# Detecciòn de la roya en el cafè caturra por medio de àrboles de decisiòn

Manuela Herrera Universidad EAFIT Colombia [mherreral@eafit.edu.co](mailto:mherreral@eafit.edu.co)

Samuel Palacios Universidad EAFIT Colombia [sdpalaciob@eafit.edu.co](mailto:sdpalaciob@eafit.edu.co)

Mauricio Toro Universidad Eafit Colombia [mtorobe@eafit.edu.co](mailto:mtorobe@eafit.edu.co)

**RESUMEN**

El problema a tratar en este proyecto es la tardía detección de la roya en los lotes de café, puntualmente el de tipo caturra, que es de mayor exportación. Si la roya se detecta a tiempo, se pueden prevenir significativas pérdidas de bultos de este, lo que en otras palabras significa impulsar la industria cafetera y elevar los ingresos económicos del país y de las familias caficultoras. Sabiendo que Colombia es un gran exportador, no sólo de café, sino también de flores, caña de azúcar y demás elementos derivados de la agricultura; si se sabe cómo interpretar las variables de los cultivos como el pH, la humedad, entre otros, entonces se podrá aumentar el índice de producción, lo que significa un beneficio tangible en la agricultura y la economía.

Así, por medio de esta solución, que implementa el árbol de decisión CART sobre las variables independientes y su variable dependiente que es; si tiene roya o no. Se logra entrenar el árbol y predecir sobre otro conjunto de variables x si estas plantas tienen o no tienen roya.

Finalzando, se obtienen resultados con un acierto del 85% en escala, siendo una buena aproximación para tomar acciones en contra del hongo, facilitando el proceso de detección y mejorando las condiciones, no sólo de los cultivos, sino de los agricultores y posibles beneficiarios de esto. Además, esta solución se puede extrapolar a varios ámbitos que tengan necesidades similares, logrando avances signicifativos de mano de la tecnología.

# PALABRAS CLAVE

Detección de la roya, algoritmos en cultivos, árboles de decisión, café caturra, filtración y análisis de datos, redes neuronales, entrenamiento de datos, modelos algorírmicos, estudios de la roya

**Palabras clave de la clasificación de la ACM**

Information systems > Database design and models > Design and analysis of algorithms > Data structures

# INTRODUCCIÓN

Colombia es un país que basa su economía en el sector primario; muchas de las familias colombianas dependen de cultivos o del ganado, por lo que, si una plaga como la roya se vuelve corpórea, afecta en gran medida la industria, generando perdidas en su mayoría, significativas.

Hasta el momento la detección de este tipo de plagas está muy mal optimizada, generando así un déficit en el desarrollo de la agricultura, lo que desemboca en no llegar a ser un país competitivo con referencia a la exportación de productos cultivados en el territorio. Por otra parte, los métodos utilizados para el control de dichos factores son poco amigables, el uso de químicos, venenos y demás elementos para el control de plagas hacen que la calidad de los cultivos disminuya, afectando no solo vendedor sino también al consumidor.

Haciendo uso de los árboles de decisión se busca prevenir la pérdida de un sustancial porcentaje de los cultivos, además de reemplazar métodos obsoletos para el control de plagas generando de este modo una alternativa que mejora en todo aspecto el factor comercial e industrial de los agricultores colombianos.

# PROBLEMA

En síntesis, la detección tardía de la roya en los cafetales, genera pérdidas inimaginables para la industria colombiana, la disminución en la producción y exportación sumado a las prácticas poco amigables con el cultivo y el consumidor como el uso de químicos y plaguicidas son factores que de ser tratados de forma oportuna pueden llegar a verse reflejados en un mejor posicionamiento del país en cuanto a exportación y aprovechamiento de recursos naturales, además de propiciar el uso de la tecnología como fundamento para el impulso de las familias agricultoras

# TRABAJOS RELACIONADOS

* 1. **Algoritmo ID3**

Este algoritmo se encarga de crear el árbol de decisión por medio de un nodo principal y unos terminales, y así, poder generar el árbol mínimo para un conjunto definido. De esta forma, por medio de preguntas se establecen para posteriormente crear un modelo “predictivo” para poder asociar los elementos del conjunto a una clase en particular.

