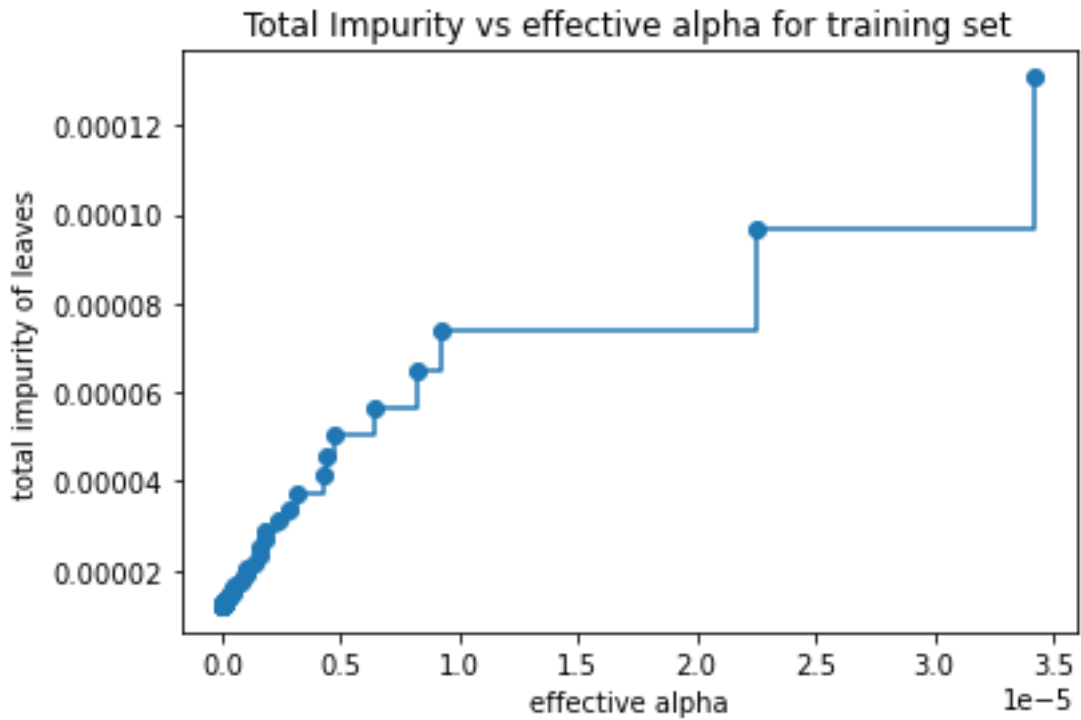


Figura 36. Total impureza vs valor efectivo de  $\alpha$ .

Se procedió a eliminar el último valor de  $\alpha$  efectivo, debido a que éste poda todo el árbol, ocasionando que quede con un solo nodo.

En la Figura 37 se muestra el número de nodos respecto a los valores de  $\alpha$ , y también la profundidad del árbol contra los valores de  $\alpha$ .

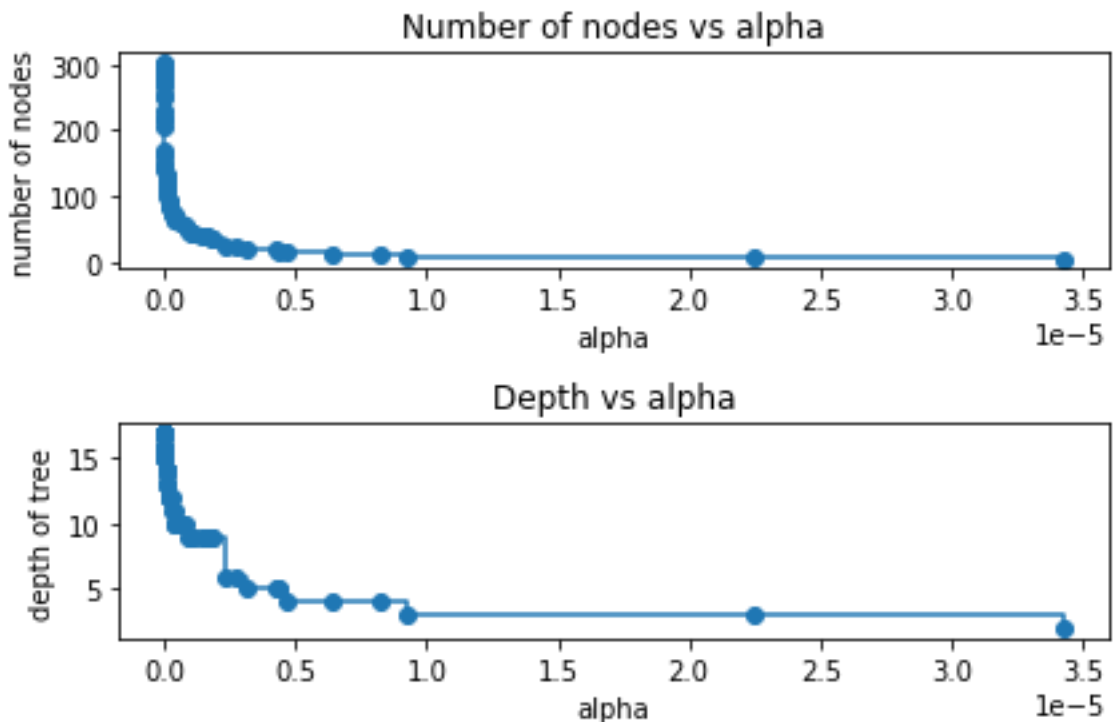


Figura 37. Número de nodos vs  $\alpha$ , y profundidad del árbol vs  $\alpha$ .

Con el objetivo de decidir el mejor valor para el parámetro *ccp\_alpha*, se graficaron los valores de *alpha* respecto al puntaje que se genera para los grupos de datos de entrenamiento y prueba. De esta forma se selecciona el mejor valor para los datos de prueba, debido a que estos no los conoce el modelo, pero disminuyendo lo menor posible el puntaje en los datos de entrenamiento que son los que conocen el resultado. Esto se presenta en la Figura 38.

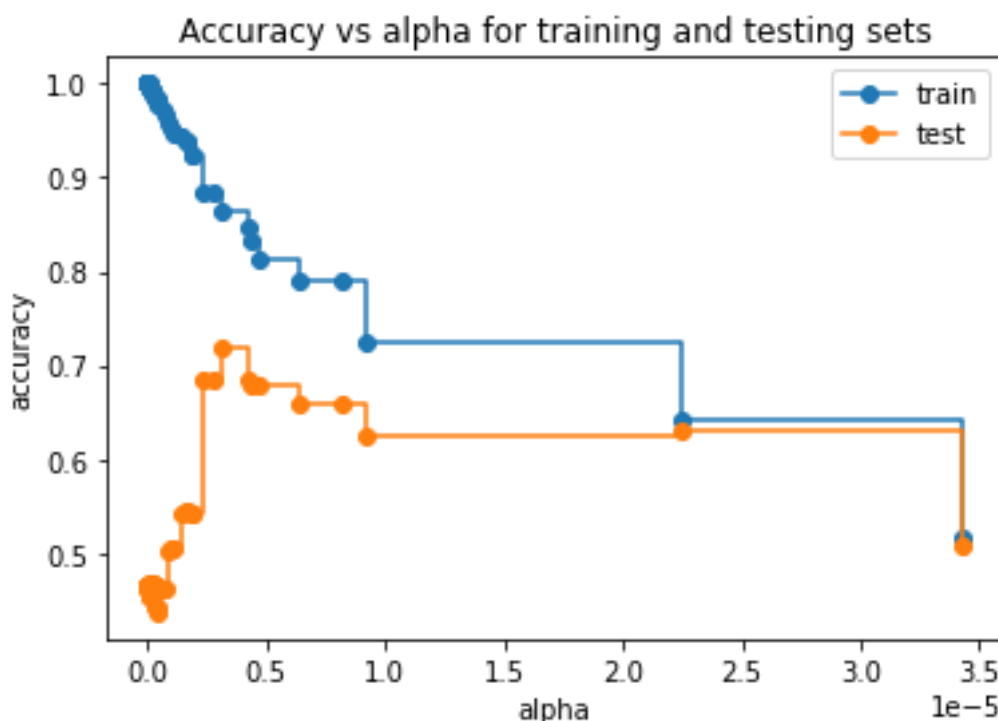
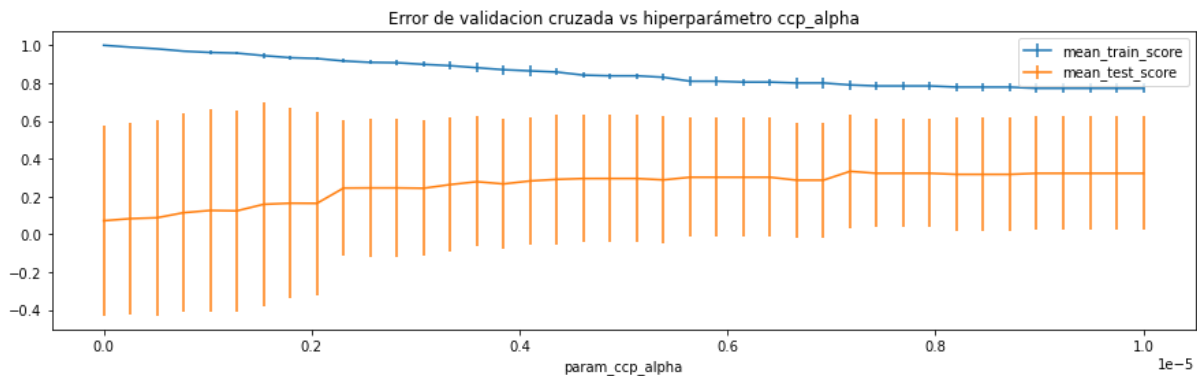


Figura 38. Puntaje vs alpha.

Adicionalmente, se hizo uso de *GridSearchCV*, esta es una técnica de validación cruzada (en inglés, *cross validation*, *CV*) que permite ejecutar un modelo introduciendo ciertos hiper-parámetros para así encontrar los mejores valores, es decir, la mejor combinación de los hiper-parámetros.

En la Figura 39 se muestra el resultado obtenido por *GridSearchCV*, donde los hiper-parámetros del modelo establecidos fueron: *max\_depth = None* (no determina una máxima profundidad), *min\_samples\_split = 2* (número de muestras mínimas para dividir un nodo), *min\_samples\_leaf = 1* (número de muestras requeridas para un nodo hoja), *random\_state = 329* (para seleccionar siempre la misma división). Los hiper-parámetros de *GridSearchCV* fueron: *param\_grid* (valores de *ccp\_alpha* que van a variar), *cv = 5* (cantidad de veces que se va a iterar la validación), *refit = True* (se usa

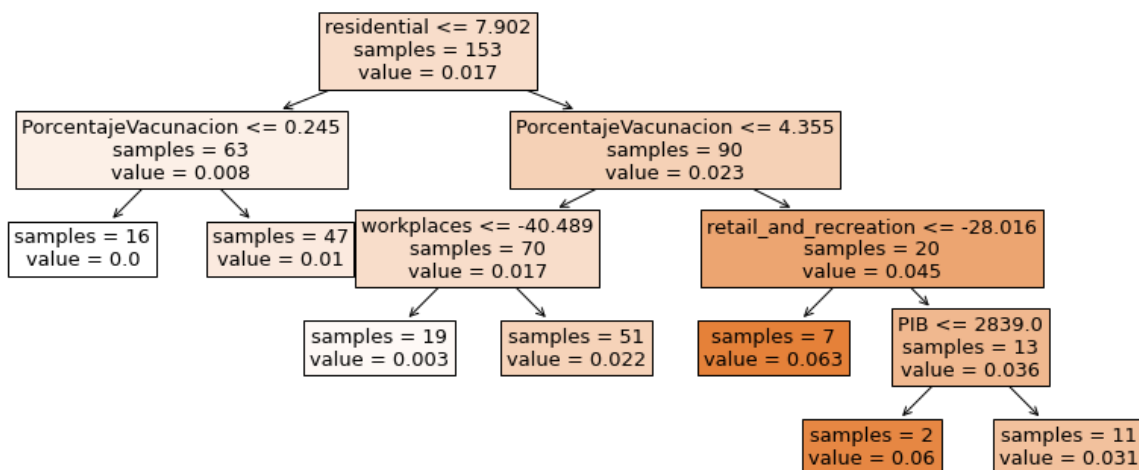
cuando se quiere hallar el mejor parámetro), *return\_train\_score = True* (para obtener información acerca de diferentes configuraciones de los hiper-parámetros).



