DETECCIÓN DE LA "ROYA" EN PLANTAS DE CAFÉ.

Camilo Cossio Alzate
Universidad EAFIT
Colombia
ccossioa@eafit.edu.co

Juan Andrés Giraldo Aristizábal
Universidad EAFIT
Colombia
jagiraldoa@eafit.edu.co

Mauricio Toro
Universidad EAFIT
Colombia
mtorobe@eafit.edu.co

RESUMEN
El objetivo de este proyecto es analizar y contribuir a la prevención de problemas de plagas en los cultivos que afectan indirectamente la economía nacional. Nuestro país es uno de los principales exportadores de café en el mundo, la economía colombiana se sustenta en gran parte gracias a los ingresos producidos por dicha actividad, sin embargo (BBC, 2013), anualmente se dan grandes perdidas a causa de la Roya, la principal plaga fitosanitaria que afecta el café.

La importancia del problema se basa específicamente en lo anteriormente mencionado, puesto que, si se logra descubrir a tiempo la presencia de esta plaga en los cultivos, es posible controlarla y así aumentar exponencialmente la cantidad de café exportado, y por consiguiente los ingresos.

A esta problemática se suman muchas otras, por ejemplo, la calidad del café disminuye, y las familias que dependen de los ingresos que este genera se ven realmente afectadas por la disminución de la producción y comercialización, adicionalmente, si esta problemática continua creciendo, los campesinos dejarán de cultivar el café porque no será rentable para ellos, debido a que podría tener como resultado más perdidas que ganancias, y deberán recurrir a otras actividades que resulten más favorables.

Palabras clave
Estructuras de datos; roya; aplicaciones de tabla hash; notación O; complejidad.

Palabras clave de la clasificación de la ACM

CCS → Theory of computation → Models of computation → Probabilistic computation

1. INTRODUCCIÓN

La humedad y el calor son las condiciones ambientales que juntas hacen que la Roya pueda crecer y reproducirse en cierto sembrado de café, la temperatura perfecta para la Roya oscila entre 18 y 24 °C, evidentemente los factores climáticos son los que más impulsan su expansión. Al principio requiere la salpicadura de la lluvia para iniciar su proceso de dispersión por toda la zona que será afectada, más tarde necesita la presencia de una capa de agua por el envés de las hojas para que el hongo germine, finalmente necesita poca presencia de luz solar y la temperatura adecuada. Para el adecuado desarrollo del hongo en el sembrado de café, se necesitan precipitaciones constantes especialmente en las horas de la tarde o noche, con cielos nublados que impidan que la temperatura sea muy alta después del medio día o muy bajas en la madrugada. Tras todo lo anteriormente explicado, el proyecto se basará en el procesamiento y análisis de datos relacionados con las condiciones anteriormente descritas, que permiten que el hongo en cuestión crezca fácil y rápidamente, para así lograr predecir si es probable que exista la Roya o no. La Roya llegó a Colombia en 1970, y tuvo un porcentaje de infestación de alrededor de 40% por una ola invernal que hizo que la humedad aumentara significativamente, favoreciendo el crecimiento del hongo, en este momento el gobierno colombiano debió invertir alrededor de 500 millones de dólares, destinados principalmente a las fincas pequeñas productoras de café.

2. PROBLEMA

El problema que intentamos resolver es el de la detección de la Roya mediante datos relacionados con las condiciones adecuadas para que la Roya crezca y se expanda por cierto cultivo, intentamos resolver este problema para favorecer la erradicación o control de esta plaga, facilitando el proceso de detección de su existencia en determinado lugar para proceder adecuadamente, disminuyendo pérdidas y aumentando así la producción e ingresos de las personas dedicadas a esta actividad.