# Algoritmo C5

El algoritmo se encarga de dar un máximo de información sobre un conjunto de reglas sobre las cuales se ejecuta el árbol de decisión, donde sencillamente se muestran las divisiones que hay en el código, de forma tal que por medio del todo que comprenden los nodos y los datos se logre predecir el comportamiento del algoritmo.

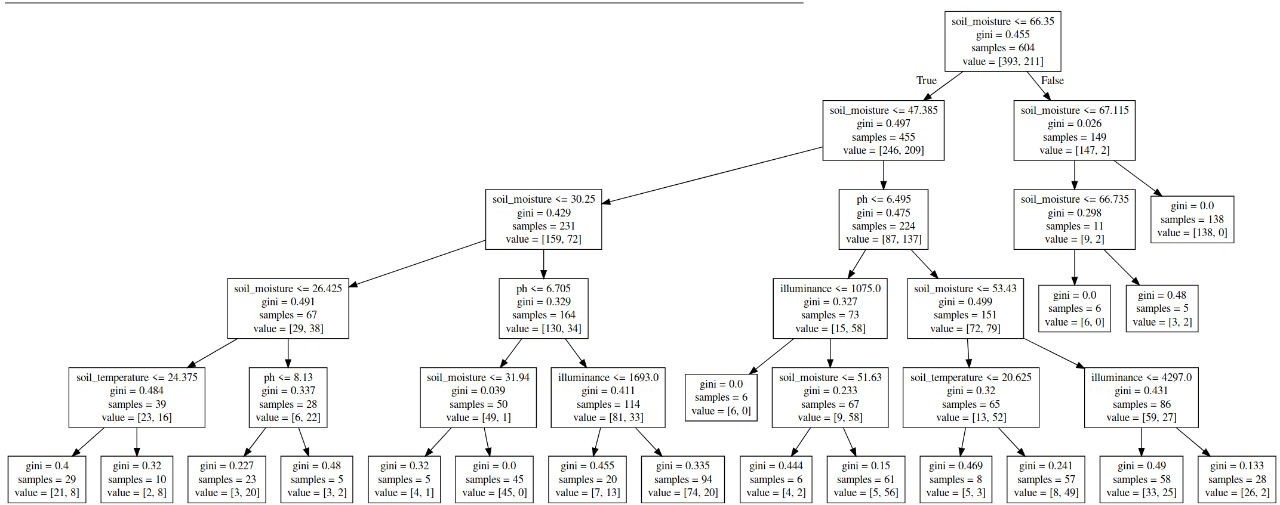
# Algoritmo CART

Conocido como classification and regression tree es un tipo de árbol estrictamente binario que funciona mediante las particiones recursivas. para su uso requiere un data set clasificado y un conjunto de entrenamiento. Sirve para la selección de características de datos tanto numéricos como categóricos utilizando datos históricos para predecir y/o clasificar nuevos datos, para definir que particiones hacer o que reglas utilizar, implementa un criterio de separación Phi “Φ” en base a promedio y sumatorias y también con el índice de Gini en base a las clases probando así todas las reglas de particionado y escogiendo las más optimas en cada caso.

# Algoritmo C4.5

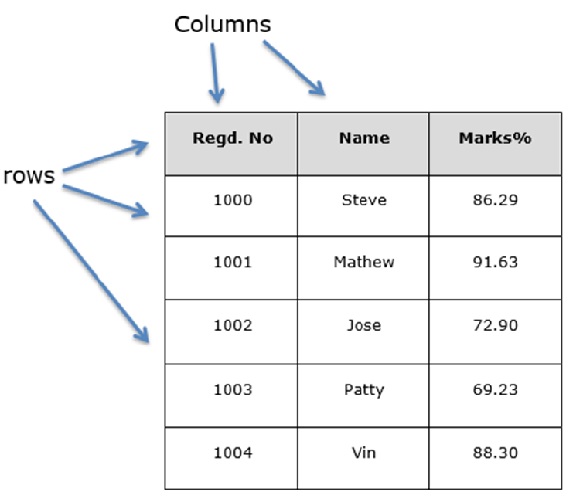
También llamado J48 es una ampliación de ID3. Es de tipo eneario lo que quiere decir que funciona haciendo particiones de forma recursiva, pero sin restricción, a diferencia del binario. Su criterio de división está fundamentado a partir de la ganancia de información, es decir, utilizando la entropía de Shannon y probando en un principio el data set sin reglas estableciendo una entropía inicial posteriormente prueba la entropía de todas las reglas de división y así obtiene el resultado óptimo.