**Figura 39.** Error de validación cruzada vs *ccp\_alpha*.

Se obtiene como resultado que el mejor valor para el hiper-parámetro *ccp\_alpha* es  $7.179487e-06$ .

Al graficar el nuevo árbol, este contiene una profundidad de 4, y 7 nodos terminales como se muestra en la Figura 40.



**Figura 40.** Árbol podado con el mejor valor de *ccp\_alpha*.

Continuando con la búsqueda de los mejores hiper-parámetros, se procedió a hallar los mejores valores para los siguientes hiper-parámetros: *criterion* (para medir la calidad de una división), *max\_features* (número de características a considerar al momento de buscar la mejor división), *max\_depth* (profundidad máxima del árbol). El resultado se observa en la Figura 41.

```

DecisionTreeRegressor
DecisionTreeRegressor(criterion='absolute_error', max_depth=4,
                      max_features='auto', random_state=329)

```

Figura 41. Mejores hiper-parámetros obtenidos para *Decision Tree Regressor*.

Evaluando los resultados que se presentan en la Figura 41, se obtuvo un valor de R-cuadrado de 0.708769 y *RMSE* de 0.009, sin embargo, al agregar el valor obtenido para el parámetro *ccp\_alpha*, el resultado fue 0.708779 para R-cuadrado y 0.009 para *RMSE*; también se pudo corroborar el valor obtenido para *max\_depth*. Este segundo resultado es un poco mayor debido a que se establece un valor para *ccp\_alpha*, y este es muy cercano al valor por *default* que es cero. En la Figura 42 se muestra el modelo final.

```

DecisionTreeRegressor
DecisionTreeRegressor(ccp_alpha=7.17948717948718e-06,
                      criterion='absolute_error', max_depth=4,
                      max_features='auto', random_state=329)

```

Figura 42. Modelo optimizado para *Decision Tree Regressor*.

Cabe resaltar que esta evaluación se hizo únicamente con los mejores hiper-parámetros hallados que se mencionaron anteriormente y no directamente con *GridSearchCV*, debido a que este cuenta con el procedimiento de *CV* internamente, donde el valor del puntaje final es el promedio del puntaje de cada división y para este caso disminuye el valor evaluado de la métrica R-cuadrado a 0.5885 y aumenta *RMSE* a 0.0104. En la Figura 43 se muestra el resultado obtenido al realizar la evaluación con *GridSearchCV*.

param_random_state	params	split0_test_score	split1_test_score	split2_test_score	split3_test_score	split4_test_score	mean_test_score	std_test_score
329	{'criterion': 'absolute_error', 'max_depth': 4, ...}	0.651022	0.621360	0.569940	0.485098	0.615272	0.588538	0.057866
329	{'criterion': 'poisson', 'max_depth': 4, 'max_...', ...}	0.680089	0.660372	0.550904	0.495334	0.392092	0.555758	0.106643
329	{'criterion': 'squared_error', 'max_depth': 13, ...}	0.523392	0.606194	0.567049	0.517308	0.479591	0.538707	0.043692

Figura 43. Resultados de *Decision Tree Regressor* con *GridSearchCV*.

### 4.6.2. Random Forest Regressor

En esta optimización la división de los datos se manejó de igual manera que para el anterior modelo, es decir, 80% entrenamiento y 20% prueba, el objetivo fue mejorar el resultado inicial obtenido para la métrica R-cuadrado igual a 0.790.

Para la optimización de este modelo se debe tener en cuenta aquellos hiperparámetros que detienen el crecimiento del árbol:  $n\_estimators$  (número de árboles incluidos en el modelo),  $max\_depth$  (profundidad máxima que puede alcanzar cada árbol),  $max\_features$  (número de predicciones a considerar para cada división),  $oob\_score$  (*out-of-bag* o R-cuadrado),  $random\_state$  (semilla para que los resultados sean reproducibles). Cabe mencionar que los gráficos obtenidos fueron resultado de la evaluación del grupo de datos del 80% correspondiente al entrenamiento.

El uso del modelo *Random Forest* tiene la ventaja de que la cantidad de árboles ( $n\_estimators$ ) no es un parámetro crítico, por el contrario, el añadir árboles puede cada vez mejorar el resultado, por lo tanto, no genera *overfitting* como sería en el caso de *Gradient Boosting*. Sin embargo, va a llegar un momento en que al añadir más árboles el modelo no mejore y se estabilice, lo que ocasiona que un exceso de árboles pueda consumir más recursos computacionales [73].

Teniendo en cuenta la métrica  $oob\_score$  (R-cuadrado), se halló la cantidad de árboles donde el error de validación se estabiliza. En las figuras 44 y 45 se observa el valor mínimo y el valor máximo, respectivamente, donde se estabiliza dicho error.

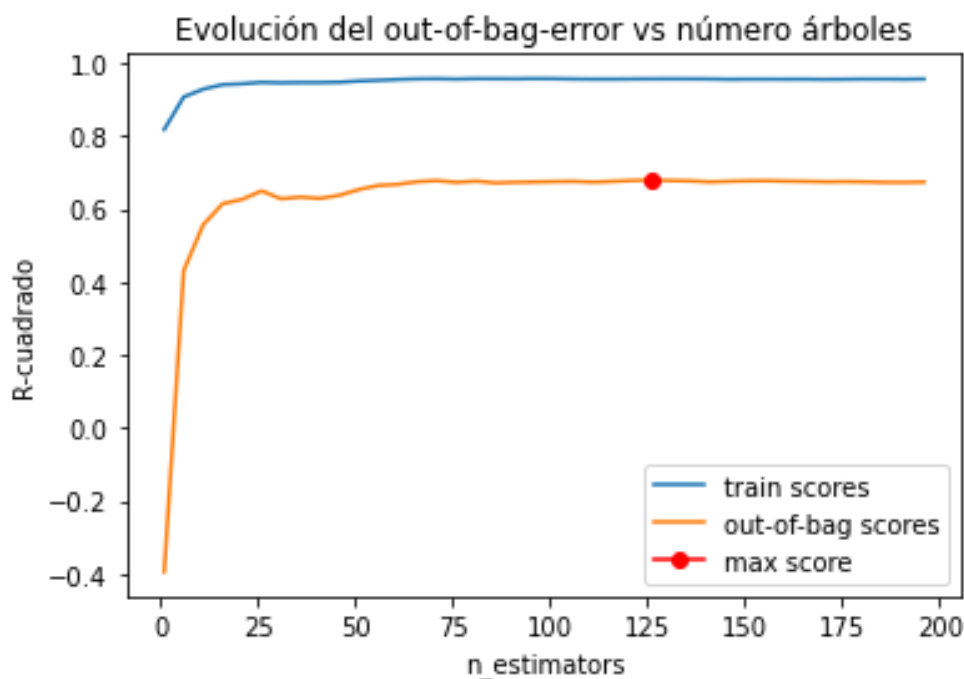


Figura 44. Valor óptimo mínimo para  $n\_estimators$  en *Random Forest*.

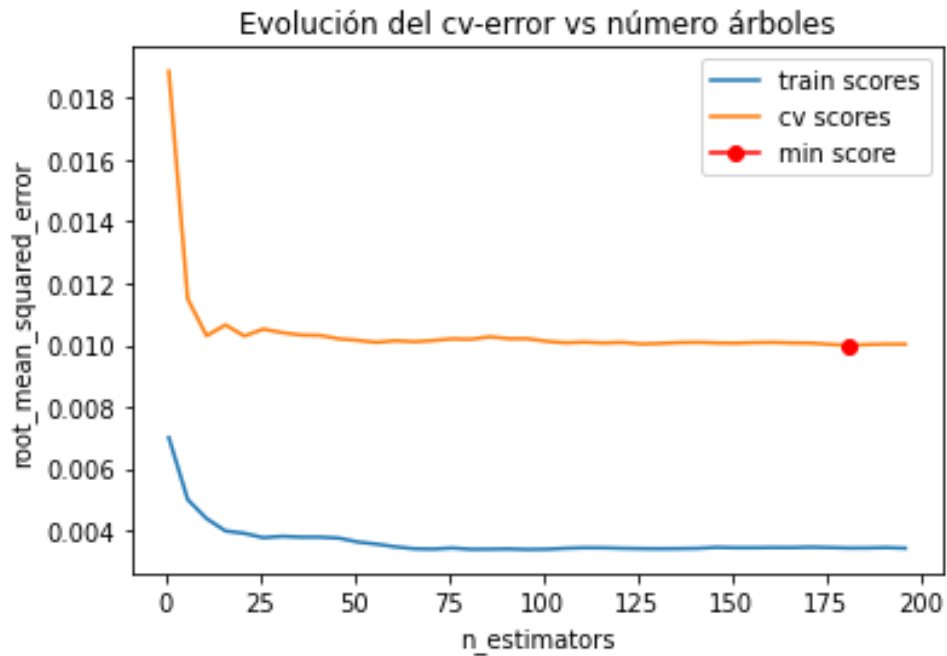


Figura 45. Valor óptimo máximo para  $n\_estimators$  en *Random Forest*.

En las Figuras 44 y 45 se observa que el valor óptimo para el hiper-parámetro  $n\_estimators$  está entre 126 y 181.

Para la optimización del hiper-parámetro  $max\_features$ , también se calcula el rango de valores donde trabaja mejor. En las Figuras 46 y 47 se observa el valor mínimo y el valor máximo, respectivamente, donde se estabiliza dicho error.

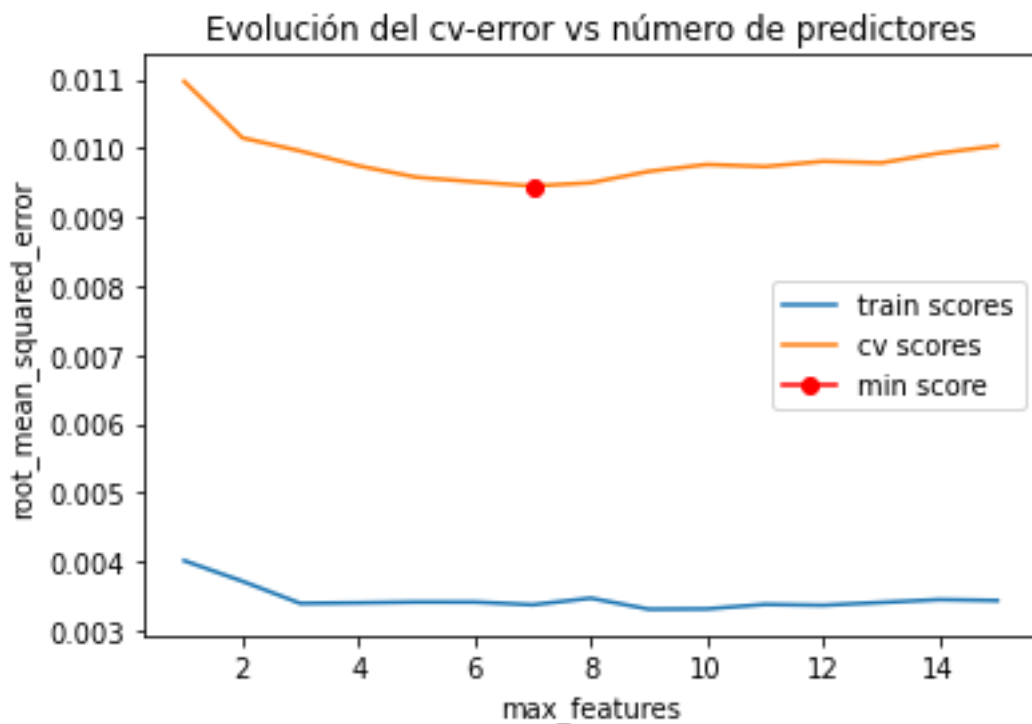


Figura 46. Valor óptimo mínimo para  $max\_features$  en *Random Forest*.