3. TRABAJOS RELACIONADOS

3.1 Algoritmo Chaid

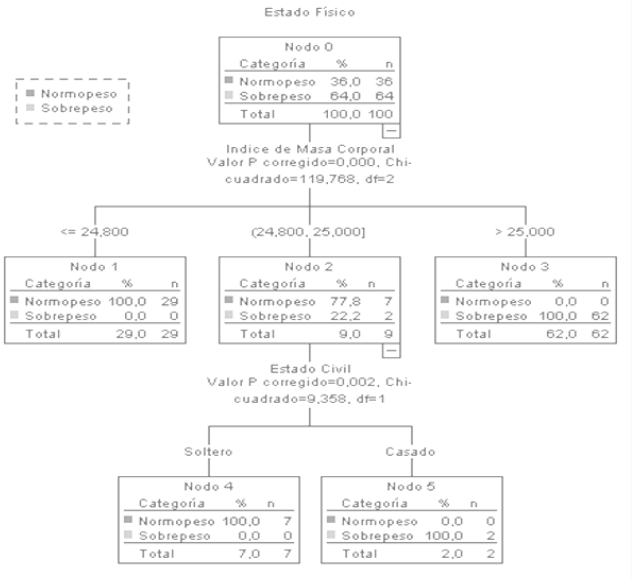


Imagen 1.1

- Procede del ámbito de la Inteligencia artificial. Desarrollado por Kass a principios de los años 80
- Asume que las variables explicativas son categóricas u ordinales. Cuando no lo son, se discretizan
- Inicialmente se diseñó para el caso de variable respuesta Y categórica. Posteriormente se extendió a variables continuas
- Utiliza contrastes de la χ 2 de Pearson y la F de Snedecor
- El corte en cada nodo es multi-vía.

Ver imagen 1.1.

3.2 Algoritmo Id3

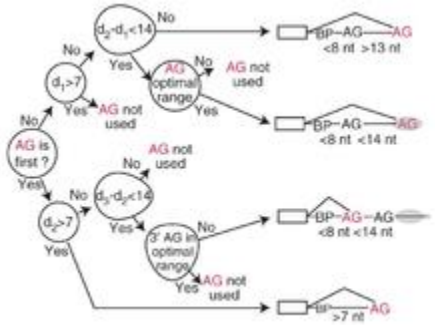


Imagen 2.1

El algoritmo ID3 es utilizado dentro del ámbito de la inteligencia artificial. Su uso se engloba en la búsqueda de hipótesis o reglas en él, dado un conjunto de ejemplos.

El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos, (el atributo a clasificar) es el objetivo, el cual es de tipo binario (positivo o negativo, sí o no, válido o inválido, etc.).

De esta forma el algoritmo trata de obtener las hipótesis que clasifiquen ante nuevas instancias, si dicho ejemplo va a ser positivo o negativo.

ID3 realiza esta labor mediante la construcción de un árbol de decisión.

Los elementos son:

- Nodos: Los cuales contendrán atributos.
- Arcos: Los cuales contienen valores posibles del nodo padre.
- Hojas: Nodos que clasifican el ejemplo como positivo o negativo.

Ver imagen 2.1.

3.3 Algoritmo C4,5 para atributos continuos y discretos.

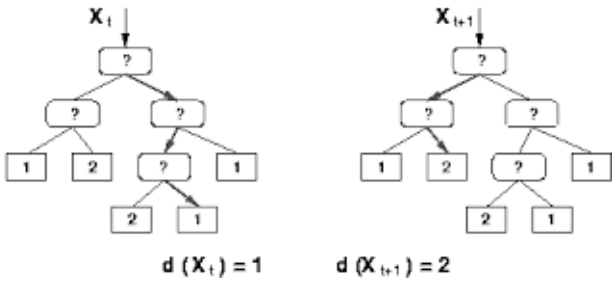


Imagen 3.1

J.R Quinlan propone una mejora del algoritmo ID3 que genera un árbol de decisión a partir de los datos mediante participaciones realizadas recursivamente. El algoritmo considera todas las posibles pruebas que puedan dividir el conjunto de datos y selecciona la prueba que le haya generado mayor ganancia de información. Las características de este algoritmo son:

- Permite trabajar con valores continuos para los atributos, separando los posibles resultados en dos ramas.
- Los árboles son menos frondosos, debido a que cada hoja cubre una distribución de clases, no una clase particular.
- Utiliza el método “divide y vencerás” para generar el árbol de decisión inicial a partir de un conjunto de datos de entrenamiento.
- Se basa en la utilización del criterio de proporción de ganancia.
- Es recursivo.

