INGENIERÍAS USBMed

https://revistas.usb.edu.co/index.php/ingUSBmed

ARTÍCULO DE INVESTIGACIÓN

Patrones de incidencia del deterioro del arbolado urbano de Bogotá

Incidence Patterns of the Deterioration of Urban Trees in Bogota

Abraham Ramírez Sánchez[1]
Cristián Camilo Hurtado Vasquez[2]
Max Alejandro Triana Gómez[3]

[1]Ingeniería Forestal, Facultad del Medio Ambiente y Recursos Naturales, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.
Email: abramirezs@correo.udistrital.edu.co

[2]Ingeniería Forestal, Facultad del Medio Ambiente y Recursos Naturales, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.
Email: cchurtadov@correo.udistrital.edu.co

[3] Ingeniería Forestal, Facultad del Medio Ambiente y Recursos Naturales, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.
Email: mtriana@udistrital.edu.co

Tipo de artículo: Investigación.
Recibido: 15-10-2019.
Revisado: 21-04-2020.
Aprobado: 11-05-2020.
Doi: 10.21500/20275846.4344

Referenciar así:
A. Ramírez, C. C. Hurtado and M. A. Triana, "Patrones de incidencia del deterioro del arbolado urbano de Bogotá," *Ingenierías USBMed*, vol. 11, no. 2, pp. 13-26, 2020.

Disponibilidad de datos:
todos los datos relevantes están dentro del artículo, así como los archivos de soporte de información.

Conflicto de intereses:
los autores han declarado que no hay conflicto de intereses.

Editor: Andrés Felipe Hernández.
Universidad de San Buenaventura, Medellín, Colombia.

Resumen. En este artículo se presentan las relaciones geoespaciales entre el arbolado urbano de Bogotá y los registros promedio anuales de contaminación, área de influencia de las vías y sitios de interés (Catastro). Se utilizaron datos para el año 2007 de las variables a relacionar. La metodología utilizada es CRISP-DM (Cross Industry Standard Process for Data Mining). Los patrones de relaciones se evaluaron bajo los algoritmos de árboles de decisión, que definen las variables relevantes para cada Dataset y el análisis clúster, el cual analiza las variables relevantes respecto a los tipos de afectaciones evaluadas, teniendo que las variables relevantes de cada Dataset fueron: especie, porcentaje de afectación general, porcentaje de afectación en tronco, sitios de interés, ancho de vías, número de carriles y localidades. Los resultados mostraron que las especies Urapán y Sauco presentaron relaciones con las variables evaluadas, siendo Urapán la que se relacionaba con las demás variables de cada Dataset. Esta investigación servirá para la toma de decisiones respecto al manejo del arbolado urbano de Bogotá.

Palabras Clave. Dataset, CRISP-DM, árboles de decisión, análisis clúster, fitosanidad, minería de datos.

Abstract. In this paper are presented geospatial relations between Bogota urban trees and the registers of average annual pollution, roads influence area and places of interest (Cadastre). For the variables of study there has been used data for year 2007. The methodology used in this case is CRISP-DM (Cross Industry Standard Process for Data Mining). The relations patterns have been evaluated under the algorithms of decision trees which define the relevant variables for each Dataset and cluster analysis which analyses the relevant variables regarding the evaluated types of affectations, assuming that the relevant variables of each Dataset were: specie, general affectation percentage, trunk affectation percentage, places of interest, roads width, number of lanes and District locations. These results showed that species like Urapán and Sauco were those that presented relations with the evaluated variables, being the Urapán the one that related most with the rest of the variables in each Dataset. This investigation will help to make decisions about the management of the urban trees in the city of Bogotá.

Keywords. Dataset, CRISP-DM, decision trees, cluster analysis, phytosanitary, data mining.

I. Introducción

Hoy en día Bogotá presenta diversas problemáticas relacionadas con la calidad del aire. Estas afectaciones recaen sobre la proyeccion de más de ocho millones de personas en 2020 [1]. Se ha demostrado que las industrias y el tráfico vehicular repercuten en la reducción de la calidad del aire y de la vida, aumentando las emisiones de material particulado respirable (Mp 10) [2]. Para Zamudio [3] la capital de Colombia es una de las ciudades más contaminadas de América Latina, este autor resalta la importancia de la vegetación como barrera protectora de la contaminación atmosférica, puesto que cumple un papel esencial en la captación de dióxido de carbono (Co2) y generación de oxígeno. Además, funciona como sumidero de partículas contaminantes del ambiente. La prestación de estos servicios ambientales mejoran la calidad de vida afectada por el incremento acelerado de la urbanizacion [4]. Sin embargo, las especies vegetales no presentan la misma resistencia a los contaminantes. Esto, y las características físico-mecánicas inherentes a cada especie, se debería tener en cuenta en la selección del arbolado urbano, con el fin de evitar problemas a futuro para las personas, la infraestructura y las redes eléctricas. Una mala selección puede manifestarse en la muerte del individuo forestal, generando riesgos de volcamiento y caída de ramas.

Para Waring [5] el deterioro de la salud de un árbol y su eventual muerte es un fenómeno complejo en el cual interactúan varios factores, los cuales se suelen clasificar en abióticos y bióticos o físico-ambientales. Entre las causas bióticas se pueden mencionar senescencia, desequilibrio mecánico, inanición, respiración herbívora, enfermedades, plagas y competencia [5]. Para las abióticas se tiene la contaminación atmosférica, viento, daños directos (malas podas, anillamiento, reducción de espacio esencial de desarrollo del individuo, entre otras). Algunas especies son más susceptibles a presentar estas afectaciones, reaccionando de forma similar a estos factores.

Dentro del campo de la ciencia de datos, resolver problemas y responder preguntas a través del análisis de información es una práctica estándar. A menudo, los científicos de datos construyen modelos para predecir resultados o descubrir patrones subyacentes, con el objetivo de obtener conocimientos [6]. De acuerdo a ello la minería de datos o Data mining (DM), a través del aprendizaje computacional, presenta diversas soluciones robustas para la ciencia de datos. El objetivo principal de la minería de datos es el análisis de grandes volúmenes de datos para la obtención de modelos y patrones predictivos o descriptivos [7].

El objetivo de este estudio es seleccionar e implementar una metodología que permita encontrar patrones de incidencia en el deterioro del arbolado urbano de Bogotá, a través de minería de datos (Data minning) y herramientas geomáticas basadas en información abierta del Distrito de Bogotá.

II. Área de estudio

Este estudio se centra en la ciudad de Bogotá D.C., que tiene una precipitación promedio anual de 866 mm, una temperatura entre los 7 °C y los 19 °C, una altura promedio sobre el nivel del mar de 2630 m y una zona de vida de bosque húmedo montano bajo (bh-MB) [8], teniendo características de lluvia hacia los cerros orientales.

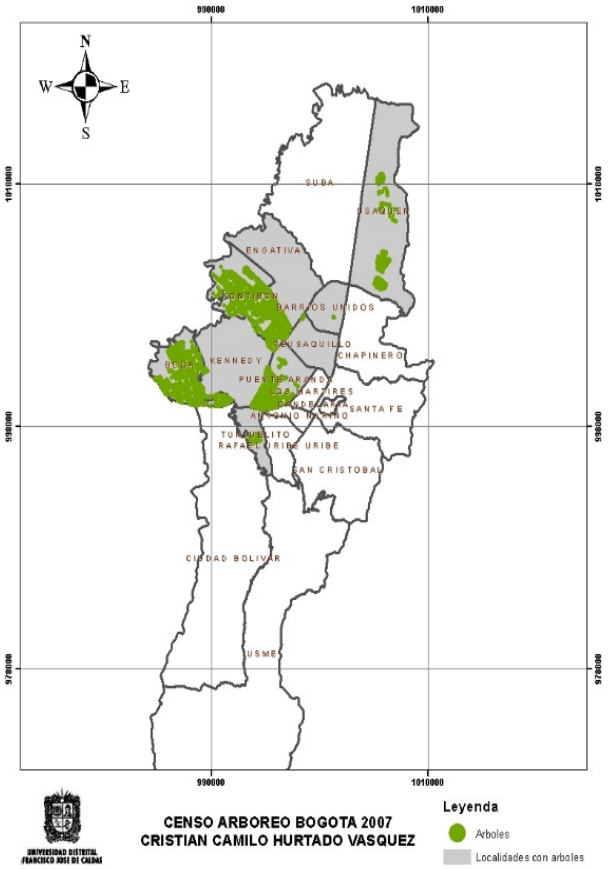

Figura 1. Distribución espacial de los árboles para el censo arbóreo del año 2007. Fuente: Autores

III. Metodología

Se utilizó la metodología *Cross Industry Standard Process for Data Mining* (Crisp DM), que es la guía metodológica más utilizada en desarrollo de proyectos de minería de datos [9]. Esta consiste en un conjunto de tareas que están organizadas en cuatro niveles de abstracción: fases, tareas generales, tareas especializadas e instancias de proceso. Dichos niveles están establecidos respetando jerarquías en tareas; inician en el nivel más general hasta llegar, finalmente, a los casos más específicos [10]. En la implementación de

esta metodología se acogen las cinco primeras fases: comprensión del problema (que se aborda en la introducción), comprensión de datos, preparación de datos, modelado y evaluación. Esta metodología presenta la ventaja de no haber sido construída de manera teórica y académica, sino de estar basada en las experiencias de otros investigadores, por lo que se ajusta plenamente al objetivo del presente estudio [11].