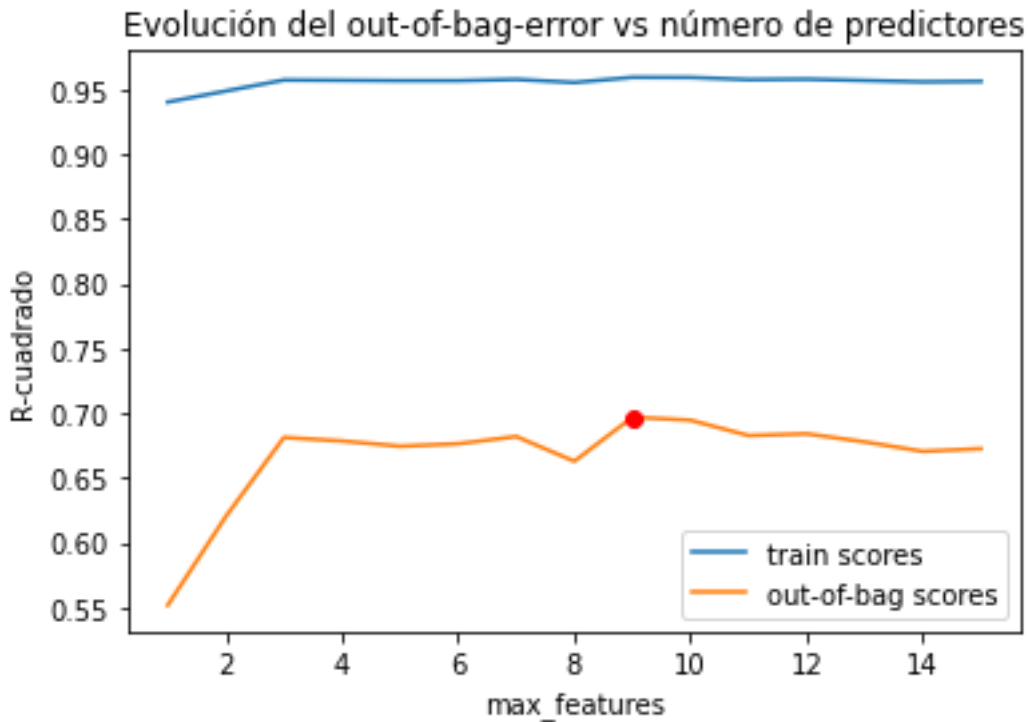


Figura 47. Valor óptimo máximo para *max\_features* en *Random Forest*.

En las Figuras 46 y 47 se visualiza el valor óptimo para el hiper-parámetro *max\_features* el cual se encuentra entre 7 y 9.

*Random Forest* cuenta con la métrica *out-of-bag* (R-cuadrado) la cual busca ser lo mejor posible, por lo tanto, se estableció un rango y se variaron los hiper-parámetros *n\_estimators*, *max\_features* y *max\_depth* para obtener el valor más alto de *out-of-bag*. En la Figura 48 se muestran los cuatro mejores resultados.

	oob_r2	max_depth	max_features	n_estimators
784	0.706507	5	9	70
2153	0.703199	15	9	165
2152	0.702576	15	9	160
2154	0.701839	15	9	170

Figura 48. Mejores hiper-parámetros para *Random Forest* según *out-of-bag*.

El análisis de los hiper-parámetros individualmente puede ayudar a comprender los rangos donde el modelo puede generar una mejor evaluación. Sin embargo, estos hiper-parámetros no actúan por separado, al contrario, conjuntamente se deben encontrar los mejores resultados, por ello se recurrió al uso de *GridSearchCV*. Los hiper-parámetros a evaluar fueron los que se usaron anteriormente (*n\_estimators*, *max\_features* y *max\_depth*). Los resultados se muestran en la Figura 49.

	param_max_depth	param_max_features	param_n_estimators	mean_test_score	std_test_score	mean_train_score	std_train_score
1395	13	7	125	0.663328	0.048617	0.953817	0.004058
1394	13	7	120	0.662530	0.047736	0.954294	0.003814
1393	13	7	115	0.662085	0.047236	0.953886	0.003786
537	5	6	185	0.662003	0.056971	0.898227	0.006803

**Figura 49.** Mejores hiper-parámetros para *Random Forest* implementando *GridSearchCV*.

El resultado obtenido para la métrica R-cuadrado fue 0.663, sin embargo, se optó por evaluar los mejores hiper-parámetros obtenidos directamente para evitar la evaluación por CV. El modelo implementado se muestra en la Figura 50.

```

RandomForestRegressor
RandomForestRegressor(max_depth=13, max_features=7, n_estimators=125,
random_state=329)

```

**Figura 50.** Modelo optimizado para *Random Forest Regressor*.

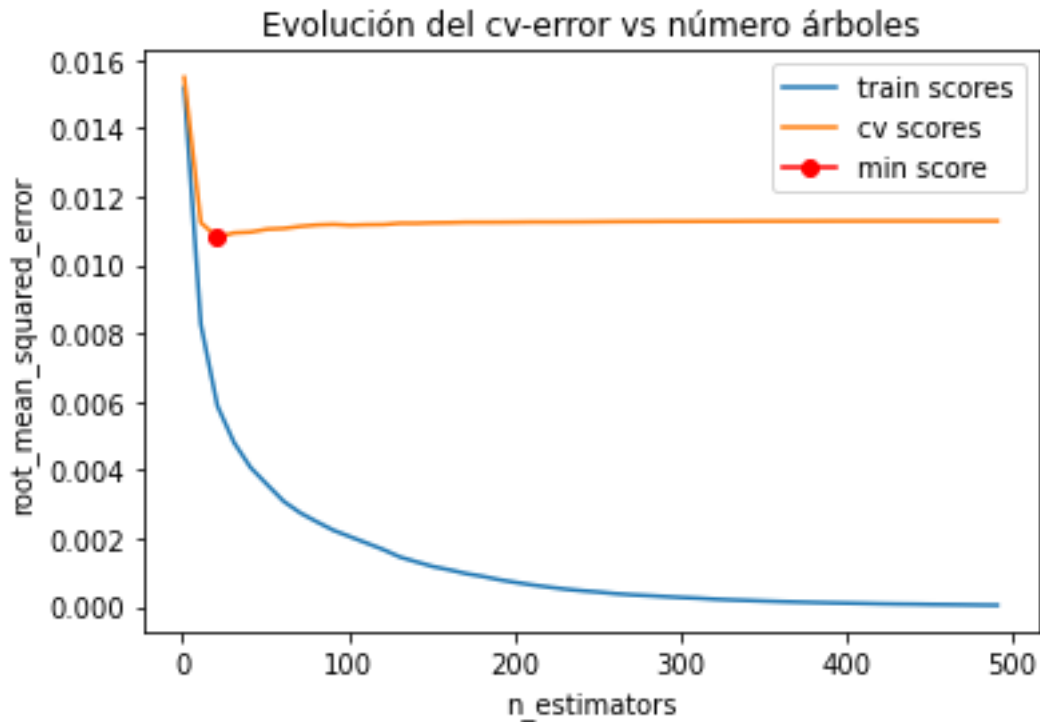
La evaluación del modelo optimizado arrojó como resultado para el parámetro R-cuadrado el valor 0.802 y para *RMSE* 0.007, mejorando el valor obtenido anteriormente sin optimizar el cual fue 0.790 para R-cuadrado y 0.007 para *RMSE*.

### 4.6.3. Gradient Boosting Regressor

Al igual que en el modelo anterior, se optó por manejar una división de los datos de 80% para entrenamiento y 20% para prueba. El objetivo de la optimización de este modelo es reducir el error. Cabe mencionar que los gráficos obtenidos fueron el resultado de la evaluación del grupo de datos del 80% correspondiente al entrenamiento.

Como no es posible conocer directamente los mejores valores para los hiper-parámetros, mediante estrategias de validación como CV, se trata de identificar los más adecuados.

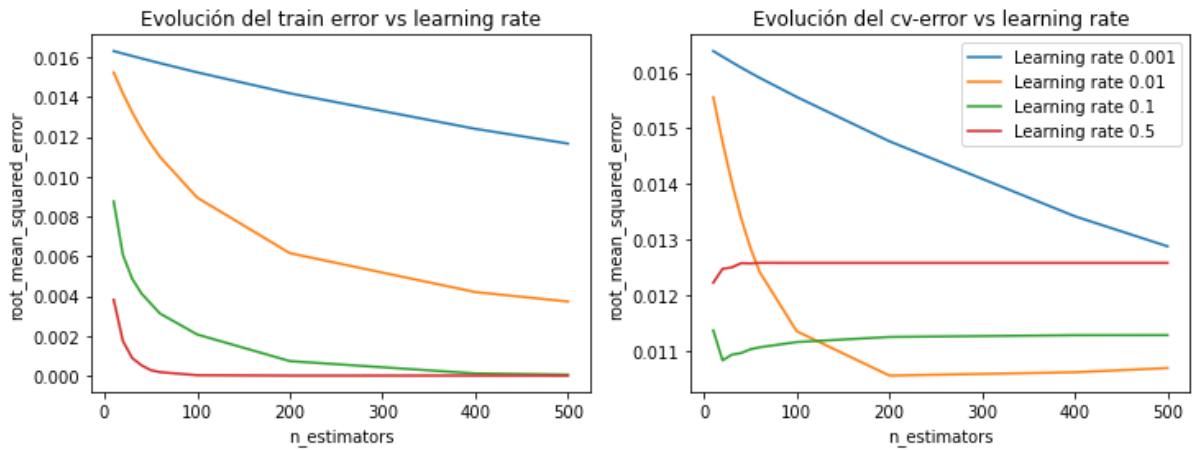
Se inició hallando el número de árboles, es decir el parámetro denominado “*n\_estimators*” para conocer el valor que genera un menor error entre el grupo de los datos de entrenamiento. Este es un hiper-parámetro crítico debido a que, a medida que se añaden árboles, se incrementa el riesgo de *overfitting* (es decir, sobre-entrenamiento del modelo) [65]. La gráfica se presenta en la Figura 51.



**Figura 51.** Evolución del *cv\_error* respecto al número de árboles para *Gradient Boosting Regressor*.

En la Figura 51 se observa que, el valor de *n\_estimators* en el que se obtiene un menor error de la validación entre los datos de entrenamiento es 21, sin embargo, un valor mayor a este logra mantener el *RMSE* entre 0.010 y 0.012.

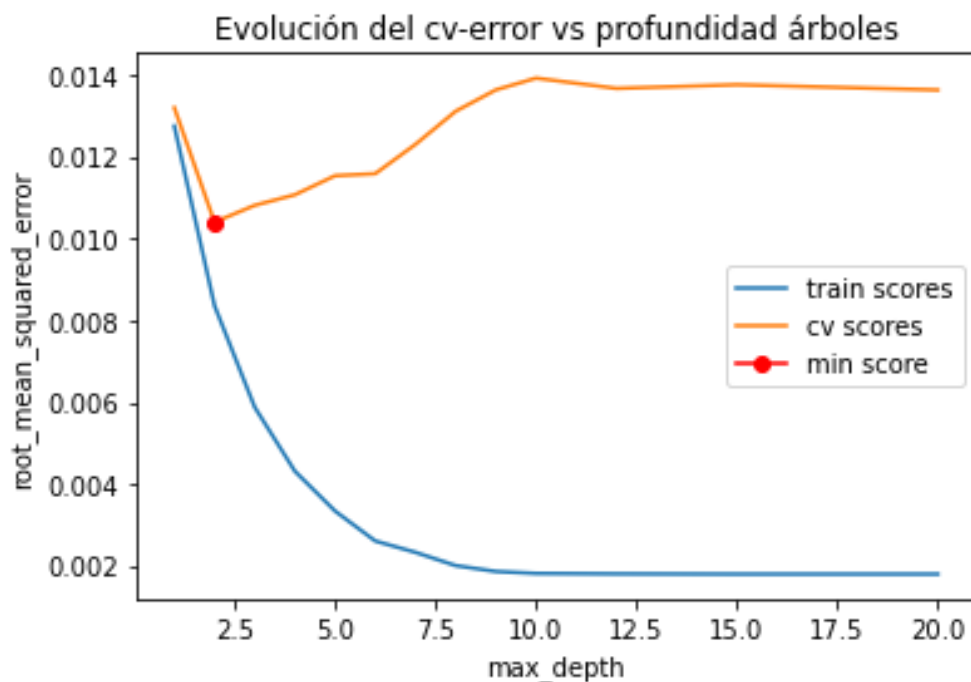
El siguiente hiper-parámetro hallado fue *learning\_rate*, este hiper-parámetro permite controlar qué tan rápido aprende el modelo, de esto también depende el riesgo de provocar *overfitting*. Estos hiper-parámetros son interdependientes, lo que quiere decir que a menor *learning\_rate*, mayor es el valor de *n\_estimators* si se desean buenos resultados, y por lo tanto habrá menor riesgo de *overfitting*. En la Figura 52 se muestra la evolución del *RMSE* a medida que incrementa el número de árboles para tres valores de *learning\_rate* (0.5, 0.1, 0.01, 0.001).