# Algoritmo CART



**Gráfica 1:** Ärbol de clasificación CART, cada planta tiene las respectivas variables de humedad del ambiente, pH, temperatura del suelo, temperatura del ambiente

# Operaciones de la estructura de datos



**Gráfica 2:** Imagen de una operación de agregardo en una matriz

# Criterios de diseño de la estructura de datos

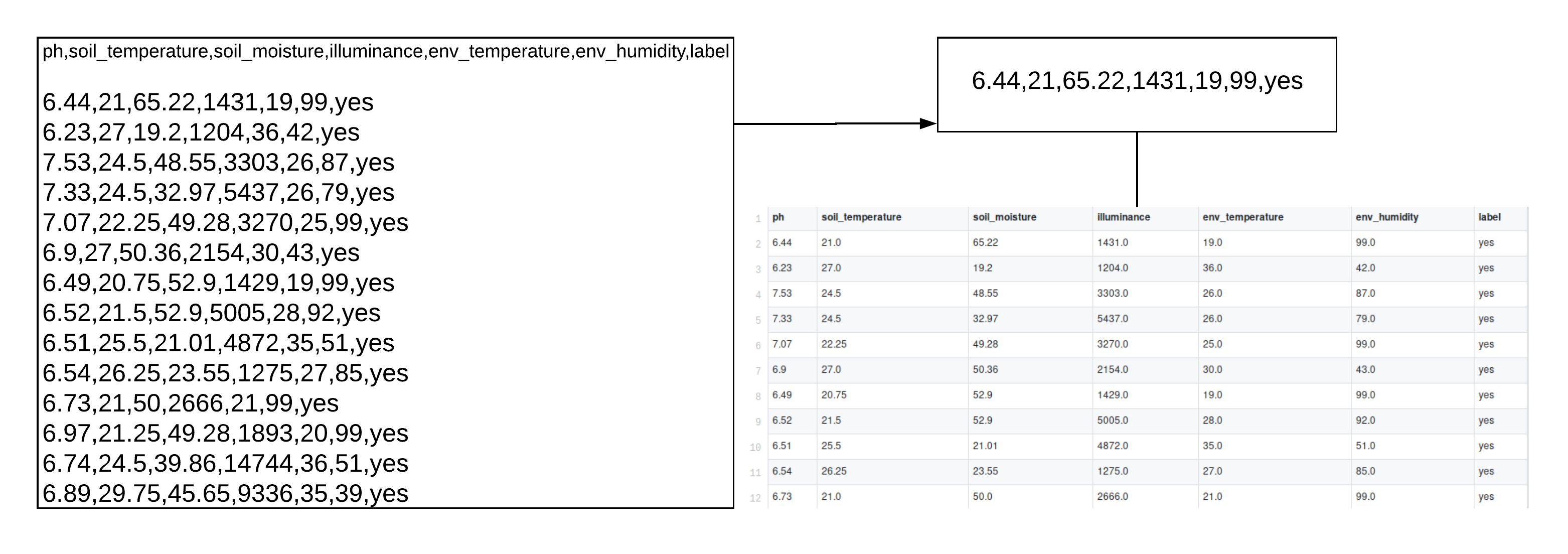
Ahora nos plantearemos el por qué de la elección del algoritmo CART; durante la búsqueda de algo que se acomodara a nuestras necesidades vimos diferentes tipos de árboles, pero en particular nos llamó la atención CART porque no teníamos que suponer cosas sobre los datos para trabajar sobre ellos, funciona bien con grandes cantidades de datos, además es fácil entender la resolución del árbol sobre si una planta tiene roya o no, y como plus a lo anterior, resulta más sencilla o intuitiva su implementación.