Dentro de la gran variedad de algoritmos descriptivos [12] se decide trabajar con árboles de decisión y análisis clúster. El primero de ellos establece las reglas entre las variables y una clase, en este caso es la enfermedad que afecta individuos arbóreos, por otra parte, permite establecer cuáles de las variables en cada Dataset eran relevantes para la clase. El análisis clúster [13] permite establecer de manera visual características de cada Dataset y define patrones de comportamiento de enfermedades de las especies arbóreas.

A. Comprensión de los datos

En esta fase se realizó la identificación de datos que relacionan el estado fitosanitario del arbolado urbano con otras variables, se identificaron otras variables como la contaminación ambiental en terminos de las industrias y su cercanía a la infraestructura vial y variables endógenas de cada especie como la predisposición específica a ciertas enfermedades [8], [14], mientras que no se encontraron estudios relacionados con la incidencia de los tipos de sitios aledaños (hospitales, parques y centros comerciales, entre otros). De acuerdo a ello, se establecieron 4 grupos de análisis, que fueron Censo del Arbolado Urbano de 2007 (en adelante Censo 2007), contaminación atmosférica por localidad, infraestructura vial y sitios de interés para evaluar patrones de incidencia de estas enfermedades. En cuanto a la informacion para cada grupo de análisis se encontró en repositorios abiertos y con información geográfica disponible.

• **Dataset:** la consecución de información de datos [15] se encontró en el Archivo Nacional de Datos (ANDA). Este es un sistema de consulta que pone a disposición de la ciudadanía un catálogo de microdatos de uso público de diferentes operaciones estadísticas, que los usuarios pueden explorar, comparar, descargar y procesar a su medida. Allí se encontraron los datos requeridos en dos tipos de formatos: en el Censo 2007 para la contaminación atmosférica en la red de monitoreo de calidad del aire de Bogotá (RMCAB), la cual presenta los datos por estación en texto plano; para infraestructura vial en formato Shapefile y para los sitios de interés en metadatos espaciales de los repositorios de la Unidad Administrativa Especial de Catastro Distrital (UAECD).

B. Preparación de los datos

En esta fase se establecieron las variables de interés, también se realizó la depuración de datos con respecto a cinco especies de árboles y tres enfermedades asociadas a estas especies, las cuales se profundizaron en la sección depuración manual de Dataset. Por otra parte, la información encontrada estaba presentaba de manera desagregada y en diversos formatos, estableciendo la necesidad de realizar geoprocesamientos que permitieran la estandarización de datos, con el fin de exportarla en un solo Dataset (Figura 2).

C. Selección de variables de interés para cada grupo de análisis

Se tenía una amplia cantidad de variables para cada Dataset, por lo que antes de realizar el geoprocesamiento se hizo para cada grupo de análisis una primera depuración manual de variables y datos en variables que no eran de interés para esta investigación. De acuerdo a lo anterior, para cada Dataset se realizó una selección precisa de estas variables, las cuales se presentan a continuación:

Censo 2007: las variables de interés del censo de arbolado urbano [16] se encuentran en la Tabla 1 de los Anexos, que se pasaron por un proceso de discretización en cuanto a nombres de enfermedades, nombres científicos, nombres comúnes, altura, Diámetro a la Altura del Pecho (DAP), porcentaje de afectación general y de tronco y diámetro de copa. En total fueron 13 variables que se tuvieron en cuenta para este Dataset.

Contaminación atmosférica: las variables ambientales se tomaron de los datos históricos promedio anuales de cada una de las variables que se encuentran en la Tabla 2 de los Anexos. Se seleccionaron los datos de 7 estaciones meteorológicas (Barrios Unidos, Kennedy, Puente Aranda, Sevillana, Suba, Tunal, Usaquén). Estas variables fueron discretizadas de acuerdo a las categorías permisibles y estados de alerta de la Red de Monitoreo de Calidad del Aire de Bogotá (RMCAB), por lo cual se tuvieron en cuenta 8 variables en total para este Dataset.

Infraestructura vial: las variables de interés de las vías presentes en los metadatos fueron ancho de carril y número de carriles para cada una de las vías. Se realizó una clasifiación de acuerdo a los valores generales de cada variable, que se presentan en la Tabla 3 de los Anexos.

Sitios de interés: los sitios de interés se presentaron como puntos georreferenciados. La variable metadata de interés fue CodSIFin, la cual contiene los valores de actividad y subactividad de cada uno de los sitios de interés, como parques, hospitales, universidades, entre otros, que se encuentra en la Tabla 4 de los Anexos.

D. Procesamiento SIG

Aunque los Dataset obtenidos tenían diferentes orígenes de información y naturaleza de datos, todos estos fueron georreferenciados mediante el uso del software Arcgis 10.7. Además se realizaron diferentes geoprocesamientos, teniendo como referencia al individuo arbóreo, para que así las relaciones obtenidas entre datos fueran de individuo arbóreo en relación con las demás variables.

De esta forma el árbol con su respectiva enfermedad sería el centro de análisis de este estudio. Todos los geoprocesamientos fueron realizados con este.

Luego de tener toda la información de los 4 grupos de análisis en polígonos (excepto Censo 2007), se realizó un *Join Espacial* para relacionar todas estas variables con cada uno de los individuos arbóreos, de esta forma se estableció el Dataset general.

En la Figura 2 se encuentra el diagrama de flujo del procesamiento de los datos SIG.

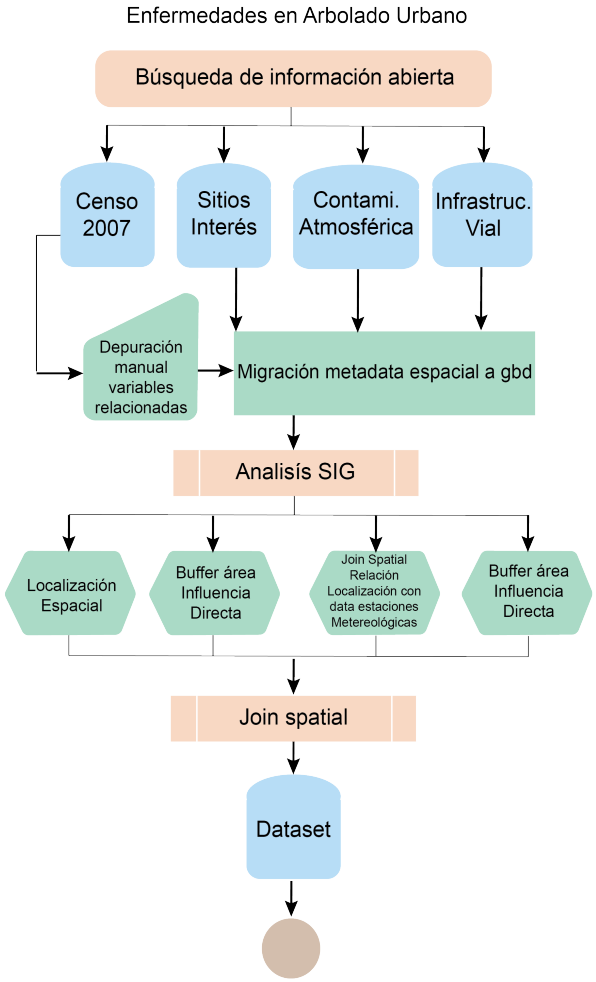

Figura 2. Diagrama de flujo y procesamiento SIG. Fuente: Autores

E. Depuración manual del Dataset

Debido a que los datos de algunas variables eran de tipo numérico, se estableció una discretización en cada una de ellas. Para los porcentajes se valoraron en rangos de 0 a 25%, de 26% a 50%, de 51 a 75% y de 76 a 100%. Las demás variables se realizaron de la siguiente forma:

Variables de Censo 2007: se realiza una discretización de las variables, que se encuentra en la Tabla 5 de los Anexos.