**Figura 52.** Evolución del *train error* y el *cv-error* respecto al número de árboles para 3 medidas de *learning rate* para *Gradient Boosting Regressor*.

En la Figura 52 se visualiza que, efectivamente a mayor número de árboles se obtiene un menor valor de *RMSE* y funciona mejor el valor de 0.5 para el *learning rate*, sin embargo, el *learning rate* de 0.1 puede funcionar mejor que los otros valores para un número de *n\_estimators* pequeño cuando se tiene en cuenta el *RMSE* debido a la *CV*.

Para calcular el mejor valor para *max\_depth*, se debe tener en cuenta que para modelos *Gradient Boosting* este valor suele ser muy bajo, debido a que el modelo hace que cada árbol aprenda una pequeña parte de la relación entre predictores y variable respuesta. La gráfica obtenida se muestra en la Figura 53.



**Figura 53.** Evolución del *cv-error* respecto a la profundidad de los árboles.

A partir de la Figura 53 se puede observar que el valor de *max\_depth* que presenta el menor *RMSE* es 2.

Al igual que en el modelo anterior se probó *GridSearchCV*, en este caso se empleó una estrategia llamada parada temprana, donde no se incluye el número de árboles como hiper-parámetro, por el contrario, se escoge un valor por defecto muy alto y se activa la parada temprana. Los hiper-parámetros de modelo se fijaron en *n\_estimators* = 1000, *random\_state* = 329, para la activación de la parada temprana se usaron los hiper-parámetros *validation\_fraction* = 0.1 (proporción de datos separados del conjunto de entrenamiento y empleados como conjunto de validación para determinar la parada temprana), *n\_iter\_no\_change* = 5 (número de iteraciones consecutivas en las que no se debe superar el *tol* para que el algoritmo se detenga) y *tol* = 0.001 (porcentaje mínimo de mejora entre dos iteraciones consecutivas por debajo del cual se considera que el modelo no ha mejorado). En la Figura 54 se muestran los 4 mejores resultados, y en la Figura 55 se muestra el modelo construido a partir de los hiper-parámetros obtenidos.

	param_learning_rate	param_max_depth	param_max_features	param_subsample	mean_test_score	std_test_score
199	0.5	2	auto	1	0.618199	0.062265
214	0.5	3	auto	1	0.602876	0.056362
229	0.5	4	auto	1	0.587081	0.197327
219	0.5	3	sqrt	1	0.581449	0.108576

**Figura 54.** Mejores hiper-parámetros obtenidos con *GridSearchCV* para *Gradient Boosting Regressor*.

```

GradientBoostingRegressor
GradientBoostingRegressor(learning_rate=0.5, max_depth=2, max_features='auto',
                           n_estimators=1000, n_iter_no_change=5,
                           random_state=329, subsample=1)

```

**Figura 55.** Modelo optimizado para *Gradient Boosting Regressor*.

La parada temprana se activó dando como resultado un total de 6 árboles para el modelo y los hiper-parámetros que se observan en la primera fila de la Figura 54. Además, se visualiza que para los mejores hiper-parámetros, el valor promedio de test (R-cuadrado) es 0.618. Adicionalmente, se corroboró el valor obtenido previamente para *max\_depth*, sin embargo, se optó por tomar los mejores hiper-parámetros hallados por medio de *GridSearchCV*, y evaluarlos evitando aplicar CV. Este procedimiento arrojó como resultado un R-cuadrado de 0.765 y *RMSE* de 0.008, superando al valor 0.758 de R-cuadrado y 0.008 para *RMSE* que se tenía antes de la optimización.

Para entrenar modelos *Gradient Boosting* en problemas de regresión, se puede hacer uso de la clase *HistGradientBoostingRegressor* el cual se enfoca en regresión, los hiper-parámetros utilizados aquí fueron  $max\_iter = 1000$  (máximo número de árboles),  $early\_stopping = True$  (detención temprana habilitada),  $validation\_fraction = 0.01$  (proporción de datos separados del conjunto de entrenamiento y empleados como conjunto de validación para determinar la parada temprana),  $n\_iter\_no\_change = 10$  (número de iteraciones consecutivas en las que no se debe superar el  $tol$  para que el algoritmo se detenga),  $tol = 1e-5$  (porcentaje mínimo de mejora entre dos iteraciones consecutivas por debajo del cual se considera que el modelo no ha mejorado), y  $random\_state = 329$  (generador de números pseudoaleatorios para controlar el submuestreo en el proceso de agrupamiento y la división de datos). En la Figura 56 se muestran los 4 mejores resultados, y en la Figura 57 el modelo construido a partir de los mejores hiper-parámetros encontrados.

	param_learning_rate	param_max_depth	mean_test_score	std_test_score	mean_train_score	std_train_score
36	0.5	3	0.591587	0.040197	0.887184	0.025204
39	0.5	6	0.591397	0.048268	0.906027	0.034392
33	0.5	None	0.591397	0.048268	0.906027	0.034392
37	0.5	4	0.591397	0.048268	0.906027	0.034392

**Figura 56.** Mejores hiper-parámetros obtenidos con *Hist Gradient Boosting*.

```

HistGradientBoostingRegressor
HistGradientBoostingRegressor(early_stopping=True, learning_rate=0.5,
                               max_depth=3, max_iter=1000, random_state=329,
                               tol=1e-05, validation_fraction=0.01)

```

**Figura 57.** Modelo optimizado para *Hist Gradient Boosting*.

En la Figura 56 se puede observar el valor obtenido para R-cuadrado ( $mean\_test\_score$ ) que arrojó como resultado 0.591. Además, el número adecuado de árboles para el modelo es 10, no obstante, evaluando los mejores hiper-parámetros sin aplicar CV, se obtuvo un R-cuadrado de 0.810 y  $RMSE$  de 0.007, siendo mejor resultado que el modelo inicial, para este modelo que fue 0.758 de R-cuadrado y 0.008 de  $RMSE$ .

#### 4.6.4. Extra Trees Regressor

Inicialmente, se procedió a hallar el número de árboles óptimo ( $n\_estimators$ ) donde se obtenía el menor error  $RMSE$  entre los datos del grupo de entrenamiento. La gráfica se muestra en la Figura 58.

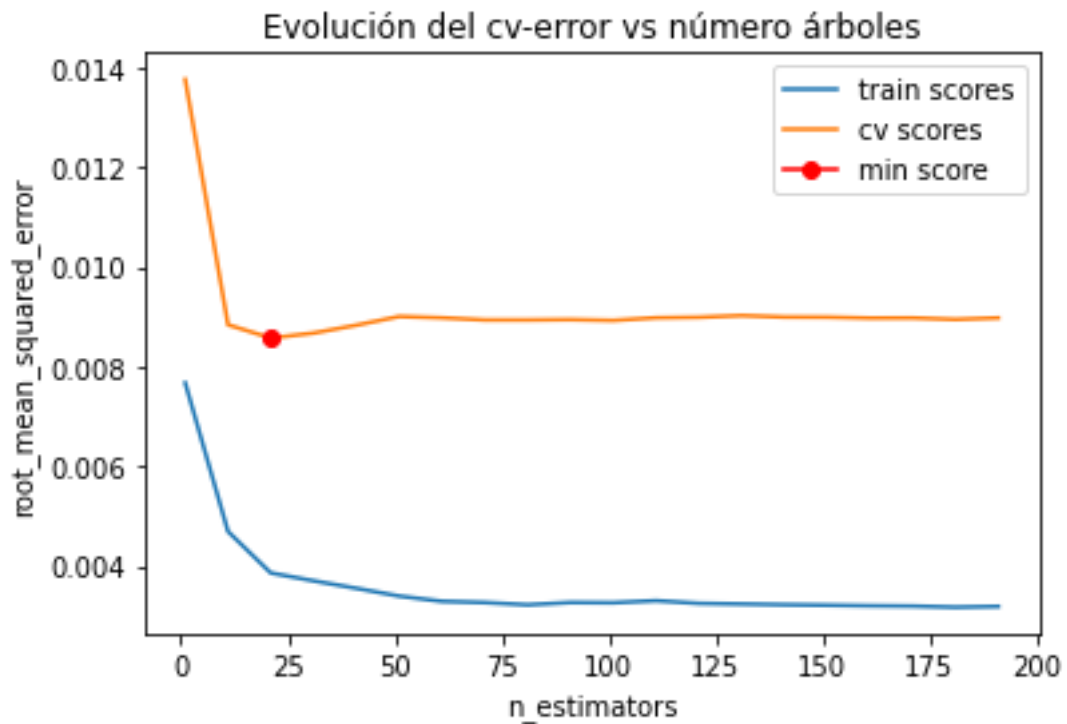


Figura 58. Evolución del  $cv\_error$  respecto al número de árboles para *Extra Trees Regressor*.

En la Figura 58 se observa que, el valor de  $n\_estimators$  en el que se obtiene un menor error de la validación entre los datos de entrenamiento es 21, aunque, un valor mayor de este logra mantener el error entre 0.008 y 0.010.

Para calcular el mejor valor para  $max\_depth$ , se graficó el comportamiento de este hiper-parámetro respecto al  $RMSE$ , tal como se muestra en la Figura 59.

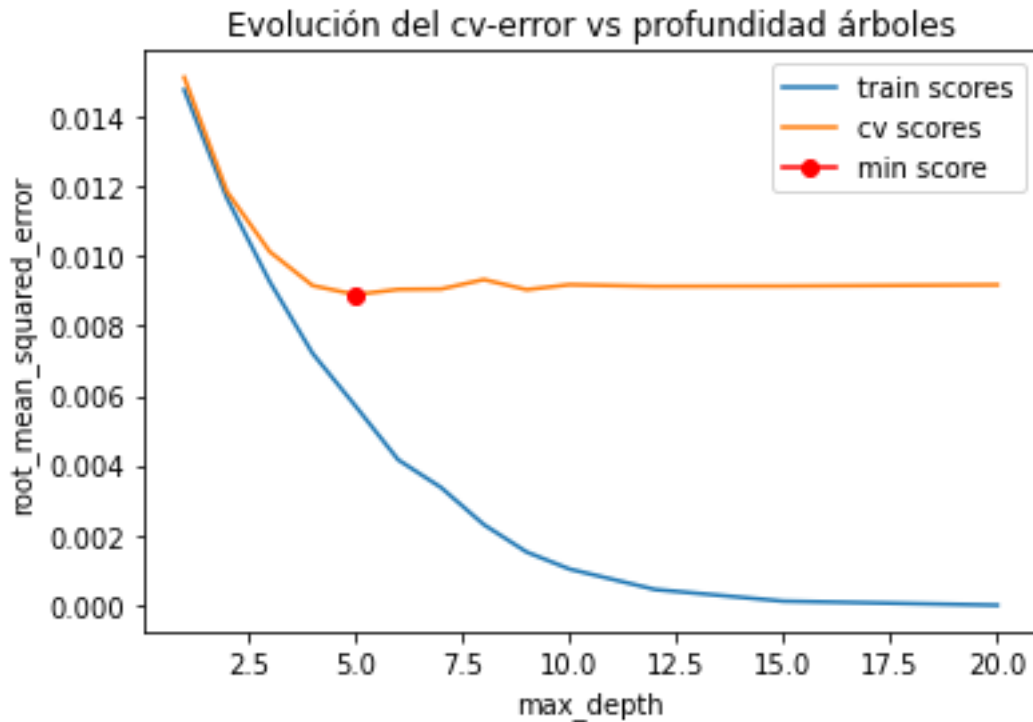


Figura 59. Evolución del *cv-error* respecto a la profundidad de los árboles para *Extra Trees Regressor*.

En la Figura 59 se nota que el valor de *max\_depth* que presenta el menor *RMSE* es 5, a pesar de ello, un valor superior a este puede mantener el *RMSE* entre 0.008 y 0.010.

