SISTEMA EXPERTO BASADO EN EMPAREJAMIENTO DE PATRONES

Universidad
del Cauca

EMMANUEL GERARDO LASSO SAMBONY

Tesis de Maestría en Ingeniería Telemática

Director:
Juan Carlos Corrales Muñoz
Doctor en Ciencias de la Computación

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telemática
Línea de Investigación en e-@mbiente
Popayán, 2016

EMMANUEL GERARDO LASSO SAMBONY

SISTEMA EXPERTO BASADO EN EMPAREJAMIENTO DE PATRONES

Tesis presentada a la Facultad de Ingeniería
Electrónica y Telecomunicaciones de la
Universidad del Cauca para la obtención del
Título de

Magíster en:
Ingeniería Telemática

Director(a):
Juan Carlos Corrales Muñoz
Doctor en Ciencias de la Computación

Popayán
2016

A mis padres, mi ejemplo y fortaleza
A mis hermanas, por su infinito amor
A mi tutor, por ser el guía en el camino
A mis amigos, por su incondicional apoyo

Resumen Estructurado

Antecedentes: Para la agroindustria, las enfermedades en cultivos constituyen uno de sus problemas más frecuentes que generan grandes pérdidas económicas y baja calidad en la producción. Por otro lado, recientes investigaciones proponen el desarrollo de sistemas expertos para resolver este problema, haciendo uso de técnicas de minería de datos e inteligencia artificial. Además, los grafos pueden ser usados para el almacenamiento de los diferentes tipos de variables que estén presentes en un ambiente de cultivos, permitiendo la aplicación de técnicas de minería de datos en grafos como el emparejamiento de patrones en los mismos. De esta manera, el desarrollo de un sistema experto para enfermedades en cultivos basado en emparejamiento de patrones en grafos, puede generar una solución para la identificación de condiciones favorables para una enfermedad en particular, como punto de partida para la toma de decisiones.

Objetivos: Desarrollar un Sistema Experto basado en Emparejamiento de Patrones en Grafos para la detección de condiciones favorables para la roya en cultivos de café en Colombia.

Métodos: es propuesto un Sistema Experto, caracterizado a partir del conocimiento de expertos en la roya, como punto de partida para la extracción de reglas que determinen condiciones favorables para la enfermedad, a partir de la inducción de árboles de decisión, aplicada a un conjunto de datos de monitorización y propiedades de cultivo. Estas reglas son expresadas como patrones de grafos, los cuales son buscados dentro de un repositorio de información de cultivos expresado como grafos, con el fin de encontrar las coincidencias de estos patrones, que determinan el estado de un cultivo de café frente a la roya.

Resultados: La presente propuesta entregó como resultados: un conjunto de variables predictivas para la roya en el café, definidas a partir del conocimiento de expertos; un conjunto de patrones de grafos para la identificación de condiciones favorables para tres tasas de infección de roya; la adaptación de un algoritmo para el emparejamiento de patrones en grafos y un sistema experto para la detección de tasas de infección de roya en el café, basado en emparejamiento de patrones en grafos.

Conclusiones: El conocimiento de expertos en el estudio de la roya en el café, permite la construcción de variables predictivas específicas para la enfermedad y su presencia dentro de modelos generados por técnicas de minería de datos . A partir de estos modelos pueden ser extraídas reglas para ser expresadas como patrones de grafos, aprovechando su

expresividad e interpretabilidad. Siendo así, la aplicación del emparejamiento de patrones en grafos da como resultado la condición de un cultivo frente a la enfermedad. Por otro lado, la carencia de una gran cantidad de datos restringe la calidad del proceso de generación de modelos y validación del sistema.

Palabras Claves: sistema experto, grafo, emparejamiento de patrón, minería de datos, cultivos, enfermedad, agricultura.

Structured Abstract

Background: For agroindustry, crop diseases constitute one of the most common problems that generate large economic losses and low production quality. On the other hand, recent research proposes the development of expert systems to solve this problem, making use of data mining and artificial intelligence techniques. Furthermore, graphs can be used for storage of different types of variables that are present in an environment of crops, allowing the application of graph data mining techniques like graph pattern matching. Therefore, the development of an expert system for crop disease based on graph pattern matching, can generate a solution for the identification of favorable conditions for a particular disease, as a starting point for decision-making.