Variables ambientales: la discretización para las variables de contaminación atmosférica se realizaron de acuerdo al criterio permisible de la RMCAB en: favorable, moderada y critica. De igual forma, la clasificación de la temperatura en categoría XIII (13.0 y 13.9 °C), categoría XIV (14.0 y 14.9 °C) y XV (15.0 y 15.9 °C), siendo estos valores promedio anuales de temperatura.

Variables vías: las variables de ancho de carril (CalAncho) y cantidad de carriles (CalNCarril) son de tipo numérico. Se realizó una discretización como se observa en la Tabla 6 de los Anexos.

Variables de sitios de interés: los valores de la variable de sitios de interés (CodSIFin) se abordaron de la misma forma que el catastro lo establece. De esta forma los valores son los mismos que se registran oficialmente para la entidad en los metadatos de los datos espaciales abiertos del Dataset.

Selección de especies arbóreas y enfermedades asociadas: se realizo la cuantificación de las especies con un total de 222 especies, de las cuales se priorizaron las 5 especies arbóreas más frecuentes y con la presencia de por lo menos una enfermedad manifestada por estas especies. Se obtiene que *Sambucus nigra* (Sauco) con 6058 individuos (8.22% del total de árboles censados) es la especie que mayor problemas fitosanitarios presentaba en Bogotá para el año 2007, seguida de *Pittosporum undulatum* (Laurel Huesito) con 3849 individuos (5.22% del total de árboles censados). Continúa *Fraxinus chinensis* (Urapán) con 3843 individuos (5.22% del total de árboles censados). En cuanto a la especie *Ficus soatensis* (Caucho Sabanero), el número es de 3799 árboles (5.16% del total de árboles censados) y por último la *Acacia melanoxylon* (Acacia) con 2783 individuos enfermos (3.78% del total de árboles censados). Entre estas cinco hubo un total de 27.6% de individuos enfermos.

El siguiente filtro realizado fue la delimitación de la investigación a solamente tres tipos de enfermedades (las más frecuentes y que podrían causar muerte en el árbol a corto, mediano y largo plazo): antracnosis (hongos), cáncer y marchitamiento.

Para darle respuesta al problema en la búsqueda de patrones que relacionen las enfermedades en el arbolado urbano y los 4 grupos de variables se seleccionaron dos algoritmos de minería de datos. Por una parte, el árbol de decisión permite que las observaciones acerca de las características de un elemento conduzcan a conclusiones acerca de un valor objetivo, que para esta investigación es la relación de las variables evaluadas con respecto a los diferentes problemas fitosanitarios del arbolado urbano de Bogotá (enfermedad), que en adelante será denominada clase.

Una pequeña explicación del funcionamiento del algoritmo según [17], [18] es: sea T un grafo acíclico dirigido, en el cual se cumple que cada nodo del grafo es:

1. Un nodo no terminal o interno si tiene p nodos hijos, p1. Los nodos internos están etiquetados con atributos.
2. Un nodo terminal u hoja si el nodo no tiene nodos hijos. Los nodos terminales están etiquetados con clasificaciones. El conjunto de todas las hojas de T se llama $T^{\wedge}$.

Cada nodo tiene exactamente un padre, a excepción del nodo superior o raíz, que no tiene padre. Cada arco o rama del grafo que sale de un nodo etiquetado con un atributo ai está a su vez etiquetado con alguno de los posibles valores v de ai. Los nodos internos equivalen a pruebas de un atributo y las ramas que salen de un nodo equivalen a los resultados para la prueba.

Notación 1.1

Se utilizará la siguiente notación:

El tamaño del árbol T se denota por |T|, y el tamaño del conjunto de terminales $T^{\wedge}$ se denota por $T^{\wedge}$.

Al conjunto de todos los atributos que conforman el árbol de decisión se le llamará A y su tamaño será |A|.

Al conjunto de todas las clases se les llama Y y cada clase tiene asignada un valor entero, de tal manera que $C = \{0, 1, ..., K-1\}$ donde $K \geq 2$.

Por lo tanto, con esta definición se tiene que la cantidad de nodos del árbol es |T|, mientras que la cantidad de hojas es $T^{\wedge}$.

La estructura del árbol de la Figura 3 muestra los nodos internos, ramas y hojas. Un nodo interno representa la prueba de un atributo, mientras que las ramas representan los diferentes valores que puede tomar el atributo. Las hojas representan los valores de clasificación para la secuencia de pruebas que va desde el nodo raíz hasta llegar a la hoja.

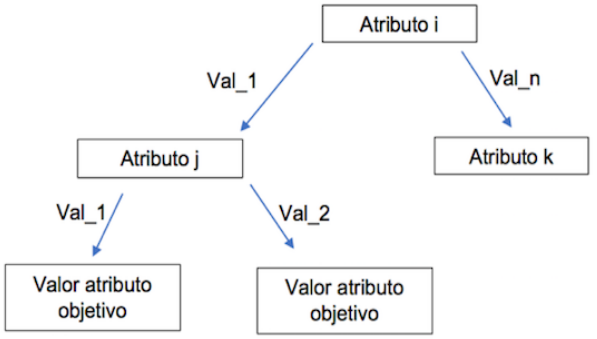

Figura 3. Estructura conceptual de un árbol de decisión. Fuente: Autores

Las entradas del árbol de decisión son un conjunto de características o atributos que pueden representar objetos o situaciones. La salida del árbol es un valor correspondiente al atributo que se desea conocer. Se puede también entender la salida del árbol como un valor booleano (Si/No) para los diversos valores de atributo de salida.

La Tabla 1 muestra una tabla de datos usual, tal como se obtiene a partir de las bases de datos originales del usuario. La fila de encabezados de las n primeras columnas muestran todos los atributos considerados y por lo tanto contienen las pruebas para los nodos internos. Las filas $1, \ldots, m$ de las primeras n columnas son valores para los atributos considerados y por lo tanto contienen las etiquetas de las ramas.

Las filas $1, \ldots, m$ de la última columna son los valores de las clasificaciones para las filas correspondientes y por lo tanto contienen las clases que se asignan para los nodos terminales. Es importante notar que algunos valores se pueden repetir, ya que no es forzoso que todos los m valores de cada columna sean diferentes.

Tabla 1. Datos típicos con m observaciones.

	Atrib.1	**Atrib.**2	...	**Atrib.**n	**Valor objetivo**
Nombres atributos	A	B	...	N	CLASE
Patrón 1	a_1	b_1	...	n_1	Y_1
Patrón 2	a_2	b_2	...	n_2	Y_2
...	...	...	...	...	...
Patrón m	a_m	b_m	...	n_m	Y_m

Para la implementación de la solución del problema se trabajó con el algoritmo J48 en la herramienta WEKA, con la clase weka.classifers.j48.J48.java.

Algunas propiedades concretas de la implementación son las siguientes:

Los tipos de atributos admitidos pueden ser simbólicos y numéricos.

El algoritmo no posibilita la generación de reglas de clasificación a partir del árbol de decisión.

El cálculo de la entropía y de la ganancia de información se realiza con las siguientes ecuaciones:

$$G(A_i) = \frac{n_{ic}}{n^2}(I - I(A_i)) \tag{1}$$

$$I = n_{ic}\log_2(n_{ic}) - \sum_{c=1}^{nc} n_c \log_2(n_c) \tag{2}$$

$$\begin{aligned} I(A_i) &= \sum_{j=1}^{nv(A_i)} n_{ij}\log_2(n_{ij}) - I_{ij}; \\ I_{ij} &= -\sum_{k=1}^{nc} n_{ijk}\log_2(n_{ijk}) \end{aligned} \tag{3}$$

F. Análisis clúster

Para realizar el análisis clúster se utiliza el algoritmo *K-Means*, que divide los objetos en un número de clústeres preespecificado, sin atender a una estructura jerárquica. Puede aplicarse para problemas de "agrupación

por similitud" y ayuda en la investigación a una comprensión cualitativa y cuantitativa de las grandes cantidades de datos N-dimensionales. Funciona de forma iterativa, dividiendo óptimamente el conjunto inicial de datos en un número (K) de clústeres, el cual se indica como parámetro. Se puede observar en la Figura 4.