Seguidamente, para hallar los mejores hiper-parámetros para R-cuadrado y *RMSE* se usó *GridSearchCV*. En la Figura 60 se muestran los resultados obtenidos.

	param_max_depth	param_max_features	param_n_estimators	mean_test_score	std_test_score
7	None	None	97	0.689497	0.008712
2	None	None	92	0.689463	0.009486
0	None	None	90	0.689295	0.009270
6	None	None	96	0.689287	0.010034

Figura 60. Mejores hiper-parámetros obtenidos con *GridSearchCV* para *Extra Trees Regressor*.

El modelo final optimizado se muestra en la Figura 61.

```

ExtraTreesRegressor
ExtraTreesRegressor(max_features=None, n_estimators=97, random_state=329)

```

Figura 61. Modelo optimizado para *Extra Trees Regressor*.



En la evaluación del modelo optimizado se obtuvo como resultado para la métrica R-cuadrado un valor de 0.829 y para *RMSE* 0.007, mejorando el valor inicial de 0.828 para R-cuadrado y 0.007 para *RMSE*.

#### 4.6.5. *AdaBoost Regressor*

Teniendo en cuenta que este modelo por defecto trabaja con el modelo de *Decision Tree Regressor*, para este caso el objetivo es mejorar la métrica R-cuadrado obtenida en la sección 4.5.2.1.

Para iniciar, se partió del modelo que presentó los hiper-parámetros con las mejores métricas (Figura 43) agregando el hiper-parámetro *n\_estimators* = 300, pero ahora con el meta-estimador, como se muestra en la Figura 62.

```

AdaBoostRegressor
└─ base_estimator: DecisionTreeRegressor
   └─ DecisionTreeRegressor
      DecisionTreeRegressor(ccp_alpha=7.179487e-06, criterion='absolute_error',
                             max_depth=4, max_features='auto', random_state=329)

```

**Figura 62.** Modelo implementado para *AdaBoost Regressor*.

Este modelo generó un resultado en la métrica R-cuadrado de 0.786, mejorando así el resultado que tenía anteriormente de 0.761.

Posteriormente se implementó *GridSearchCV* para hallar el mejor valor para el hiper-parámetro *n\_estimators*, arrojando un valor de 247 con el cual se obtuvo una métrica R-cuadrado de 0.659. En la Figura 63 se observa la tabla de resultados.

	param_n_estimators	mean_test_score	std_test_score
27	247	0.659562	0.048024
26	246	0.652194	0.035947
7	227	0.651492	0.051445
23	243	0.650674	0.035800

**Figura 63.** Mejores hiper-parámetros obtenidos con *GridSearchCV* para *AdaBoost Regressor*.

En la Figura 64 se muestra el modelo optimizado para *AdaBoost Regressor*.

```
AdaBoostRegressor
AdaBoostRegressor(base_estimator=DecisionTreeRegressor(ccp_alpha=7.179487e-06,
                                                       criterion='absolute_error',
                                                       max_depth=4,
                                                       max_features='auto',
                                                       random_state=329),
                 n_estimators=247)
  base_estimator: DecisionTreeRegressor
    DecisionTreeRegressor
    DecisionTreeRegressor(ccp_alpha=7.179487e-06, criterion='absolute_error',
                          max_depth=4, max_features='auto', random_state=329)
```

**Figura 64.** Modelo optimizado implementado para *AdaBoost Regressor*.

Utilizando los hiper-parámetros de la Figura 64 se realizó la evaluación del modelo para así evitar la *CV*, el resultado para la métrica R-cuadrado fue 0.811 y para *RMSE* 0.007, mejorando así el resultado sin optimización que fue de 0.761 para R-cuadrado y 0.008 para *RMSE*.

Por último, se presenta la Tabla 8, la cual contiene los resultados para cada modelo optimizado.

**Tabla 8.** Resultados obtenidos en la optimización de los hiper-parámetros

<b>Modelo</b>	<b>Hiper-parámetros</b>	<b>R-cuadrado Inicial</b>	<b>RMSE Inicial</b>	<b>R-cuadrado optimizado</b>	<b>RMSE optimizado</b>
<i>Decision Tree Regressor</i>	<i>criterion='absolute_error' max_depth=4 max_features='auto' random_state=329 ccp_alpha=7.179e-06</i>	0.611	0.010	0.708	0.009
<i>Random Forest Regressor</i>	<i>max_depth=13 max_features=7 n_estimators=125 random_state=329</i>	0.790	0.007	0.802	0.007
<i>Gradient Boosting Regressor</i>	<i>learning_rate=0.5 max_depth=2 max_features='auto' n_estimators=1000 n_iter_no_change=5 random_state=329 subsample=1</i>	0.758	0.008	0.765	0.008
<i>Hist Gradient Regressor</i>	<i>learning_rate=0.5 max_depth=3</i>	No aplica	No aplica	0.810	0.007
<i>Extra Trees Regressor</i>	<i>n_estimators=97 max_features=None random_state=329</i>	0.828	0.007	0.829	0.007
<i>AdaBoost Regressor*</i>	<i>n_estimators=247</i>	0.761	0.008	0.811	0.007

\* El modelo internamente trabaja con *Decision Tree Regressor*.

La Tabla 8 permite observar que el modelo que mejor resultados presenta es *Extra Trees Regressor* tanto antes como después de la optimización de los hiper-parámetros, logrando así un R-cuadrado de 0.829; valor bastante cercano a uno, y un *RMSE* de 0.007; valor bastante cercano a 0. Considerando lo anterior, se puede dar por cumplido el segundo objetivo específico de este trabajo: “Construir un modelo de aprendizaje automático que permita calcular un índice de vulnerabilidad de la COVID-19, con las variables que se determinen como las más relevantes”. Cabe resaltar que un índice mide la exposición a un riesgo en específico, por lo que, para este trabajo se está midiendo la exposición a contagiarse de COVID-19.

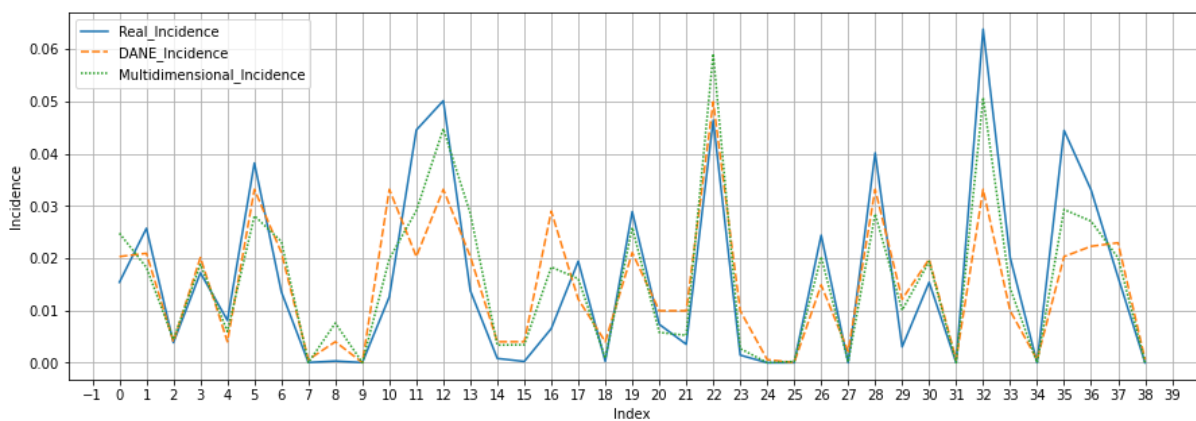
Teniendo en cuenta el modelo que mejores métricas presentó para el índice multidimensional y para el índice del DANE, se procedió a predecir los valores de incidencia considerando las filas comunes en el conjunto de prueba de cada *dataset*, con el objetivo de evitar sesgo en los resultados. A continuación, se presentan las Figuras 65-67 en las cuales se pueden observar los resultados obtenidos.

Index	City	Year	Quarter	Real_Incidence	DANE_Incidence	Multidimensional_Incidence
0	Valledupar	2020	4	0.015410	0.020321	0.024793
1	Armenia	2021	1	0.025731	0.020944	0.018241
2	Bogota	2020	2	0.003851	0.004024	0.004465
3	Santamarta	2020	3	0.017207	0.020144	0.018831
4	Cartagena	2020	2	0.008119	0.004024	0.005976
5	Neiva	2021	2	0.038239	0.033183	0.028120
6	Cartagena	2021	1	0.013558	0.021089	0.023214
7	Neiva	2020	1	0.000058	0.000597	0.000032
8	Sanandres	2020	2	0.000331	0.004024	0.007598
9	Cali	2020	1	0.000035	0.000034	0.000025
10	Florencia	2021	2	0.012629	0.033183	0.019916
11	Neiva	2020	4	0.044538	0.020321	0.029111
12	Valledupar	2021	2	0.050153	0.033183	0.044783
13	Santamarta	2020	4	0.013797	0.020321	0.028564
14	Medellin	2020	2	0.000825	0.004024	0.003386
15	Bucaramanga	2020	2	0.000250	0.004024	0.003446
16	Sincelejo	2020	4	0.006542	0.029000	0.018318
17	Medellin	2021	3	0.019432	0.012161	0.016081
18	Manizales	2020	2	0.000273	0.004024	0.000875
19	Tunja	2021	1	0.028914	0.021089	0.025863
20	Riohacha	2021	4	0.007358	0.009959	0.005806

**Figura 65.** Tabla valores predichos vs valores reales – parte 1.

Index	City	Year	Quarter	Real_Incidence	DANE_Incidence	Multidimensional_Incidence
21	Popayan	2021	4	0.003535	0.009959	0.005252
22	Cartagena	2021	2	0.046409	0.049935	0.059131
23	Monteria	2021	4	0.001464	0.009959	0.002714
24	Pasto	2020	1	0.000005	0.000597	0.000016
25	Ibague	2020	1	0.000017	0.000034	0.000242
26	Ibague	2021	1	0.024407	0.014890	0.020278
27	Sincelejo	2020	1	0.000003	0.002175	0.000121
28	Quibdo	2021	2	0.040171	0.033183	0.028498
29	Manizales	2021	4	0.003040	0.012161	0.010067
30	Riohacha	2020	3	0.015393	0.019812	0.019371
31	Popayan	2020	1	0.000028	0.000591	0.000021
32	Manizales	2021	2	0.063856	0.033183	0.050754
33	Bucaramanga	2021	3	0.020167	0.009959	0.014315
34	Monteria	2020	1	0.000002	0.000591	0.000019
35	Manizales	2020	4	0.044464	0.020321	0.029324
36	Medellin	2020	4	0.033072	0.022291	0.027097
37	Ibague	2020	3	0.016368	0.022960	0.020077
38	Riohacha	2020	1	0.000005	0.000597	0.000015

**Figura 66.** Tabla valores predichos vs valores reales – parte 2.



**Figura 67.** Gráfica de valores predichos vs valores reales

La Figura 67 permite notar que los valores predichos por el índice multidimensional en su mayoría son más cercanos a los valores reales a diferencia de los valores predichos por el índice del DANE. Considerando lo anterior, se puede dar por cumplido el tercer objetivo específico de este trabajo: “Evaluar el modelo creado para calcular un índice de vulnerabilidad de la COVID-19 obtenido, realizando una comparación entre el índice calculado y un índice de referencia.”

Además, se puede dar por cumplido el objetivo general de este trabajo que indica que: “Proponer un modelo de aprendizaje automático, que permita calcular un índice de vulnerabilidad de la COVID-19 que considere de manera integrada factores de riesgo humanos, ambientales, sociodemográficos y socioeconómicos”.