# Análisis de Complejidad

| **Operaciones** | **Complejidad** |
| --- | --- |
| Construcción del árbol por | O(m(log n)) |
| nodo |  |
|  | M: cantidad de |
|  | características |
|  | N: cantidad de la muestra |
| Construcción total | O(2^n) |
| Lectura y almacenamiento de datos | O(n\*m)  n: número de columnas  m: número de filas |
| Precisión del subconjunto | O(n) |
| split | O(n/2) |

# Tiempos de Ejecución

|  | Conjunto de datos 1 | Conjunto de datos 2 |
| --- | --- | --- |
| creación | 6s | 4.5s |

**Tabla 2:** Tiempos de ejecución de las operaciones de la estructura de datos con diferentes conjuntos de datos

# Memoria

|  | Conjunto de datos 1 | Conjunto de datos 2 |
| --- | --- | --- |
| Consumo de memoria | 15 MG | 20 MG |

**Tabla 3:** Consumo de memoria de la estructura de datos con diferentes conjuntos de datos

# Análisis de los resultados

Así, viendo cómo funciona el algoritmo, notamos que consume más memoria y se demora más en ejecutarse el conjunto de datos 2, debido a que tiene mayor número de registros; aún así podemos evidenciar que, con muchos valores, sigue siendo relativamente rápido

|  | | Conjunto datos 1 | de | Conjunto datos 2 | de |
| --- | --- | --- | --- | --- | --- |
| Tiempo creación | de | 3.5s | | 6s | |
| Consumo memoria | de | 15 MG | | 20 MG | |

**Table 4:** Análisis de los resultados obtenidos con la implementación de la estructura de datos

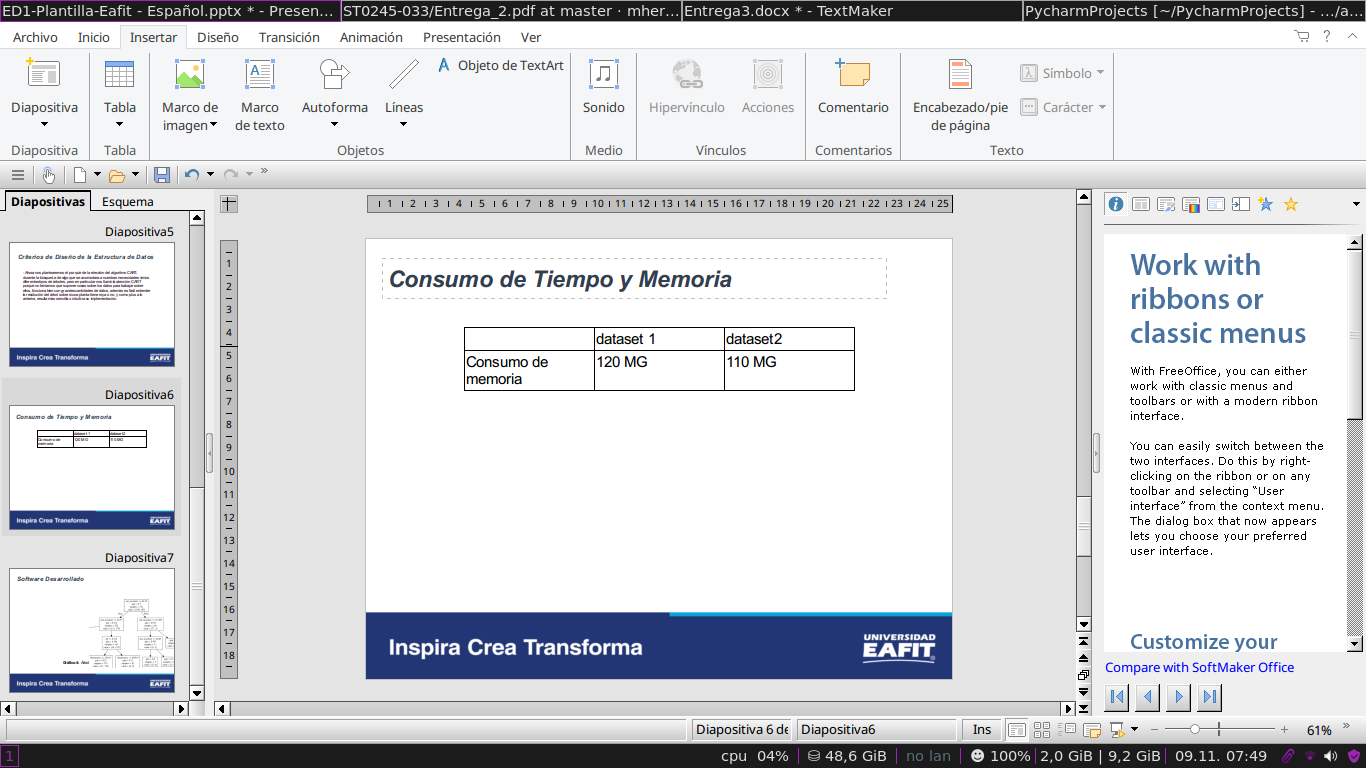
# Detección de la roya por medio de árboles de decisión