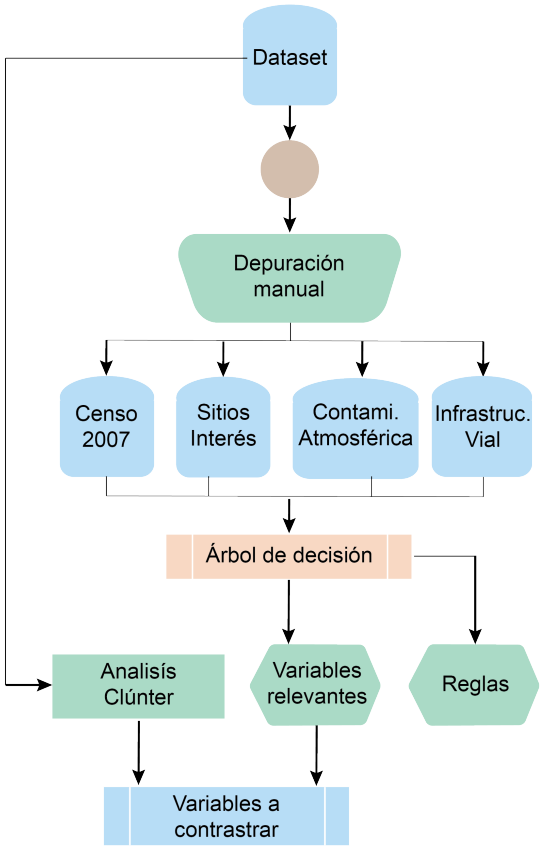

Figura 4. Diagrama de flujo fase modelamiento y evaluación. Fuente: Autores

IV. Resultados

A. Modelado y evaluación

1) Censo 2007

Para el análisis de Censo 2007 que tiene como objetivo buscar los patrones de las variables endógenas de cada una de las cinco especies evaluadas, siendo las variables en total 14, al implementar el algoritmo J48, se estableció que las variables: nombre común, porcentaje afectación general, interferencia en redes y altura discretizada son las variables relevantes para la clase enfermedad, aunque no se encontraron patrones relacionados a la enfermedad cáncer. Las reglas generadas a partir del árbol de decisiones del grupo de análisis de Censo 2007 que se establecen para la variable Especie se presenta como un patrón para las especies Laurel Huesito, Acacia, Caucho Sabanero y Sauco. Se relacionan directamente con una enfermedad, específicamente con marchitamiento. Con ello se concluye que para las variables endógenas (exceptuando Urapán) se relacionan directamente con marchitamiento, las otras variables no son relevantes.

En cuanto a la especie Urapán, se encontró que es un árbol que relaciona sus enfermedades con el grado de afectación general, en donde si presenta porcentajes de afectación mayores a 50% se relaciona con la enfermedad antracnosis, pero si presenta porcentajes entre 26% y 50% presenta marchitamiento. Para los valores de afectación general menor a 25% se encontró relación con la variable interferencia en redes, donde si hay interferencia en redes se relaciona con la variable de Categoría de altura. De acuerdo a esto, si hay un Urapán de tipo arbolito y árbol mediano con interferencia en redes y un porcentaje de afectación menor a 25% presenta marchitamiento. Si este es del tipo árbol pequeño y árbol grande presenta antracnosis.

Análisis de matriz de confusión: los valores de rendimiento para el árbol de decisiones realizado con el Dataset Censo 2007 presento 1644 instancias clasificadas correctamente y solamente 267 mal clasificadas, obteniendo un rendimiento de 86.02% y un índice Kappa de 47.7%. La variable cáncer se estableció como indefinida en la clasificación.

Análisis clúster

Enfermedad-Especie-Grado Afectación General: en la Figura 6 se observa que para la especie Sauco entre el 0% y el 50% de porcentaje general de afectación presenta principalmente la enfermedad marchitamiento (Azul). Para la especie Urapán entre un 51% y 100% padece mayor afectación de la enfermedad antracnosis (Rojo) en Bogotá. La especie Acacia es la que menos afectación de enfermedades posee, presentándose solo algunos casos por marchitamiento.

El Caucho sabanero, que se observa en la Figura 6, es la única especie que padece cáncer (Verde) específicamente notorio en el porcentaje de afectación (del 76% al 100%). Por otra parte, se observa cómo la especie Acacia es la que menor cantidad de individuos afectados por enfermedades presenta relacionada con los porcentajes generales de afectación. Por último, en el porcentaje de afectación general, entre 76% y 100% presenta menor cantidad de individuos enfermos, mientras que la cantidad mayor de individuos enfermos se agrupan en los porcentajes de 0% a 25%.

Enfermedad-Especie-Interferencia en redes (Interfe): en la Figura 7 se observa a nivel general que la variable interferencia en redes incide en las enfermedades de todas las especies, esto se confirma en el árbol de decisiones de la Figura 5, donde solamente la especie Urapán es susceptible a la interferencia en redes.

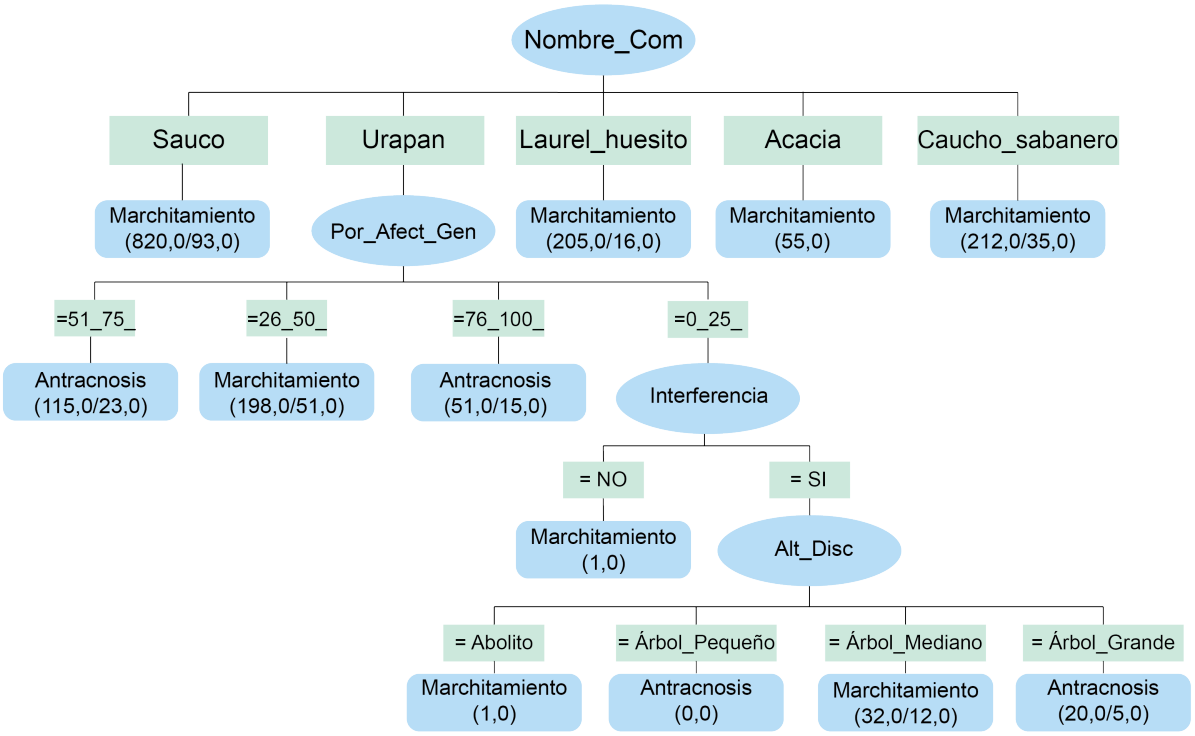

Figura 5. Árbol de decisiones para el Dataset Censo 2007

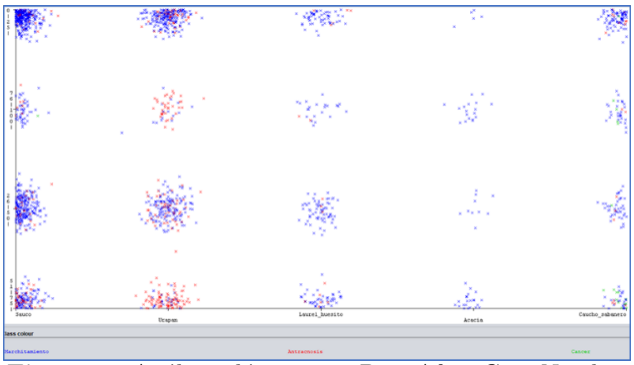

Figura 6. Análisis clúster para Por_Afec_Gen, Nombre común y Enfermedad. Salida programa WEKA

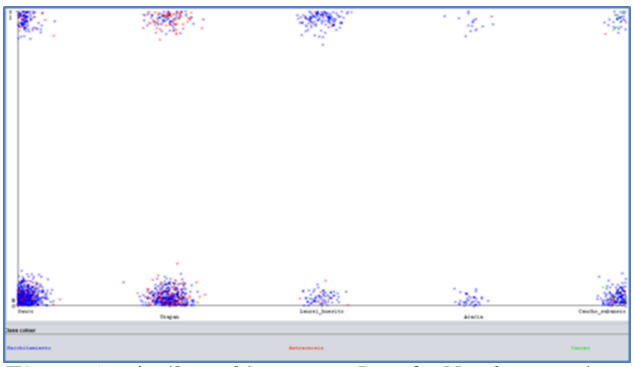

Figura 7. Análisis clúster para Interfe, Nombre común y Enfermedad. Salida programa WEKA

Enfermedad–Especie–Altura Discriminación (Alt. Disc): en el análisis de la Figura 8 se observa cómo en las especies Laurel Huesito y Sauco no se encuentran relaciones con el tamaño del árbol de tipo árbol grande, lo que podría deberse a la baja cantidad de individuos en esta categoría. Para la especie Sauco se encuentran mayor cantidad de individuos enfermos, específicamente por marchitamiento en la clase árbol mediano.