# Capítulo 5.

## 5. Análisis de resultados

En relación con la evaluación del *dataset* inicial se puede afirmar lo siguiente:

- El algoritmo *Decision Tree Classifier* presentó el mejor valor de la métrica *F1-score* (la cual calcula un promedio entre la precisión y la sensibilidad) para el índice inicial. Sin embargo, la métrica tiene un valor bajo (fue de 0.429), evidenciando así que el índice inicial no era el adecuado para determinar la vulnerabilidad.
- Para los dos escenarios ejecutados (80%-20% y 70%-30% en entrenamiento y prueba respectivamente), los valores de las métricas fueron muy cercanos, para todos los modelos propuestos del índice inicial. En algunos casos presentaron el mismo valor, como se presentó para los modelos que utilizaron los algoritmos: *Quadratic Discriminant Analysis*, *KNeighbors Classifier* y *Support Vector Machine*, en dichos modelos el valor de *F1-score* y de *accuracy* fue igual. Lo anterior, permite afirmar que, a pesar de que se variaron los escenarios de entrenamiento y prueba, el índice inicial propuesto no fue óptimo.
- Los algoritmos empleados presentaron en su mayoría un valor alto para la métrica *accuracy*, no obstante, el valor de la métrica *F1-score* fue muy bajo. Esto debido a que *F1-score* es el promedio entre las métricas *precision* y *recall*. Por lo tanto, las predicciones correctas respecto a las predicciones correctas previstas son bajas. Debido a esto no se consideran modelos adecuados para calcular un índice de vulnerabilidad de la COVID-19.
- Para el índice inicial fueron consideradas varias variables sociodemográficas y socioeconómicas, a pesar de ello, el índice no obtuvo resultados óptimos. Esto podría deberse a que no fueron considerados los factores de riesgo humanos, es decir las comorbilidades (para las cuales no se tuvo acceso a los datos), las cuales parece ser que son bastante relevantes.
- Los valores de correlación entre las variables independientes y la variable dependiente fueron nulos, por lo que, se puede afirmar que las variables

utilizadas no eran las apropiadas para un realizar un buen ejercicio que permitiera calcular un índice de vulnerabilidad.

Respecto a la evaluación del *dataset* multidimensional propuesto se puede afirmar lo siguiente:

- El algoritmo *Support Vector Machine* arrojó el peor valor de la métrica R-cuadrado en el índice multidimensional, lo cual puede deberse a que no se obtuvo el hiper-plano adecuado para el modelo. El valor obtenido para dicha métrica (R-cuadrado) fue -2.697.
- El algoritmo *Extra Trees Regressor* para el escenario 80%-20% presentó los mejores valores de las métricas R-cuadrado y *RMSE*. Lo cual permite notar que el modelo multidimensional propuesto predice favorablemente la incidencia de la COVID-19, siendo así, un índice de vulnerabilidad adecuado.
- Después de realizar la optimización de los hiper-parámetros nuevamente el algoritmo *Extra Trees Regressor* obtuvo las mejores métricas, esto puede deberse a que es un meta-estimador, el cual va calculando a partir de varios árboles de decisión las mejores predicciones.
- Después del ajuste de los hiper-parámetros del algoritmo *Decision Tree Regressor* el valor de la métrica R-cuadrado mejoró. Además, se obtuvo un árbol en el cual se redujeron considerablemente la cantidad de nodos y de hojas. Por lo tanto, conjuntamente se lograron mejores resultados y la cantidad de recursos computacionales utilizados fue menor.
- La *CV* interna que realiza la función *GridSearchCV* ocasionaba que el valor de la métrica de R-cuadrado de los algoritmos optimizados no superara el valor de la métrica antes de la optimización. Debido a esto, fue necesario realizar manualmente las evaluaciones, para así evitar la *CV* interna de la función *GridSearchCV* pero, considerando en la evaluación los valores de los hiper-parámetros que arrojaba la función.
- En general, los resultados obtenidos para el índice multidimensional creado se consideran muy buenos, dado que, se obtuvieron valores de *RMSE* cercanos a 0 y valores de R-cuadrado cercanos a 1. Esto indica que el índice tiene en cuenta variables que están relacionadas con la incidencia al COVID-19.

Para el desarrollo de este trabajo, inicialmente se determinaron 3 tipos de factores de riesgo: factores de riesgo humanos, factores de riesgo sociodemográficos y socioeconómicos, y factores de riesgo ambientales. Teniendo en cuenta lo anterior, a continuación, se presenta la clasificación de las variables utilizadas en el índice multidimensional:

- Variable vulnerabilidad: esta variable es de tipo humano y sociodemográfico puesto que incluye los datos de vulnerabilidad ya calculados a partir de las comorbilidades y características de las personas como edad y hacinamiento en hogares.
- PIB: factor de riesgo sociodemográfico y socioeconómico.
- Porcentaje de desempleo: factor de riesgo sociodemográfico y socioeconómico.
- Porcentaje de vacunación: factor de riesgo humano.
- Datos de movilidad: factores de riesgo sociodemográficos y socioeconómicos.
- Temperatura: factor de riesgo ambiental.
- Precipitación: factor de riesgo ambiental.

Lo anterior permite notar que para el índice multidimensional se consideraron variables de diferentes tipos.

Para el índice multidimensional se presentan las Figuras 68-72, las cuales muestran las tablas de importancia de las variables para cada algoritmo, obtenidas mediante el atributo *features\_importances\_* permitiendo visualizar que las variables de porcentaje de vacunación y de movilidad “*residential*” fueron las de mayor importancia en la predicción realizada por la mayoría de los algoritmos.



	predictor	importancia
12	residential	0.445722
11	workplaces	0.288363
13	PorcentajeVacunacion	0.147125
8	grocery_and_pharmacy	0.061674
3	PorcentajeDesempleo	0.030142
6	PIB	0.016249
7	retail_and_recreation	0.006693
2	Trimestre	0.004032
0	CodDepartamento	0.000000
1	Ano	0.000000
4	Temperatura	0.000000
5	Precipitacion	0.000000
9	parks	0.000000
10	transit_stations	0.000000
14	Vulnerabilidad_numero	0.000000

**Figura 68.** Tabla de importancia para el algoritmo *Decision Tree Regressor*.

	predictor	importancia
13	PorcentajeVacunacion	0.200493
12	residential	0.195868
7	retail_and_recreation	0.137760
11	workplaces	0.102729
8	grocery_and_pharmacy	0.089317
9	parks	0.063697
3	PorcentajeDesempleo	0.034238
5	Precipitacion	0.032986
2	Trimestre	0.031703
4	Temperatura	0.031669
6	PIB	0.028556
0	CodDepartamento	0.019073
10	transit_stations	0.017382
1	Ano	0.012330
14	Vulnerabilidad_numero	0.002199

**Figura 69.** Tabla de importancia para el algoritmo *Random Forest Regressor*.

	predictor	importancia
2	Trimestre	0.180776
12	residential	0.177252
13	PorcentajeVacunacion	0.162240
7	retail_and_recreation	0.103196
11	workplaces	0.072263
1	Ano	0.070885
8	grocery_and_pharmacy	0.049209
9	parks	0.039343
5	Precipitacion	0.033362
3	PorcentajeDesempleo	0.032704
4	Temperatura	0.021991
0	CodDepartamento	0.017969
10	transit_stations	0.014853
6	PIB	0.014760
14	Vulnerabilidad_numero	0.009198

**Figura 70.** Tabla de importancia para el algoritmo *Extra Trees Regressor*.

	predictor	importancia
12	residential	0.392922
13	PorcentajeVacunacion	0.365747
11	workplaces	0.120991
8	grocery_and_pharmacy	0.066739
3	PorcentajeDesempleo	0.018035
9	parks	0.016274
10	transit_stations	0.007274
6	PIB	0.004525
7	retail_and_recreation	0.004190
0	CodDepartamento	0.003303
1	Ano	0.000000
2	Trimestre	0.000000
4	Temperatura	0.000000
5	Precipitacion	0.000000
14	Vulnerabilidad_numero	0.000000

**Figura 71.** Tabla de importancia para el algoritmo *Gradient Boosting Regressor*.

	predictor	importancia
13	PorcentajeVacunacion	0.196731
12	residential	0.186605
7	retail_and_recreation	0.103750
11	workplaces	0.080952
9	parks	0.079101
8	grocery_and_pharmacy	0.067650
5	Precipitacion	0.059690
3	PorcentajeDesempleo	0.052787
4	Temperatura	0.048848
2	Trimestre	0.031923
6	PIB	0.031807
0	CodDepartamento	0.030958
10	transit_stations	0.022218
14	Vulnerabilidad_numero	0.005504
1	Ano	0.001475

Figura 72. Tabla de importancia para el algoritmo *AdaBoosting Regressor*.

Respecto a la evaluación realizada al *dataset* del DANE, los resultados presentados en la sección 4.5.2.2. permiten realizar las siguientes afirmaciones:

- El algoritmo *AdaBoost Regressor* presentó las mejores métricas para el escenario 70% entrenamiento y 30% prueba con un *RMSE* de 0.010 y un *R-cuadrado* de 0.610. Sin embargo, para el escenario 80% entrenamiento y 20% prueba el valor de la métrica *R-cuadrado* empeoró, ocasionado que el algoritmo con mejor *R-cuadrado* fuera *Random Forest*.
- Para el algoritmo *Support Vector Machine* nuevamente se observaron valores de las métricas deficientes.

Debido a la alta dispersión de los datos que componen los *datasets*, no fue posible para ninguna de las pruebas realizadas, obtener alguna métrica con un valor perfecto. A pesar de eso, los valores de las métricas obtenidas fueron buenos en general para el índice multidimensional. Sin embargo, se consideró necesario realizar la optimización de los hiper-parámetros, para así obtener mejores modelos y revisar qué tan aceptable era el modelo multidimensional propuesto.

Los resultados permitieron evidenciar que el modelo multidimensional propuesto es el que mejor desempeño obtuvo para la métrica R-cuadrado (comparándolo con el modelo propuesto por el DANE), lo que indica que, el uso de datos de diferentes factores de riesgo arroja un índice con una mejor predicción para la incidencia de casos de la COVID-19 en Colombia. Además, es importante mencionar que la periodicidad de los datos también juega un factor clave, puesto que, el índice del DANE es estático, a diferencia del índice multidimensional creado el cual tiene una periodicidad trimestral. Cabe anotar también, que el modelo multidimensional propuesto fue entrenado para calcular la incidencia de la COVID-19 por lo que, presenta mejores opciones para tareas de predicción, que las del índice desarrollado por el DANE. Lo anterior, se pudo determinar gracias al ejercicio de predicción realizado para los dos índices, el cual arrojó mejores resultados para el modelo multidimensional propuesto.

Del trabajo realizado se está elaborando un artículo para la revista *Journal of Personalized Medicine (JPM)* de *MPDI*, Anexo F denominado “Artículo para la revista JPM de MDPI”, en la cual se exponen el proceso y resultados más importantes del presente trabajo. Cabe resaltar, que este artículo es una versión extendida del artículo que se encuentra en el Anexo B, el cual fue solicitado después de participar en la conferencia de *pHealth 2022* que se llevó a cabo en Oslo, Noruega.

Es de resaltar que, todo el trabajo, recursos, y desarrollo realizados por el DANE Colombia son muy importantes, y constituyen una fuente relevante para los resultados de este trabajo. Los datos recopilados de forma minuciosa, rigurosa y detallada en diferentes municipios del país hacen que el índice de vulnerabilidad del DANE sea una fuente relevante para este trabajo.

Este trabajo es un primer paso para futuras investigaciones, en las cuales se puede tener en cuenta una mayor cantidad de variables consideradas como factores de riesgo de la COVID-19, para así buscar un índice que cada vez realice mejores predicciones de la incidencia de la COVID-19. También es relevante para el cálculo de índices de vulnerabilidad de otros eventos de notificación obligatoria, como por ejemplo el dengue.

# Capítulo 6.

## 6. Conclusiones

A partir de la hipótesis planteada para el presente trabajo, la cual afirma que: “mediante la implementación de un modelo de aprendizaje automático, se puede determinar un índice de vulnerabilidad a la COVID-19 en Colombia, teniendo en cuenta la base de datos de casos históricos de la COVID-19 y los factores de riesgo más relevantes, con el fin de ayudar en la toma de decisiones en los programas de salud pública”; y teniendo en cuenta las etapas de desarrollo del trabajo, fue posible concluir lo siguiente:

Durante la realización del módulo de recolección de datos del índice inicial, se evidenció que el acceso a los datos era muy limitado para calcular un índice que pudiese considerar varios tipos de factores de riesgo. Además, el pre-procesamiento y limpieza de datos requirió un trabajo considerable para poder implementar los modelos de *ML*. A pesar del trabajo realizado, la falta de acceso a datos de variables como comorbilidades arrojó como resultado que los algoritmos utilizados obtuvieran valores bajos en las métricas seleccionadas. Por lo cual, la mejor opción era partir de los datos de índice de vulnerabilidad que ya había calculado el DANE y que incluían ya datos de comorbilidades que no son de dominio público, y agregar a esto datos otras fuentes relevantes.