Ver imagen 3.1.

3.4 Algoritmo C5.

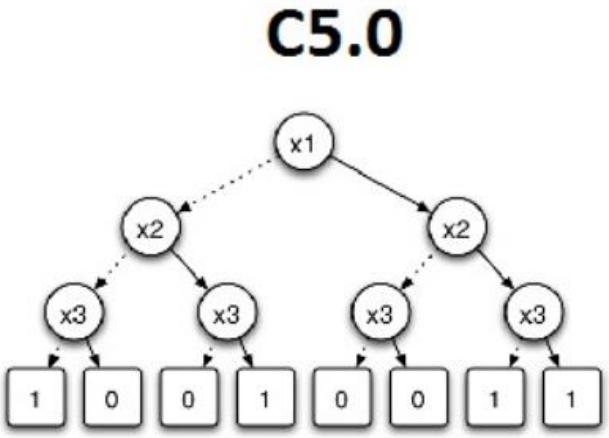


Imagen 4.1

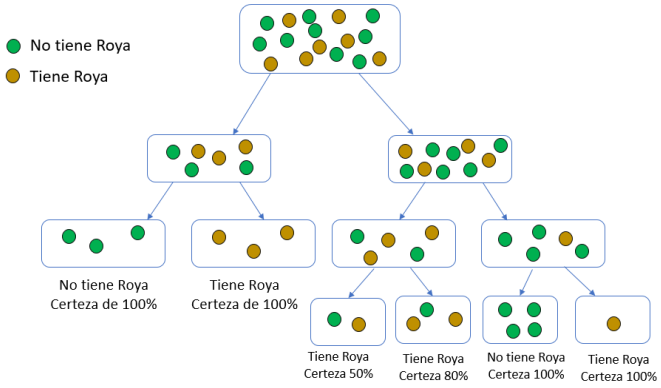
C5.0 es el algoritmo sucesor de C4.5, ambos publicados por Quinlan, con el objetivo de crear árboles de clasificación. Entre sus características, destacan la capacidad para generar árboles de predicción simples, modelos basados en

reglas, *ensembles* basados en *boosting* y asignación de distintos pesos a los errores. Las características de este algoritmo son:

- La medida de pureza empleada para las divisiones del árbol es la entropía.
- El podado de los árboles se realiza por defecto, y el método empleado se conoce como *pessimistic pruning*.
- Los árboles se pueden convertir en modelos basados en reglas.
- Emplea un algoritmo de boosting más próximo a *AdaBoost* que a *Gradient Boosting*.
- Por defecto, el algoritmo de *boosting* se detiene si la incorporación de nuevos modelos no aporta un mínimo de mejora.
- Incorpora una estrategia para la selección de predictores (*Winnowing*) previo ajuste del modelo.
- Permite asignar diferente peso a cada tipo de error.

Ver imagen 4.1.

4. Estructura de datos

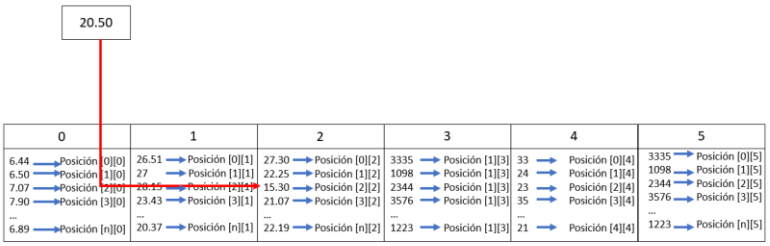


Gráfica1: Árbol de decisión que evalúa la presencia de roya de acuerdo con las variables pH, humedad del suelo, temperatura del suelo, humedad del ambiente, iluminación y temperatura del ambiente.