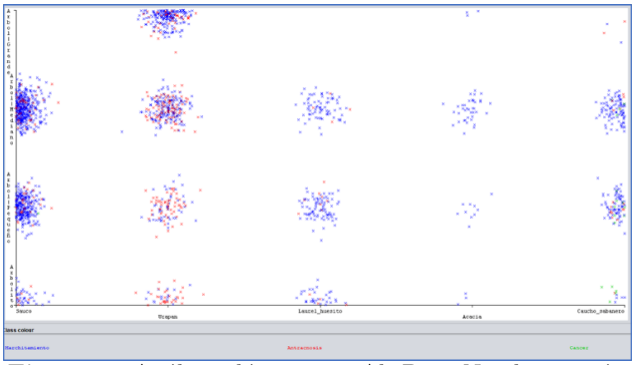

Figura 8. Análisis clúster para Alt.Disc, Nombre común y Enfermedad. Salida programa WEKA

Otra característica predominante que se resalta en este análisis clúster es que el número de individuos enfermos por antracnosis se presenta con mayor frecuencia en Urapán, seguido de Sauco.

2) Sitios de interés

Para el árbol de decisiones de este grupo de análisis (Figura 9) se tiene como objetivo encontrar algunos pa-

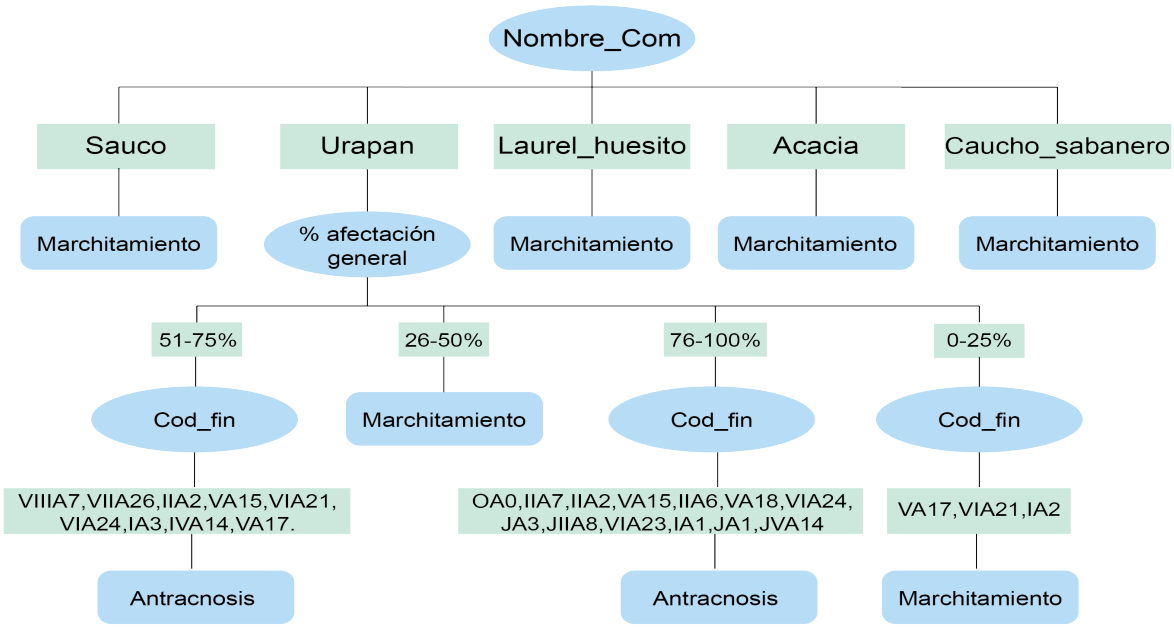

Figura 9. Árbol de decisiones para el Dataset sitios de interés

trones que relacionen las enfermedades de los árboles evaluados con respecto a los sitios de interés. Allí se encontró, al igual que en el anterior grupo de análisis, que las especies de individuos son las variables más relevantes con respecto a la ocurrencia de enfermedades, que, a excepción del Urapán, se relacionan directamente con marchitamiento.

La especie Urapán se relaciona con el porcentaje de afectación general, que si presenta un número de 26% a 50% se relaciona con marchitamiento, mientras que si el porcentaje se encuentra entre 0% y 25% se relaciona con los sitios VA17 (Bibliotecas), VIA21 (Edificios Civiles) y IA2 (Almacenes de cadena) y su enfermedad es marchitamiento. Para los demás sitios se relaciona con antracnosis si la afectación general es mayor a 50%.

Análisis de matriz de confusión: los valores de rendimiento para el árbol de decisiones realizado con el Dataset Sitios de Interés presentó 1628 instancias clasificadas correctamente y solamente 283 mal clasificadas, obteniendo un rendimiento de 85.19% y un índice Kappa de 40.6%. La categoría cáncer se estableció como indefinida en la clasificación.

Análisis clúster

Enfermedad-Especie-Sitios de interés (Cod. Fin): la relación de las enfermedades con los sitios de interés brinda resultados muy interesantes, ya que relaciona una concentración de la mayoría de individuos enfermos en los sitios 0A0 (Residencial), IIIA7 (Parques) y VIIA26 (Institución educativa o académica).

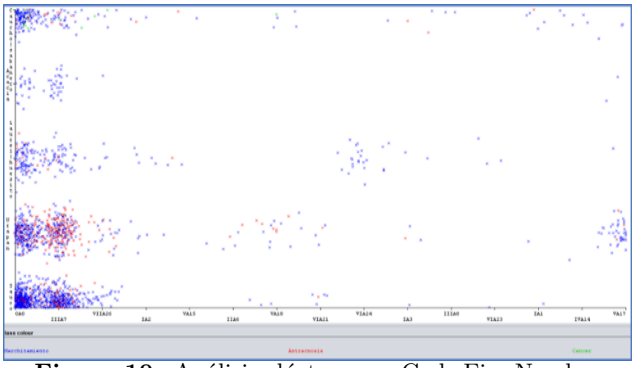

Figura 10. Análisis clúster para Cod. Fin, Nombre común y Enfermedad. Salida programa WEKA

En el análisis de la Figura 10 la especie Laurel Huesito presenta también patrones de relación de enfermedad de marchitamiento con el sitio VIA24 (Industria). La enfermedad cáncer para la especie Caucho Sabanero se relaciona con los sitios 0A0, IIIA7 y IIA26. Los sitios que menos se relacionan con enfermedades son VA15 (Templo o lugar de oración), IIA6 (Aeropuerto), IIIA8 (Complejo deportivo) y IA1 (Centro comercial).

3) Infraestructura vial

En el análisis de la cercanía a vías que relacionen patrones de enfermedades se tuvieron en cuenta solamente dos variables. Por una parte, ancho de carril y por otra, la cantidad de carriles. Estas están relacionadas con la velocidad de circulación vehicular.