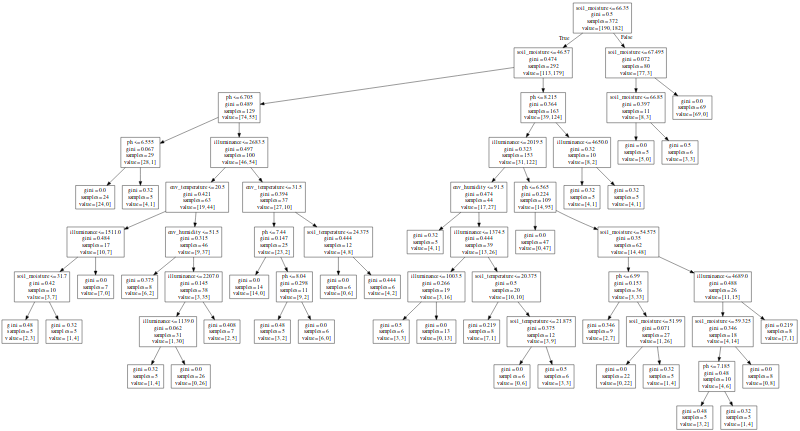


# Operaciones de la estructura de datos

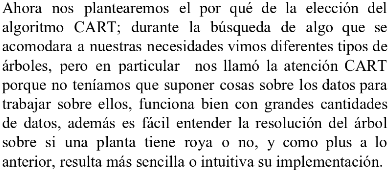
**Gráfica 4:** Imagen de una operación de borrado de una lista encadenada