Gráfica 2: Arreglo bidimensional que almacena cada columna y fila del archivo CSV en forma de matriz.

4.1 Operaciones de la estructura de datos

Insertar nuevo dato.

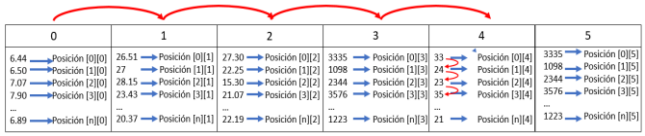


Gráfica 3: Insertar el elemento 20.50 en la posición [2][2], en este caso como existe un elemento en dicha posición, se reemplaza, pero en caso de no haber elemento, solo se inserta.

Eliminar.

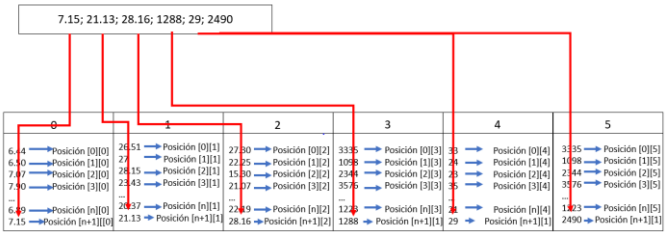
Gráfica 4: Eliminar los elementos en las posiciones [2][0], [1][3], [3][5].

Buscar.



Gráfica 5: Buscar el elemento “35” en la categoría temperatura del ambiente.

Insertar nueva fila.



Gráfica 6: Luego de leer el archivo CSV, se le asigna a la nueva fila su respectiva posición dentro del arreglo bidimensional a cada dato que contiene esta.

4.2 Criterios de diseño de la estructura de datos

Elegimos esta estructura de datos porque se acomoda perfectamente a los requerimientos del proyecto, pues al organizar los datos en forma de tabla de la misma forma en

que son ingresados se hace más fácil insertar y acceder a estos para realizar el respectivo análisis y determinar si la roya existe o no en esta planta. La complejidad de operaciones es otro factor que se tuvo en cuenta, porque acceder y eliminar posiciones de memoria específicas es O(1), otras operaciones como insertar una fila o buscar un dato específico tiene complejidad O(n), con n = Número de filas (Porque el número de columnas es constante), no obstante, esto no será un gran obstáculo porque las entradas no tienen una cantidad de datos exageradamente grande, lo que hace que el algoritmo no se tarde mucho en realizar cualquier operación Adicionalmente, la estructura no ocupa una gran cantidad de memoria RAM en el computador en que se ejecute, porque se trata de un arreglo bidimensional, el cual va creando posiciones a medida que se ocupan las que estaban disponibles.

4.3 Análisis de complejidad.

Operación	Complejidad
Buscar	O(n)
Insertar nuevo dato	O(1)
Eliminar	O(1)
Insertar nueva fila	O(n)

Tabla 1: Complejidad de las operaciones de la estructura de datos

4.4 Tiempos de ejecución.

Tamaño	Mejor Tiempo [seg]	Peor Tiempo [seg]	Tiempo Promedio [
457	0.000297785	0.000964403	0.000411267
373	0.000287056	0.002040386	0.00040689
673	0.000524521	0.001213312	0.00058816
301	0.000231504	0.001347065	0.000339568

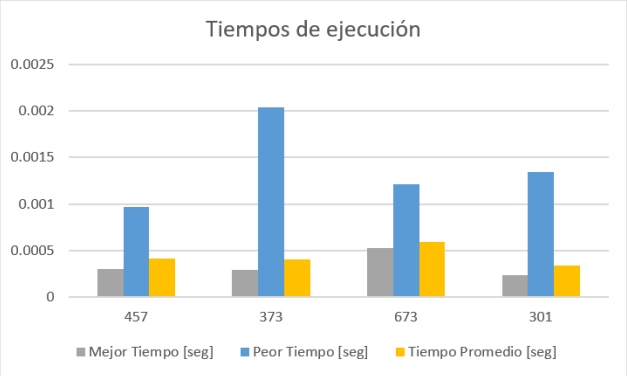
Tabla 2: Tiempos de ejecución con conjuntos de datos de diferentes tamaños.