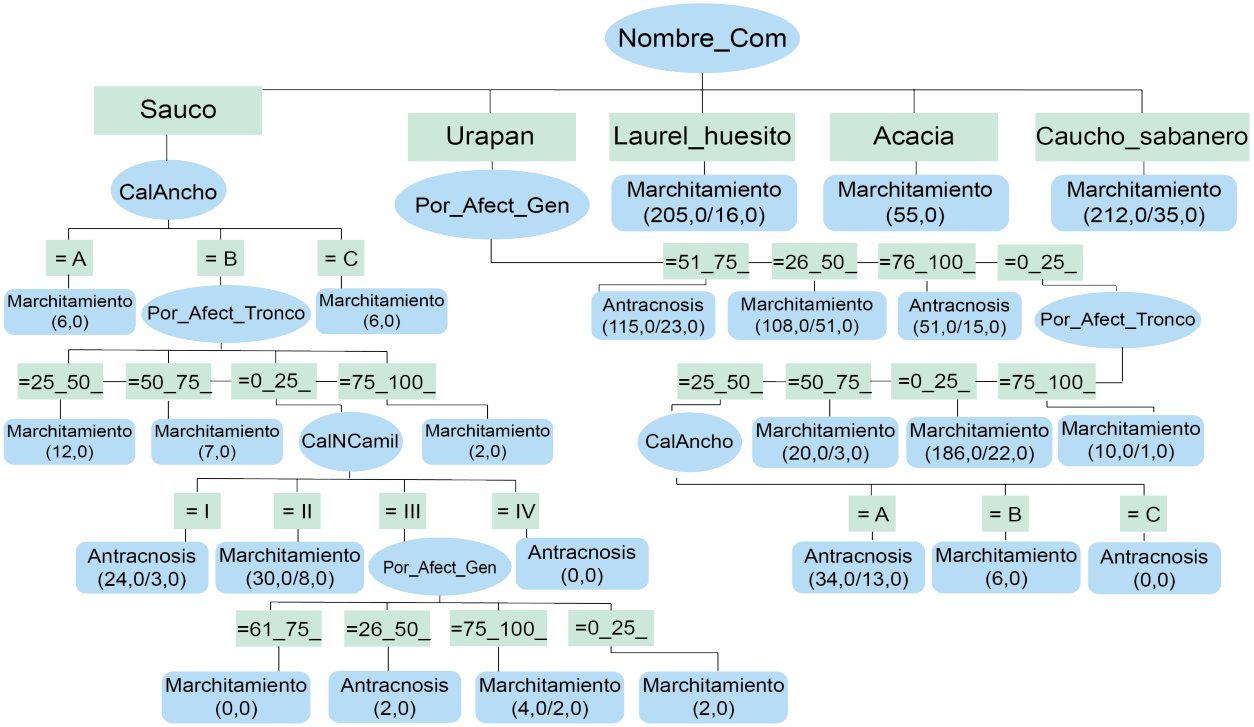

Figura 11. Árbol de decisiones para el Dataset infraestructura vial

En la Figura 11 se observan los resultados obtenidos, al igual que en los dos análisis anteriores. La especie es la primera variable relevante en patrones de enfermedades, pero, en este caso, las especies Laurel Huesito, Acacia y Caucho sabanero no establecen patrones de relación con la variable vías, mientras que la especie Sauco se ve afectada principalmente por el ancho del carril, donde si es de tipo A (< 5m) y C (10.1–15 m) se relaciona con la enfermedad marchitamiento, y si es de tipo B (5.1–10 m) se relaciona con el porcentaje de afectación al tronco. Si la afectación es mayor a 25% se relaciona con marchitamiento.

Si el porcentaje de afectación al tronco es menor a 25% se ve afectado por el número de carriles, donde se establecen de tipo I (1 carril) y V (5 carriles). Esto se relaciona con antracnosis, mientras que si es de tipo II (2 carriles) se relacionan con marchitamiento. Para el tipo de carril III (3 carriles) influye el porcentaje de afectación general, donde si es de 26% a 50% se relaciona con antracnosis; si es cualquier otro valor de porcentaje se relaciona con marchitamiento.

Para la especie Urapán se encuentran relaciones similares a los anteriores árboles de decisión, donde el grado de afectación general es la variable más relevante. Si presenta un porcentaje mayor a 50% de afectación se relaciona con antracnosis y para uno de 26% a 50% con marchitamiento. Cuando el porcentaje es menor a 25% la variable afectación del tronco presenta relación. Si el porcentaje de afectación es mayor a 50% y menor a 25% se relaciona con marchitamiento, mientras que si el porcentaje de afectación está entre 25% y 50% se relaciona con la variable ancho de carril, donde A y C se relacionan con antracnosis y B con marchitamiento.

Análisis de matriz de confusión: los valores de rendimiento para el árbol de decisiones realizado con el Dataset infraestructura vial presentó 1644 instancias clasificadas correctamente y solamente 267 mal clasificadas, obteniendo un rendimiento de 86,.02% y un índice Kappa de 47.7%. En la categoría cáncer se estableció como indefinido en la clasificación.

Análisis clúster

Especie-Enfermedad-Número de Carriles (CalNCarril): el resultado del análisis clúster para número de carriles de la Figura 12 permitió establecer que las enfermedades están relacionadas con vías de entre 1 y 4 carriles, más específicamente con las vías de solo un carril (tipo I), en donde la mayoría de individuos enfermos para las 5 especies están relacionados. A medida que aumentan los carriles es menor la cantidad de individuos enfermos.

La enfermedad cáncer se relaciona con las vías de tipo I y II en la especie Caucho Sabanero.

Especie-Enfermedad-Ancho de carril (CalAncho): se observa en la Figura 13 la relación de las enfermedades y las especies con el ancho del carril, estableciendo que no hay relación con vías mayores a 15 m de ancho. Para vías tipo A (menores a 5 m) es donde más se concentran los individuos enfermos para todas las especies evaluadas.

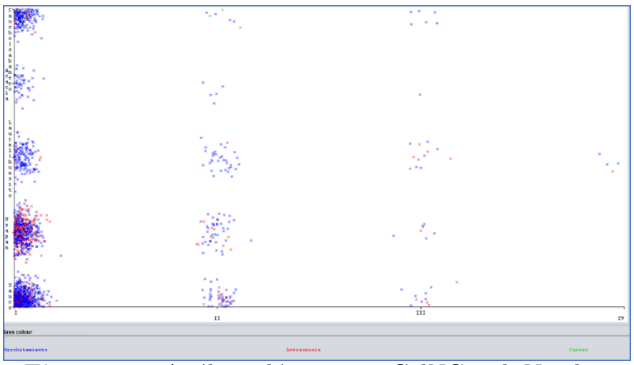

Figura 12. Análisis clúster para CalNCarril, Nombre común y Enfermedad. Salida programa WEKA

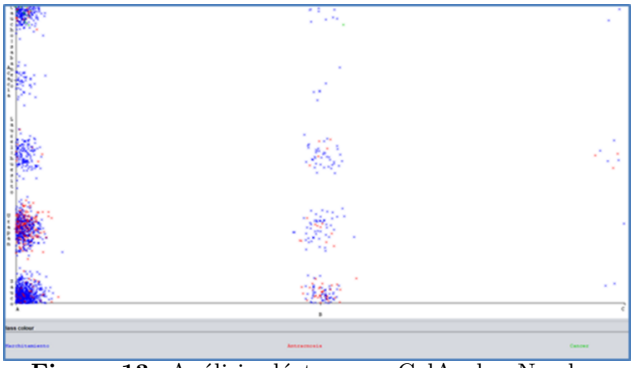

Figura 13. Análisis clúster para CalAncho, Nombre común y Enfermedad. Salida programa WEKA

La especie Sauco se relaciona principalmente a las vías tipo A con la enfermedad marchitamiento, mientras que para las vías tipo B se relaciona con la enfermedad antracnosis.

4) Contaminación atmosférica

En el clúster para el grupo de contaminación atmosférica en general se encontró que los materiales contaminantes de promedio anual no presentan relación con las enfermedades de los árboles. La única variable relevante fue la localidad para la especie Urapán, esto se debió a otras variables exógenas que no se abordaron en esta.

En la Figura 14 se observa que para la especie Urapán las localidades de Puente Aranda, Bosa, Tunjuelito y Kennedy se relacionan con marchitamiento, en Usaquén con antracnosis y para Fontibón se establece una relación con el porcentaje de afectación general, donde si es menor a 50% se relaciona con marchitamiento y si es mayor a 50% con antracnosis.

Análisis de matriz de confusión: los valores de rendimiento para el árbol de decisiones realizado con el Dataset contaminación ambiental presentaron 1661 instancias clasificadas correctamente y solamente 250 mal clasificadas, obteniendo un rendimiento de 86.91%, y un índice Kappa de 48.76%. La categoría cáncer se estableció como indefinida en la clasificación.

Análisis clúster

Especie-Enfermedad-Localidad: mediante el análisis de la Figura 15 para la especie Sauco se encontró que Tunjuelito no presentó ningún individuo enfermo y para Kennedy se presentó solamente un individuo enfermo, mientras que para la localidad de Fontibón fue más alto el número de individuos de Sauco enfermos.

Para la especie Urapán se encontró que en Usaquén está la mayor cantidad de los árboles con antracnosis. Para las especies Caucho sabanero y Acacia en relación con las enfermades es indiferente la localidad en donde se encuentren. A nivel general, las localidades que menos individuos enfermos presentaron fueron Kennedy y Tunjuelito.