Calcular un índice multidimensional para la COVID-19 en Colombia fue complejo, ya que se realizó en un país donde a pesar de que existe una política de acceso a datos abiertos, son poco los datos que se encuentran públicos, actualizados y desagregados (como la información obtenida del CNPV donde se tenían divididos los datos entre personas, hogares, vivienda, fallecidos y marco de georreferenciación).

Este trabajo permitió comparar el desempeño del índice de vulnerabilidad realizado por el DANE con el multidimensional propuesto en este trabajo de grado, mediante la predicción de la incidencia de casos de la COVID-19. En ese sentido, se observó que el índice multidimensional propuesto presentó mejores valores que el índice del DANE

en las métricas seleccionadas para la predicción de la incidencia. Por lo cual, se puede afirmar que el índice multidimensional tiene un mejor desempeño, al haber entrenado el modelo para predecir la incidencia. También es importante tener en cuenta que, el índice propuesto considera el valor calculado por el DANE como una de las entradas de datos, y adiciona otras fuentes relevantes de diferentes tipos.

En este trabajo se demuestra cómo el uso de fuentes de datos abiertas permite construir un *dataset* multidimensional, es decir un *dataset* que integra factores de riesgo de diferentes tipos, la propuesta de este *dataset* generó un gran esfuerzo en la etapa de pre-procesamiento debido a la carencia de datos de algunas de las variables seleccionadas para diferentes periodos de tiempo. A pesar de las limitaciones, es una solución interesante que permitió identificar y evaluar otro tipo de factores de riesgo que se encontraron como relevantes en la literatura, como factores importantes en el cálculo de un índice de vulnerabilidad de la COVID-19 como fue el caso de las variables de porcentaje de vacunación y de movilidad “*residential*”, que fueron las de mayor importancia en la predicción realizada por la mayoría de los modelos evaluados.

El trabajo también incluye la evaluación de diversos modelos de *ML*, implementados con varios tipos de algoritmos. Dicha evaluación, fue optimizada, tratando de mejorar los resultados. Todo este proceso arrojó que el modelo que utilizó el algoritmo *Extra Trees Regressor* fue el mejor para predecir la incidencia de la COVID-19 en las ciudades capitales del país.

Finalmente, este trabajo permite evidenciar la necesidad que hay respecto al acceso a datos en Colombia y a los protocolos para datos abiertos y anonimizados, ya que, de esa manera podría ser recolectada información por diferentes grupos u organizaciones con fines investigativos. Además de ello, es habitual la tardanza en la publicación de los datos, así como la liberación de estos en formatos inconsistentes como PDF. Estos casos se presentaron en el desarrollo de este trabajo, ocasionando que para algunas variables fuera necesario añadir manualmente los datos al *dataset*.

En relación a los trabajos futuros, este trabajo presenta las diferentes variables consideradas como factores de riesgo, dando así la posibilidad de revisar qué otras

variables podrían ser añadidas en un nuevo índice de vulnerabilidad multidimensional. De esta forma, se podría proponer un índice que contemple una mayor cantidad de variables consideradas como factores de riesgo de la COVID-19. Adicional a ello, existen otros algoritmos de aprendizaje supervisado que por diversos factores no pudieron ser ejecutados en este trabajo, por lo que, es importante revisar qué tipos de algoritmos aún no se han utilizado para entrenar los modelos y que podrían ser implementados. Inclusive, se podrían evaluar los hiper-parámetros para cada modelo, con mayor profundidad, ya que en este trabajo hubo una limitante respecto a los recursos computacionales disponibles.

## Referencias

- [1] «Información básica sobre la COVID-19». <https://www.who.int/es/news-room/questions-and-answers/item/coronavirus-disease-covid-19> (accedido 10 de febrero de 2022).
- [2] C. Mitchell y <https://www.facebook.com/pahowho>, «OPS/OMS | La OMS caracteriza a COVID-19 como una pandemia», *Pan American Health Organization / World Health Organization*, 11 de marzo de 2020. [https://www3.paho.org/hq/index.php?option=com\\_content&view=article&id=15756:who-characterizes-covid-19-as-a-pandemic&Itemid=1926&lang=es](https://www3.paho.org/hq/index.php?option=com_content&view=article&id=15756:who-characterizes-covid-19-as-a-pandemic&Itemid=1926&lang=es) (accedido 10 de febrero de 2022).
- [3] «El Coronavirus en Colombia». <https://coronaviruscolombia.gov.co/Covid19/> (accedido 12 de septiembre de 2021).
- [4] «COVID-19: cronología de la actuación de la OMS». <https://www.who.int/es/news/item/27-04-2020-who-timeline---covid-19> (accedido 9 de noviembre de 2022).
- [5] «Qué significa que la OMS haya clasificado al coronavirus como pandemia», *BBC News Mundo*. Accedido: 9 de noviembre de 2022. [En línea]. Disponible en: <https://www.bbc.com/mundo/noticias-internacional-51842708>
- [6] «Coronavirus». <https://www.who.int/es/health-topics/coronavirus> (accedido 9 de noviembre de 2022).
- [7] «informe\_cov19\_pti\_salud\_global\_csic\_v2\_1.pdf». Accedido: 5 de noviembre de 2022. [En línea]. Disponible en: [https://www.csic.es/sites/default/files/informe\\_cov19\\_pti\\_salud\\_global\\_csic\\_v2\\_1.pdf](https://www.csic.es/sites/default/files/informe_cov19_pti_salud_global_csic_v2_1.pdf)
- [8] «Casos positivos de COVID-19 en Colombia | Datos Abiertos Colombia». <https://www.datos.gov.co/Salud-y-Proteccion-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr> (accedido 9 de noviembre de 2022).
- [9] «Diccionario de Acción Humanitaria». <https://www.dicc.hegoa.ehu.eus/listar/mostrar/228> (accedido 22 de septiembre de 2021).
- [10] «Medición de la Vulnerabilidad Multidimensional del Estudiante», *JUNAEB*. <https://www.junaeb.cl/medicion-la-vulnerabilidad-ivm> (accedido 22 de septiembre de 2021).

- [11] de Groot, Nynke & Bonsel, Gouke & Birnie, Erwin & Valentine, Nicole. (2019). Towards a universal concept of vulnerability: Broadening the evidence from the elderly to perinatal health using a Delphi approach. PLOS ONE. 14. e0212633. 10.1371/journal.pone.0212633.
- [12] «Nota\_metodologica\_indice\_de\_vulnerabilidad.pdf». Accedido: 9 de noviembre de 2022. [En línea]. Disponible en: [https://www.dane.gov.co/files/comunicados/Nota\\_metodologica\\_indice\\_de\\_vulnerabilidad.pdf](https://www.dane.gov.co/files/comunicados/Nota_metodologica_indice_de_vulnerabilidad.pdf)
- [13] D. A. C. Rodriguez, «Comunicado de Prensa», p. 3, 2020.
- [14] K. Pastor-Sierra *et al.*, «Factores ambientales en la transmisión del SARS-CoV-2/COVID 19: panorama mundial y colombiano», *Rev. Univ. Ind. Santander Salud*, vol. 53, dic. 2021, doi: 10.18273/saluduis.53.e:21037.
- [15] «Protegiendo a las comunidades indígenas de la COVID-19 - OPS/OMS | Organización Panamericana de la Salud». <https://www.paho.org/es/historias/protegiendo-comunidades-indigenas-covid-19> (accedido 9 de noviembre de 2022).
- [16] «Lo que debes saber sobre las vacunas contra la COVID-19». <https://www.unicef.org/es/coronavirus/lo-que-debes-saber-sobre-vacuna-covid19> (accedido 9 de noviembre de 2022).
- [17] «“Las vacunas han reducido la mortalidad de manera sustancial”: vicesalud». <https://www.minsalud.gov.co/Paginas/Las-vacunas-han-reducido-la-mortalidad-de-manera-sustancial-vicesalud.aspx> (accedido 9 de noviembre de 2022).
- [18] «En Colombia, no vacunados tienen de 4 a 9 veces más riesgo de morir por covid-19». <https://www.minsalud.gov.co/Paginas/En-Colombia-no-vacunados-tienen-de-4-a-9-veces-mas-riesgo-de-morir-por-covid-19-.aspx> (accedido 9 de noviembre de 2022).
- [19] Economic Commission for Latin America and the Caribbean, *La prolongación de la crisis sanitaria y su impacto en la salud, la economía y el desarrollo social*. United Nations, 2021. doi: 10.18356/9789210016377.
- [20] «¿Qué es PMBOK en gestión de proyectos?». <https://www.wrike.com/es/project-management-guide/faq/que-es-pmbok-en-gestion-de-proyectos/> (accedido 9 de noviembre de 2022).
- [21] «Guidelines for conducting systematic mapping studies in software engineering: An update - ScienceDirect». <https://www.sciencedirect->



com.acceso.unicauca.edu.co/science/article/pii/S0950584915000646?via%3Dihub (accedido 10 de noviembre de 2022).

[22] «Lección 3.1: Principales bases de datos en ciencias de la salud». [https://evidenciaencuidados.es/wp-](https://evidenciaencuidados.es/wp-content/uploads/MOOC/C2/Curso_02_U03_D01_web_b.html)

[content/uploads/MOOC/C2/Curso\\_02\\_U03\\_D01\\_web\\_b.html](https://evidenciaencuidados.es/wp-content/uploads/MOOC/C2/Curso_02_U03_D01_web_b.html) (accedido 17 de marzo de 2023).

[23] M. Suárez Lastra *et al.*, «Índice de vulnerabilidad ante el COVID-19 en México», *Investig. Geográficas*, n.º 104, 2021, doi: 10.14350/rig.60140.

[24] A. Barrera, A. Bonilla, S. Espinosa, J. González, C. Santelices, y J. Villavicencio, «Índice de vulnerabilidad y trayectorias espaciales del COVID-19 en el Distrito Metropolitano de Quito», *Geopolíticas Rev. Estud. Sobre Espac. Poder*, vol. 12, n.º 1, Art. n.º 1, mar. 2021, doi: 10.5209/geop.70908.

[25] A. Tiwari, A. V. Dadhania, V. A. B. Rangunathrao, y E. R. A. Oliveira, «Using machine learning to develop a novel COVID-19 Vulnerability Index (C19VI)», *Sci. Total Environ.*, vol. 773, p. 145650, jun. 2021, doi: 10.1016/j.scitotenv.2021.145650.

[26] «Censo Nacional de Población y Vivienda 2018». <https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018> (accedido 10 de noviembre de 2022).

[27] «COLOMBIA - Censo Nacional de Población y Vivienda - CNPV - 2018 - Data Dictionary».

[http://microdatos.dane.gov.co/index.php/catalog/643/data\\_dictionary#page=F9&tab=data-dictionary](http://microdatos.dane.gov.co/index.php/catalog/643/data_dictionary#page=F9&tab=data-dictionary) (accedido 10 de noviembre de 2022).

[28] V. G. Cortina, «Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario».

[29] «Jupyter Notebook: qué es y cómo se usa». <https://ebac.mx/blog/jupyter-notebook> (accedido 9 de noviembre de 2022).

[30] «Python: qué es, para qué sirve y cómo se programa | Informática Industrial», *aula21 | Formación para la Industria*, 8 de octubre de 2020. <https://www.cursosaula21.com/que-es-python/> (accedido 9 de noviembre de 2022).

[31] «Python y la Ciencia de Datos», *Auribox Training*, 7 de marzo de 2017. <https://blog.auriboxtraining.com/business-intelligence/python-y-la-ciencia-de-datos/> (accedido 9 de noviembre de 2022).