4.5 Memoria.

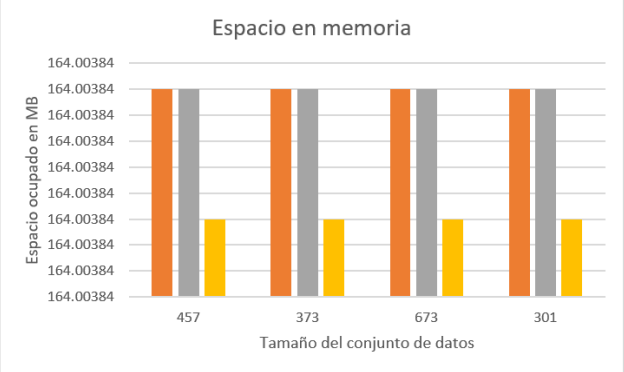
Tamaño	Mejor Memoria [MB]	Peor Memoria [MB]	Memoria Promedio [M
457	164.00384	164.00384	164.00384
373	164.00384	164.00384	164.00384
673	164.00384	164.00384	164.00384
301	164.00384	164.00384	164.00384

Tabla 3: Consumo de memoria para conjuntos de datos de diferentes tamaños

4.6 Análisis de los resultados.



Gráfica 6: Tiempos de ejecución de conjuntos de datos de varios tamaños.



Grafica 7: Espacio en memoria de conjuntos de datos de varios tamaños.

De los gráficos anteriores podemos concluir que el tiempo de ejecución es muy mínimo y este crece o decrece de acuerdo al tamaño de la base de datos que se le ingrese. La cantidad de memoria que se consume es baja, lo que favorece la ejecución del algoritmo.

REFERENCIAS

Referencias

Arevalillo, J. M. (s.f.). *Data Mining con Árboles*. Obtenido de Data Mining con Árboles: https://web.fdi.ucm.es/posgrado/conferencias/JorgeMartin-slides.pdf

BBC. (24 de Enero de 2013). *BBC news*. Obtenido de BBC news: https://www.bbc.com/mundo/noticias/2013/01/130123_despiadado_enemigo_cafe_centroamerica

D. (s.f.).

Federacion Nacional de Cafeteros de Colombia. (s.f.). *Federacion Nacional de Cafeteros de Colombia*. Obtenido de Federacion Nacional de Cafeteros de Colombia: https://www.federaciondecafeteros.org/particulares

/es/programas_para/11621_manejo_adecuado_de_la_roya_del_cafeto/

Federación Nacional de Cafeteros de Colombia. (s.f.). *Federación Nacional de Cafeteros de Colombia*. Obtenido de Federación Nacional de Cafeteros de Colombia: https://www.federaciondecafeteros.org/particulares/es/programas_para/11621_manejo_adecuado_de_la_roya_del_cafeto/116211_practicas_y_recomendaciones_para_el_manejo_de_la_roya-1/

Rodrigo, J. A. (02 de 2017). *Ciencia de Datos.* Obtenido de Ciencia de Datos: https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting#introducci%C3%B3n

Scribd. (s.f.). *Scribd*. Obtenido de Scribd: https://es.scribd.com/doc/57484779/Algoritmo-c45-Arboles-de-Decision

Wikipedia. (s.f.). *Wikipedia La encicloedia Libre*. Obtenido de https://es.wikipedia.org/wiki/Algoritmo_ID3