V. Conclusiones

Los datos abiertos que se usaron para la elaboración de este artículo fueron fundamentales. La forma de procesarlos y prepararlos para el análisis se logró mediante la aplicación de herramientas geomáticas, con el fin de establecer relaciones geoespaciales para todas las variables trabajadas en este estudio. De esta forma se establece el gran potencial que tienen estas herramientas para análisis similares.

De los resultados encontrados se tiene que la especie Urapán es la que presenta una relación más relevante en todos los Dataset analizados, siendo la especie que presenta mayor grado de afectación. Esto servirá para la toma de decisiones respecto al manejo de esta especie en el arbolado urbano de Bogotá.

A nivel general todos los árboles de decisión presentaron como primera variable relevante a las especies, donde a excepción del Urapán y el Sauco, estas se relacionaban directamente con el marchitamiento. Para el árbol de decisión del Censo 2007 se obtuvo un índice global de 86.62%, que se considera una clasificación favorable. Para el árbol de decisión de sitios de interés se resalta la relación del Urapán con el porcentaje de afectación general y sitios de interés. Entre 0% y 25% se relacionó con los sitios VA17 (Bibliotecas), VIA21 (Edificios Civiles) y IA2 (Almacenes de cadena) y con la enfermedad de marchitamiento, mientras que para los demás sitios se relaciona con antracnosis si la afectación general es mayor a 50%. Este Dataset tuvo el rendimiento de 85.19%.

El árbol de decisión de infraestructura vial tuvo un índice de rendimiento global de 86.02%, donde solamente las especies Sauco y Urapán relacionan enfermedades con ancho de carril y número de carriles. Por último, para contaminación ambiental se obtuvo un valor de índice global de 86.91%, siendo el más alto de todos. Sin embargo, solamente para la especie Urapán se encontró relación con las localidades.

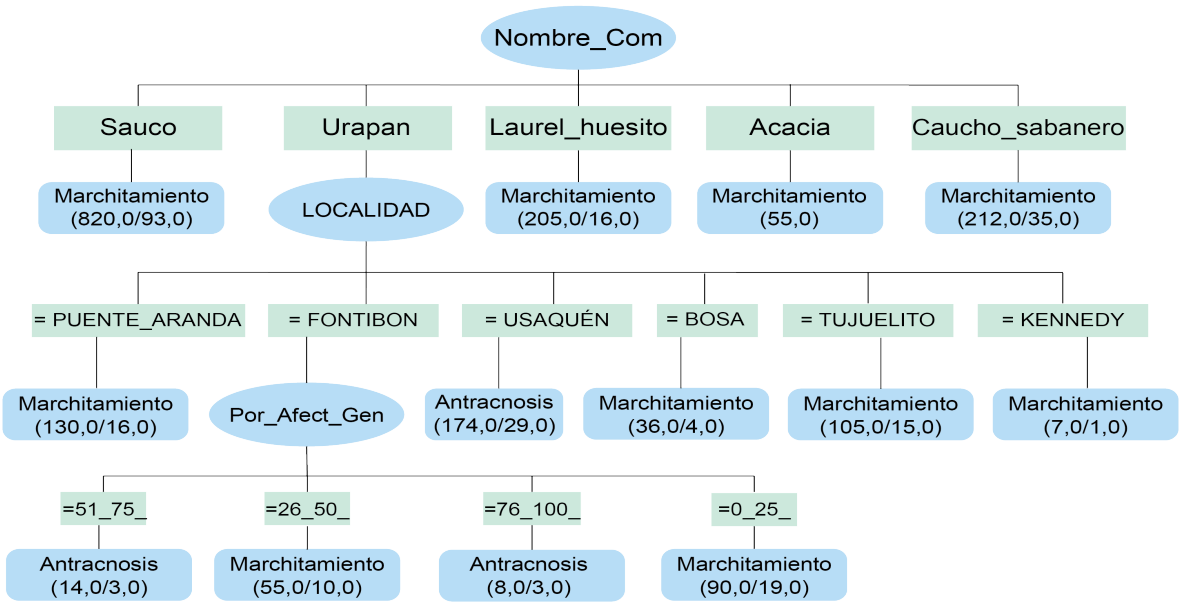

Figura 14. Árbol de decisiones para el Dataset Contaminación Ambiental

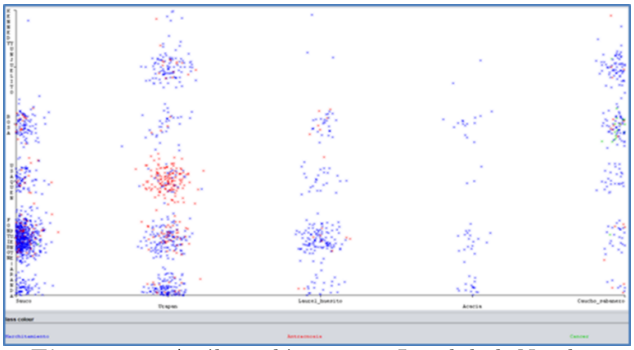

Figura 15. Análisis clúster para Localidad, Nombre común y Enfermedad. Salida programa WEKA

En cuanto a los análisis clúster se encontró para el análisis de Censo 2007 que en el porcentaje de afectación general entre 76% y 100% se encuentran la menor cantidad de individuos enfermos, mientras que la mayor cantidad se encuentra en un porcentaje menor a 25%. Para interferencia en redes solamente se encontró afectación con el Urapán.

En la variable altura discretizada se encontró que para la categoría de árbol pequeño y mediano hubo mayor cantidad de individuos enfermos. En el análisis de clúster para el Dataset de sitios de interés se encontró que los sitios que menor cantidad de individuos enfermos presentaron son templos, centros comerciales y complejos deportivos, mientras que los sitios que relacionan la mayor cantidad de individuos enfermos son instituciones educativas, parques y sectores residenciales.

En cuanto al Dataset de vías se identificó que para vías con un ancho mayor a 15 metros y más de 4 carriles no se presentó ninguna relación con las enfermedades en los árboles. De igual forma se obtuvo que entre más ancha sea la vía y entre más carriles se tengan, menor cantidad de individuos enfermos. Por último, para el análisis clúster de contaminación se identificó que las localidades de Kennedy y Tunjuelito son las que menos individuos enfermos tienen. En Urapán se presentaron diferencia entre tipos de enfermedad, mientras que la mayor cantidad de individuos enfermos de Sauco se encontró en Fontibón.

Las especies Caucho Sabanero, Laurel Huesito y Acacia se relacionan directamente con la enfermedad de marchitamiento en todos los grupos de análisis. Para la especie Sauco se encontró una relación en el Dataset de infraestructura vial, donde se relacionan las enfermedades con respecto al ancho del carril y al número de carriles.

Para la especie Urapán se encontró relación con algunas variables de todos los Dataset. Por ejemplo, para el Censo 2007 se encontró una relación entre porcentaje de afectación general. Si esta presentaba porcentajes de afectación mayores a 50% se relacionaba con la enfermedad antracnosis, pero si presentaba porcentajes entre 26% y 50% mostraba marchitamiento.

Para los valores de afectación general menor a 25% se encuentra relación con la variable interferencia en redes, donde si esta aparece se relaciona con la variable de categoría de altura. De acuerdo con esto, un Urapán de tipo arbolito y árbol mediano con interferencia en redes y un porcentaje de afectación menor a 25% presenta marchitamiento; si es de tipo árbol pequeño y árbol grande

presenta antracnosis. Para el Dataset de interferencia en redes se tiene que: si presentaba un porcentaje mayor a 50% de afectación se relacionaba con antracnosis y si era de 26% a 50% con marchitamiento. Cuando el porcentaje era menor a 25% la variable afectación del tronco presentaba relación; si el porcentaje de afectación era mayor a 50% y menor a 25% se relacionaba con marchitamiento, mientras que si el porcentaje de afectación estaba entre 25% y 50% y se relacionaba con la variable ancho de carril, para A y C se relacionaba con antracnosis y para B con marchitamiento.

En cuanto al Dataset de sitios de interés se encontró que hay relación con el porcentaje de afectación general, donde si presentaba un porcentaje de 26% a 50% se relacionaba con marchitamiento, mientras que si el porcentaje se encontraba entre 0% y 25% se relacionaba con los sitios VA17 (Bibliotecas), VIA21 (Edificios Civiles) y IA2 (Almacenes de cadena) y la enfermedad de marchitamiento. Para los demás sitios se relacionaba con antracnosis si la afectación general era mayor a 50%.