# 5.5 Memoria



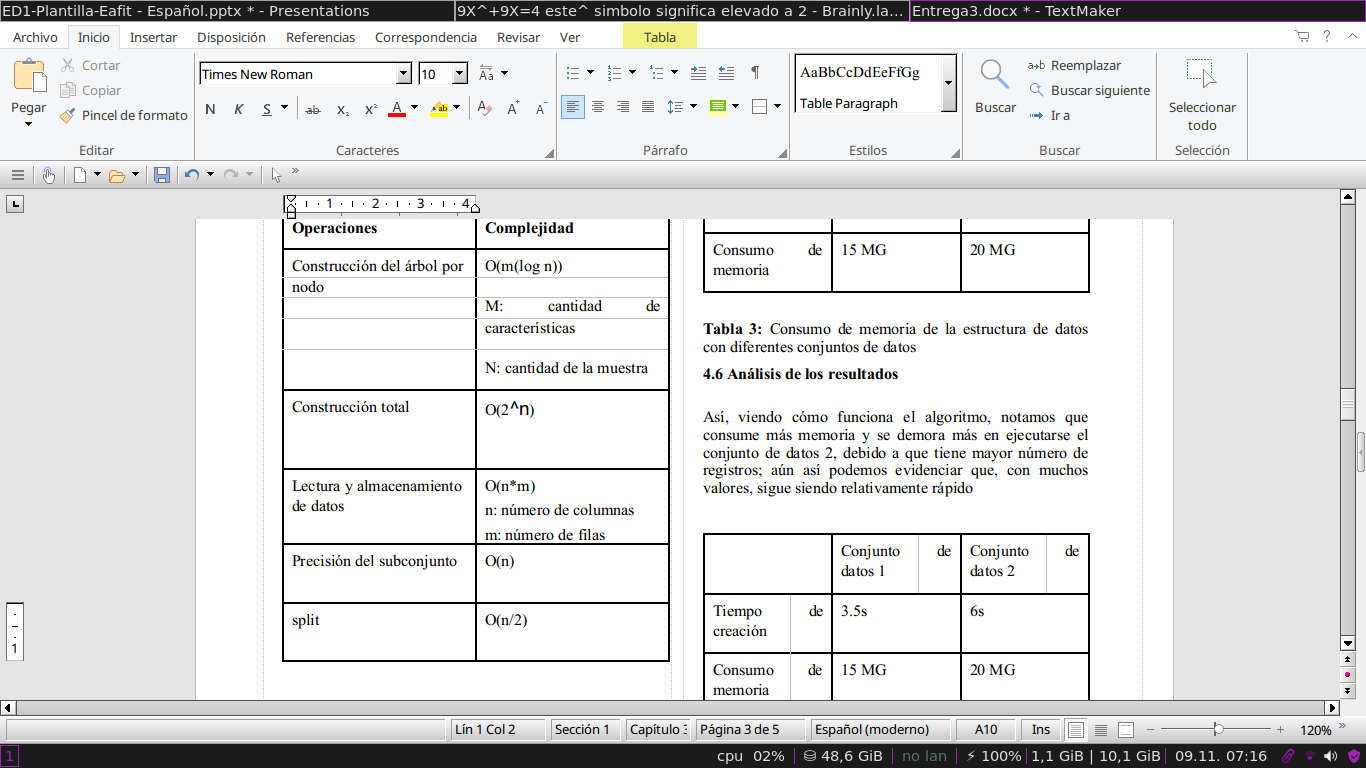


* 1. **Criterios de diseño de la estructura de datos**

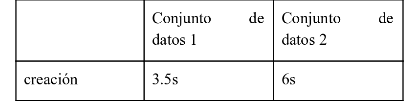


# Análisis de la Complejidad

:

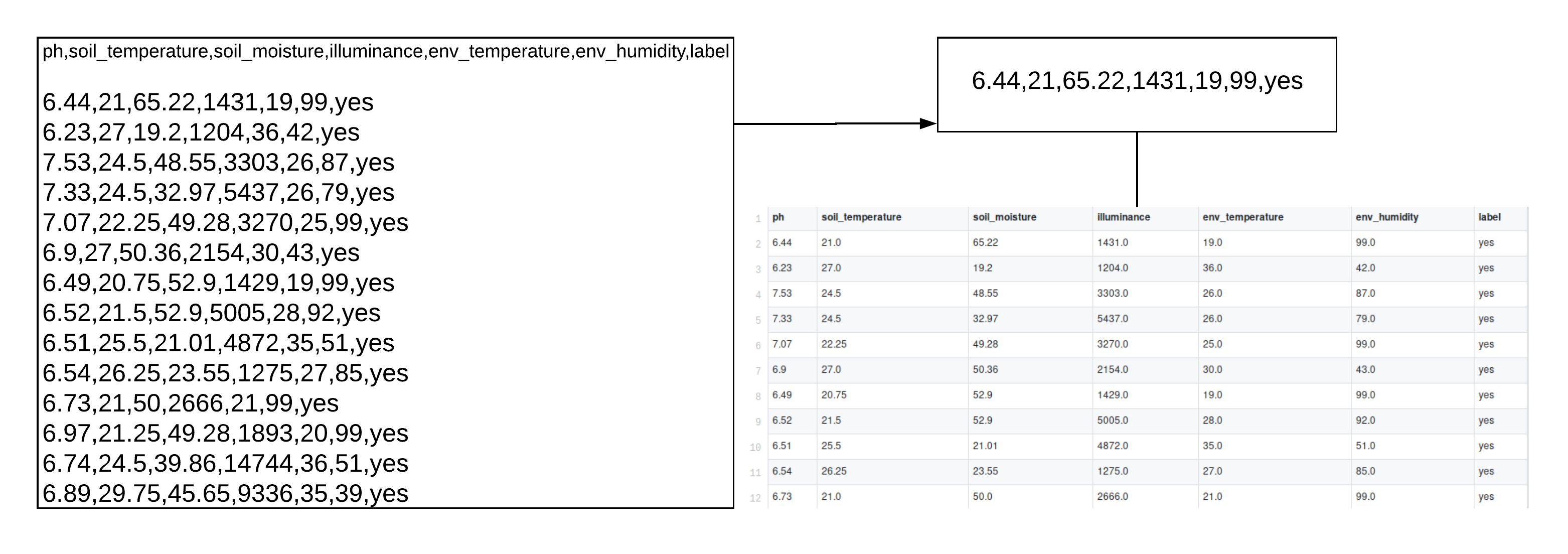


# Tiempos de Ejecución



# 5.6 Análisis de los resultados

Expliquen los resultados obtenidos. Hagan una gráfica con los datos obtenidos, como por ejemplo:



# CONCLUSIONES

La roya es un problema que logra afectar a todo el sector primario, pues debido a esta, en el último siglo se han generado grandes pérdidas, que equivalen al menos al 30% del cultivo, siendo esto un porcentaje importante que probablemente pueda ser mayor dado a que no hay muchas técnicas certeras de detección de esta plaga, y cuando se logra evidenciar, resulta varias veces inútil, pues termina siendo tardío el diagnóstico.

Siendo el anterior problema, mediante la solución se logra, por medio del análisis de las variables (pH, iluminación, humedad y la temperatura del ambiente y del suelo) entrenar un árbol de decisión, que predice si una planta, dadas sus características tiene la plaga o no, teniendo un acierto aproximado del 85%, lo que puede tener un impacto significativo, pues generaría pérdidas más bajas de bultos del café caturra.

Con la solución final se logra predecir el 100% de otro set de datos, lo que no se lograba con la primera versión, además, se generan productos más organizados, siendo este un documento donde se establece cuáles plantas tienen roya y cuáles no.

# Trabajos futuros

En un futuro sería interesante implementar árboles n-arios para una predicción quizá un poco más precisa, resolviendo problemas cada vez más complejos por medio del entrenamiento de árboles más concretos y complejos, siempre con el fin de responder a una problemática presente en el contexto por medio de la tecnología.

# AGRADECIMIENTOS

# Esta investigación fue soportada en su mayoría por la beca de Generación E de la que somos beneficiarios. También, por la monitora Isabel Piedrahita, quien nos orientó con un par de cosas sobre las librerías de pandas y sckitlearn, además de clarificar algunos conceptos sobre árboles necesarios para llevar a cabo el proyecto.

# REFERENCIAS

# Referenciar las fuentes usando el formato para referencias de la ACM. Léase en [http://bit.ly/2pZnE5](http://bit.ly/2pZnE5g)g Vean un ejemplo:

# Adobe Acrobat Reader 7, Asegúrense de justificar el texto. <http://www.adobe.com/products/acrobat/>.

# Fischer, G. and Nakakoji, K. Amplifying designers’ creativity with domainoriented design environments. in Dartnall, T. ed. Artificial Intelligence and Creativity: An Interdisciplinary Approach, Kluwer Academic Publishers, Dordrecht, 1994, 343-364.

# López, B. ALGORITMO ID3. Recuperado de <http://www.itnuevolaredo.edu.mx/takeyas/>

# IBM knowledge center. Nodo C5.0. Recuperado de <https://www.ibm.com/support/knowledgecenter/es/SS3RA7>[\_sub/modeler\_mainhelp\_client\_ddita/clementine/c50node\_](https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/c50node_general.html) [general.html](https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/c50node_general.html)