- [32] «Scikit-Learn, herramienta básica para el Data Science en Python», *Máster en Data Science*, 6 de agosto de 2018. <https://www.master-data-scientist.com/scikit-learn-data-science/> (accedido 9 de noviembre de 2022).
- [33] P. A. Rosero, J. S. Realpe, C. D. Farinango, D. S. Restrepo, R. Salazar-Cabrera, y D. M. Lopez, «Risk Factors for COVID-19: A Systematic Mapping Study», *PHealth* 2022, pp. 63-74, 2022, doi: 10.3233/SHTI220964.
- [34] «Vulnerabilidad». <https://geoportal.dane.gov.co/visor-vulnerabilidad/> (accedido 1 de diciembre de 2022).
- [35] «ML | One Hot Encoding to treat Categorical data parameters», *GeeksforGeeks*, 12 de junio de 2019. <https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/> (accedido 18 de marzo de 2023).
- [36] «Glosario - Gran encuesta integrada de hogares». <https://www.dane.gov.co/index.php/28-espanol/sociales/mercado-laboral/422-glosario-gran-encuesta-integrada-de-hogares> (accedido 20 de marzo de 2023).
- [37] «Páginas - Ciclo de Vida». <https://www.minsalud.gov.co/proteccionsocial/Paginas/cicloVida.aspx> (accedido 18 de noviembre de 2022).
- [38] «Python estadísticas.median() Método». [https://www.w3schools.com/python/ref\\_stat\\_median.asp](https://www.w3schools.com/python/ref_stat_median.asp) (accedido 28 de diciembre de 2022).
- [39] «tutorial EDA para Data Science». <https://kaggle.com/code/micheldc55/tutorial-eda-para-data-science> (accedido 21 de marzo de 2023).
- [40] «Una comparación de los métodos de correlación de Pearson y Spearman». <https://support.minitab.com/es-mx/minitab/20/help-and-how-to/statistics/basic-statistics/supporting-topics/correlation-and-covariance/a-comparison-of-the-pearson-and-spearman-correlation-methods/> (accedido 9 de noviembre de 2022).
- [41] J. F. V. Rueda, «Aprendizaje supervisado y no supervisado», *healthdataminer.com*, 4 de agosto de 2019. <https://healthdataminer.com/data-mining/aprendizaje-supervisado-y-no-supervisado/> (accedido 21 de marzo de 2023).
- [42] «Una Guía Para Principiantes Sobre La Regresión Lineal En Python Con Scikit-Learn», *DataSource.ai*, 10 de diciembre de 2020. <https://www.datasource.ai/es/data-science-articles/view-source:https://www.datasource.ai/es/data-science-articles/una->

guia-para-principiantes-sobre-la-regresion-lineal-en-python-con-scikit-learn

(accedido 9 de noviembre de 2022).


[43] Na8, «Sets de Entrenamiento, Test y Validación», *Aprende Machine Learning*, 3 de marzo de 2020. <https://www.aprendemachinelarning.com/sets-de-entrenamiento-test-validacion-cruzada/> (accedido 20 de marzo de 2023).

[44] P. R. de los S. E. en generación de contenidos tecnológicos para los canales digitales de T. T. A. of T. L. en C. F. y M. en T. E. A. por las “tecnologías para la vida” y L. Q. N. H. L. V. M. F. P. L. Pedagogía, «Datos de entrenamiento vs datos de test», *Think Big*, 24 de enero de 2022. <https://empresas.blogthinkbig.com/datos-entrenamiento-vs-datos-de-test/> (accedido 20 de marzo de 2023).

[45] «1.2. Linear and Quadratic Discriminant Analysis», *scikit-learn*. [https://scikit-learn/stable/modules/lda\\_qda.html](https://scikit-learn/stable/modules/lda_qda.html) (accedido 1 de diciembre de 2022).


[46] «ANÁLISIS DISCRIMINANTE LINEAL Y CUADRÁTICO». [https://rstudio-pubs-static.s3.amazonaws.com/389151\\_3cfc2588daff4989b0ce8da8b3d5ab01.html](https://rstudio-pubs-static.s3.amazonaws.com/389151_3cfc2588daff4989b0ce8da8b3d5ab01.html) (accedido 24 de noviembre de 2022).

[47] R. Python, «The k-Nearest Neighbors (kNN) Algorithm in Python – Real Python». <https://realpython.com/knn-python/> (accedido 24 de noviembre de 2022).

[48] L. Gonzalez, «Aprendizaje Supervisado: Decision Tree Classification»,  *Aprende IA*, 23 de marzo de 2018.


[49] P. Sharma, «Implementing Gaussian Naive Bayes in Python», *Analytics Vidhya*, 29 de noviembre de 2021. <https://www.analyticsvidhya.com/blog/2021/11/implementation-of-gaussian-naive-bayes-in-python-sklearn/> (accedido 24 de noviembre de 2022).

[50] S. Ray, «SVM | Support Vector Machine Algorithm in Machine Learning», *Analytics Vidhya*, 12 de septiembre de 2017. <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/> (accedido 08 de enero de 2023).

[51] «Evaluando el error en los modelos de clasificación -  Aprende IA». <https://aprendeia.com/evaluando-el-error-en-los-modelos-de-clasificacion-machine-learning/> (accedido 05 de enero de 2023).

[52] «Producto interno bruto (PIB) | Banco de la República». <https://www.banrep.gov.co/es/glosario/producto-interno-bruto-pib> (accedido 18 de noviembre de 2022).

- [53] «U5 pp 122 regla de tres.pdf». Accedido: 24 de noviembre de 2022. [En línea]. Disponible en: <https://www.mineduc.gob.gt/DIGECADE/documents/Telesecundaria/Recursos%20Digitales/2o%20Recursos%20Digitales%20TS%20BY-SA%203.0/06%20MATEMATICA/U5%20pp%20122%20regla%20de%20tres.pdf>
- [54] «Google Earth Engine». <https://earthengine.google.com> (accedido 3 de julio de 2022).
- [55] «INDICADORES IDEAM». <http://bart.ideam.gov.co/indiecosistemas/ind/temperatura.html> (accedido 21 de marzo de 2023).
- [56] «Cómo calcular la temperatura media anual». <http://es.scienceaq.com/Chemistry/100315406.html> (accedido 21 de marzo de 2023).
- [57] «Microsoft Power BI». <https://app.powerbi.com/view?r=eyJrljoiNTNmZTJmZWYtOWFhMy00OGE1LWFiNDAtMTJmYjM0NDk1NGY2IiwidCI6ImJmYjdlMTNhLTdmYjctNDAxNi04MzBjLWQzNzE2ZThkZDhiOCJ9> (accedido 5 de agosto de 2022).
- [58] «Empleo y desempleo». <https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo> (accedido 18 de noviembre de 2022).
- [59] «COVID-19 Community Mobility Report», *COVID-19 Community Mobility Report*. <https://www.google.com/covid19/mobility?hl=en> (accedido 18 de noviembre de 2022).
- [60] «divipola2007.pdf». Accedido: 24 de noviembre de 2022. [En línea]. Disponible en: <https://www.dane.gov.co/files/investigaciones/divipola/divipola2007.pdf>
- [61] «secme-21228.pdf». Accedido: 20 de marzo de 2023. [En línea]. Disponible en: <http://ri.uaemex.mx/bitstream/handle/20.500.11799/32032/secme-21228.pdf?sequence=1&isAllowed=y>
- [62] «Linear Regression in Machine learning - Javatpoint». <https://www.javatpoint.com/linear-regression-in-machine-learning> (accedido 08 de enero de 2023).
- [63] «¿Qué es el algoritmo de k vecinos más cercanos? | IBM». <https://www.ibm.com/co-es/topics/knn> (accedido 08 de enero de 2023).
- [64] «What is Random Forest? | IBM». <https://www.ibm.com/topics/random-forest> (accedido 29 de diciembre de 2022).

- [65] «Gradient Boosting con Python». [https://www.cienciadedatos.net/documentos/py09\\_gradient\\_boosting\\_python.html](https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html) (accedido 9 de diciembre de 2022)
- [66] «sklearn.ensemble.ExtraTreesRegressor», *scikit-learn*. <https://scikit-learn/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html> (accedido 05 de enero de 2023).
- [67] «sklearn.ensemble.AdaBoostRegressor», *scikit-learn*. <https://scikit-learn/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html> (accedido 05 de enero de 2023).
- [68] L. Gonzalez, «Evaluando el error en los modelos de regresión»,  *Aprende IA*, 23 de noviembre de 2018. <https://aprendeia.com/evaluando-el-error-en-los-modelos-de-regresion/> (accedido 21 de marzo de 2023).
- [69] S. DELSOL, «▷ Coeficiente de determinación ¿Qué es?», 3 de junio de 2019. <https://www.sdelsol.com/glosario/coeficiente-de-determinacion/> (accedido 21 de marzo de 2023).
- [70] «How do you evaluate the trade-off between accuracy and simplicity in decision tree pruning?» <https://www.linkedin.com/advice/3/how-do-you-evaluate-trade-off-between-accuracy> (accedido 21 de marzo de 2023).
- [71] «Arboles de decision python». [https://www.cienciadedatos.net/documentos/py07\\_arboles\\_decision\\_python.html](https://www.cienciadedatos.net/documentos/py07_arboles_decision_python.html) (accedido 05 de enero de 2023).
- [72] «Post pruning decision trees with cost complexity pruning», *scikit-learn*. [https://scikit-learn/stable/auto\\_examples/tree/plot\\_cost\\_complexity\\_pruning.html](https://scikit-learn/stable/auto_examples/tree/plot_cost_complexity_pruning.html) (accedido 9 de diciembre de 2022).
- [73] «Random Forest con Python». [https://www.cienciadedatos.net/documentos/py08\\_random\\_forest\\_python.html](https://www.cienciadedatos.net/documentos/py08_random_forest_python.html) (accedido 08 de enero de 2023).

## Anexo A. Revisión índices de vulnerabilidad

El Anexo A presenta la revisión de los índices de vulnerabilidad realizados en diferentes países.

Disponible en el siguiente enlace:

[https://docs.google.com/presentation/d/13UPAeW6-9IPvc2vCPFwPKexSksb0GQrmsHbgjp\\_rWbM/edit?usp=sharing](https://docs.google.com/presentation/d/13UPAeW6-9IPvc2vCPFwPKexSksb0GQrmsHbgjp_rWbM/edit?usp=sharing)

## Anexo B. Artículo generado con la revisión de literatura

El Anexo B presenta el artículo científico desarrollado durante la elaboración del presente trabajo de grado, el artículo presentado fue incluido en las memorias del evento *Phealth 2022* (realizado en Oslo, Noruega) publicadas en acceso abierto en la serie de prensa de IOS "*Studies in Health Technology and Informatics*". La serie es indexada en *Elsevier's EMCare*, *Elsevier's SciVerse Scopus*, *Google Scholar / Google Books*, *MEDLINE*, *PubMed*, *Thomson Reuters' Book Citation Index*, and *Thomson Reuters' Conference Proceedings Citation Index*. El artículo fue presentado en el evento *Phealth 2022*, el día 9 de noviembre de 2022, a través de videoconferencia por parte del Co-Director del trabajo de grado, Profesor Diego Mauricio López G.

Disponible en el siguiente enlace:

<https://ebooks.iospress.nl/volumearticle/61373>

## Anexo C. Repositorio de Github

En el Anexo C se encuentran alojados todos los códigos realizados para la limpieza y pre-procesamiento de los datos así como el diccionario de datos de los *datasets* individuales. Considerando lo anterior, en este repositorio se pueden observar todos los códigos realizados para llegar al *dataset* inicial, así como los códigos empleados para obtener el índice multidimensional. Adicional a ello, están las evaluaciones realizadas con los diferentes algoritmos que se tuvieron en cuenta para el desarrollo del trabajo.

Disponible en el siguiente enlace:

[https://github.com/Sebas-Realpe/Indice\\_vul\\_COVID19.git](https://github.com/Sebas-Realpe/Indice_vul_COVID19.git)



## Anexo D. *Dataset* inicial

El Anexo D presenta el *dataset* para el índice de vulnerabilidad inicial.

Disponible en el siguiente enlace:

<https://www.kaggle.com/datasets/sebastianrgonzalez/initial-dane-covid19-dataset>

## Anexo E. *Dataset* del índice multidimensional

El Anexo E presenta el *dataset* para el índice de vulnerabilidad multidimensional.

Disponible en el siguiente enlace:

<https://www.kaggle.com/datasets/sebastianrgonzalez/covid19-colombia>

## Anexo F. Artículo para la revista JPM de MDPI

En el Anexo F se presenta el borrador del artículo que se está preparando como versión extendida del artículo que se presenta en el Anexo B de este documento.

Disponible en el siguiente enlace:

<https://drive.google.com/file/d/1ubOZcdPGIEd7FsXT4N8EFZgpe3gsi3fi/view?usp=sharing>