Este artículo servirá para la toma de decisiones respecto al manejo del arbolado urbano en Bogotá, debido a que relaciona los sitios en los cuales se presentan mayores afectaciones para las especies.

En futuras investigaciones se pueden realizar estos tipos de análisis para determinar cuáles son las especies que presentan menos afectación y así establecer especies aptas para el arbolado urbano de Bogotá.

VI. Agradecimientos

Los autores agradecen a la Universidad Distrital Francisco José de Caldas y al semillero de investigación Semillero Producción y Manejo Forestal (PROMAFOR).

Referencias

[1] Secretaria Distrital De Planeación (SDP), "Análisis Demográfico y Proyecciones Poblacionales Bogotá, " Bogota : Secretaria Distrital de Planeación, 2018.

[2] Secretaria Distrital de Ambiente (SDA). "Informe Anual de Calidad del Aire en Bogota 2018," Bogota : Secretaria Distrital de Ambiente, 2019.

[3] A. Zamudio, "Estrategias Para Mitigar La Contaminación del Aire en Zonas Aledañas a Grandes Avenidas de Bogotá," Msc, Universidad Nacional de Colombia-Sede Bogotá, 2017.

[4] F. A. Seoane and J. M. Evans, "Beneficios del arbolado urbano evaluación del balance entre secuestro, demanda energética y otros impactos", Av. En Energ. Renov. Medio Ambiente, vol. 5, 2001.

[5] R. H. Waring, "Characteristics Of Trees Predisposed To Die: Stress Causes Distinctive Changes In Photosynthate Allocation," *Bioscience*, vol. 8, no. 37, pp. 569–574, 1987.

[6] IBM (2015, jun.), "Metodología Fundamental Para La Ciencia De Datos," [Online]. Available: https://www.ibm.com/downloads/cas/6RZMKDN8.

[7] J. E. R. Rodirguez, "Fundamentos de mineria de datos," Bogotá: Universidad Distrital Francisco José de Caldas, 2010.

[8] Corporación Autónoma Regional de Cundinamarca-CAR and A. Guzman Gonzalez, "Zonas De Vida o Formaciones Vegetales Area Jurisdiccional C.A.R," Bogotá, D.C., Colombia: CAR, pp. 5–12, 1996.

[9] T. Aluja, "La minería de datos, entre la estadística y la inteligencia artificial," *Qüestiió: quaderns d'estadística i investigació operativa*, vol. 3, no. 25, 2001.

[10] D. Pyle, "Data Preparation for Data Mining," San Fransisco: Morgan Kaufmann Publishers, Inc. 1999.

[11] J. A. Gallardo Arancibia, "Metodología para la definición de requisitos en proyectos de data mining," phd, Facultad de Informática (UPM), 2009.

[12] J. Gironés Roig, "Minería de datos: modelos y algoritmos," Barcelona : Editorial UOC, pp. 1–273, 2017.

[13] C. G. Cambronero and I. G. Moreno, "ALGORITMOS DE APRENDIZAJE: KNN & KMEANS," *Inteligencia de Redes de la Telecomunicación*, vol. 23, p. 8, 2006.

[14] O. H. I. Restrepo, H. Moreno, and E. Hoyos, "Incidencia del Deterioro Progresivo del Arbolado Urbano en el Valle de Aburrá, Colombia," *Colombia Forestal*, vol. 18, no. 2, pp. 225–240, 2014.

[15] E. Pardo and D. Cruz, "Introducción al Análisis de Datos Textuales," Presentado en el XXII Simposio Internacional de Estadistica, Bucaramanga, Colombia, Jul. 2012.

[16] Dirección De Regulación, Planeación, Estandarización y Normalización Estadística (Dirpen), "COLOMBIA-Censo del Arbolado Urbano Bogotá D.C, - CAU 2005 - 2007, FASES I-IV SEPTIEMBRE 2005 SEPTIEMBRE 2007," Bogotá: Dirección De Regulación, Planeación, Estandarización y Normalización Estadística, 2014.

[17] L. J. Moscovitz, "Un Modelo Conceptual para el Desarrollo de Árboles de Decisión con Programación Genética," Esp., Fundación Universitaria Konrad Lorenz, Bogotá, D.C., Colombia, 2007.

[18] S. R. Timarán-Pereira, I. Hernández-Arteaga, S. J. Caicedo-Zambrano, VA. Hidalgo-Troya and J. C. Alvarado Pérez, "El Proceso De Descubrimiento De Conocimiento En Bases De Datos," In *Las Competencias Genéricas De La Formación Profesional.* Bogotá: Ediciones Universidad Cooperativa De Colombia, 2016, pp.63-86.

Anexos

Tabla 2. Variables, siglas y descripción dataset censo 2007.

Variable	Sigla	Descripción
Nombre común	Nombre_Com	Nombre común del árbol
Tipo de árbol	Tipo_Arbol	Clasificación fisiológica del árbol
Estado fitosanitario	Estado_Fit	Estado fitosanitario del árbol
Porcentaje afectación general	Por_Afect_General	Porcentaje afectación de todo el árbol
Enfermedad	Enfe	Tipo de enfermedad identificada
Porcentaje afectación tronco	Por_Afec_Tronco	Porcentaje afectación del tronco
Porcentaje afectación raíz	Por_Afec_Raiz	Porcentaje afectación de la raíz
Raíces expuestas	Raices_Exp	Presencia o ausencia de raíces expuestas
Altura discretizada	Alt_Disc	Categorías de altura discretizadas
Categoría diamétrica	Categoria_Diam	Categorías de DAP discretizadas
Categoría de ancho de copa	Categoria_Copa	Categorías de ancho de copa discretizadas
Tipo de afectación al suelo	Diagnostic	Categorías de afectación al suelo
Interferencia en redes	Interferen	Presencia o ausencia de redes aéreas respecto al árbol

Tabla 3. Variables, siglas y descripción dataset contaminación ambiental.

Variable	Sigla	Descripción
Localidad	Localidad	División territorial y administrativa de Bogotá
Materia particulada 10	PM10	Pequeñas partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento o polen, dispersas en la atmósfera y cuyo diámetro varía entre 2.5 y 10 μm
Monóxido de carbono	CO	Gas altamente tóxico. Se produce cuando se queman materiales combustibles como gas, gasolina, keroseno, carbón, petróleo, tabaco o madera en ambientes de poco oxígeno
Ozono	OZONO	Gas altamente reactivo. Se considera como uno de los contaminantes de mayor preocupación. Es altamente oxidante y afecta a los tejidos vivos
Óxidos de nitrógeno	NO NO2 NOX	Uno de los principales contaminantes entre los varios óxidos de nitrógeno, subproducto en los procesos de combustión a altas temperaturas como en vehículos motorizados y plantas eléctricas
Óxido de azufre	SO2	Gas que en contacto con el aire y la humedad se convierte en trióxido de azufre. En agua se disuelve formando una disolución ácida.
Temperatura	Temperatura	Temperatura Categorizada

Tabla 4. Variables, siglas y descripción dataset infraestructura vial.

Variable	Sigla	Descripción
Número de carriles	CalNCarril	Categorización del número de carriles de cada vía
Ancho del carril	CalAncho	Categorización del número del ancho del carril de cada vía

Tabla 5. Variables, siglas y descripción dataset desitios de interés.

Variable	Sigla	Descripción
Sitios de interés	CodSIFin	Códigos de actividades de los sitios de acuerdo al catastro

Tabla 6. Categorías y siglas para dataset censo 2007.

Altura		DAP		Diámetro copa	
Valores (m)	Categoría	Valores (cm)	Categoría	Valores (m)	Categoría
≤ 1	Arbusto	0–10	I	0–1	I
1.01–2	Arbolito	10.01–20	II	1–2	II
2.01–4	Árbol pequeño	20.01–30	III	2–3	III
4.01–10	Árbol Mediano	30.01–40	IV	3–4	IV
> 10	Árbol Grande	40.01–50	V	4–5	V
		50.01–60	VI	5–6	VI
		60.01–70	VII	6–7	VII
		70.01–80	VIII	7–8	VIII
		80.01–90	IX	8–9	IX
		90.01–100	X	9–10	X
		100.01–110	XI	>10	XI
		>110	XII		

Tabla 7. Categorías y siglas para dataset infraestructura vial.

Ancho de carril		**Cantidad de carriles**	
Valores (m)	Categoría	Valores (unidad)	Categoría
1–5	A	1	I
5.1–10	B	2	II
10.1–15	C	3	III
15.1–20	D	4	IV
20.1–25	E	5	V
25.1–30	F	6	VI
>30	G	7	VII
		8	VIII
		9	IX