Goals: Develop an expert system based on graph pattern matching to detect favorable conditions for coffee rust in Colombian crops.

Methods: This work proposes an expert system, characterized from expert knowledge in coffee rust, as a starting point for the extraction of rules that determine conditions favorable for this disease, from induction of decision trees, applied to a dataset of monitoring and cultivation properties. These rules are expressed as patterns of graphs, which are sought within an information repository crop expressed as graphs, in order to find the similarities of these patterns, which determine the state of a crop of coffee against rust.

Results: A set of predictive variables for coffee rust, defined from expert's knowledge; a set of graph patterns to identify three favorable conditions for rust infection rates; adaptation of an algorithm for graph pattern matching and an expert system for detecting coffee rust infection rates, based on graph pattern matching.

Conclusions: Expert knowledge in coffee rust allows the construction of specific predictive variables for the disease and include it within models generated by data mining techniques. From these models, can be extracted rules to be expressed as graph patterns, using their expressiveness and interpretability. Thus, the application of graph pattern matching results in the condition of a crop against disease. Moreover, the lack of a large amount of data restricts the quality of model generation process and the system validation.

Keywords: expert system, graph, pattern matching, data mining, crop, disease, agriculture.

Contenido

Lista de figuras

Lista de tablas

Lista de ecuaciones

1 Introducción

En este capítulo son presentadas las consideraciones que motivan al desarrollo del presente trabajo de investigación. De forma seguida, son presentados los objetivos que buscan generar una solución al problema encontrado. Por último, son descritas las contribuciones del proyecto y el contenido de la monografía.

1.1. Planteamiento del Problema

En Colombia, según el Plan Nacional de Ciencia, Tecnología e Innovación, la Agroindustria está posicionada como un sector de producción tradicional en el país, en el que la productividad puede mejorar significativamente a través del uso y aplicación de las Tecnologías de la Información y la Comunicación (TIC) [1]. El objetivo de este sector es obtener productos de alta calidad para hacer frente a los mercados globales. Para lograrlo, deben afrontar los problemas más frecuentes y causantes de grandes pérdidas, como por ejemplo, las enfermedades en los cultivos. Por este motivo, diferentes Instituciones del sector agrícola han invertido recursos con el fin de implementar soluciones desde las TIC hacia sus problemáticas. Además, cuentan con expertos que dedican sus investigaciones a identificar las causas de las enfermedades y proponer tratamientos para su control.

Por otro lado, desde las ciencias de la computación han surgido propuestas para disminuir los efectos causados por diversas enfermedades y plagas en cultivos. Una de las soluciones informáticas utilizadas para este fin son los sistemas expertos, los cuales poseen información de uno o más especialistas en un área específica [2]. El objetivo de los sistemas expertos es generar soluciones a un problema dado haciendo uso de una base de conocimiento del área de aplicación. Para este fin, son utilizadas diferentes técnicas de la inteligencia artificial, como los árboles de decisión, redes neuronales, entre otras. Asimismo, existe una tendencia dentro del área de la informática hacia el uso de grafos como estructura de datos, los cuales tienen una naturaleza generalmente dinámica y consisten en un conjunto de nodos que están relacionados a través de aristas. A partir de esto, la representación de información que brindan los grafos puede ser aprovechada para el almacenamiento y análisis de los distintos tipos de variables [9] que están presentes en un ambiente de cultivos .

Diferentes investigaciones proponen el desarrollo de sistemas expertos para la detección y tratamiento de enfermedades en cultivos. Varias de ellas hacen uso de los árboles de decisión como modelo de predicción, dada una base de conocimiento, que categoriza una serie de condiciones, de manera similar a reglas de inferencia, para llegar a la solución de un problema. Por otra parte, la representación de información basada en grafos permite una gran interpretabilidad y expresividad en su estructura, convirtiéndose en una tecnología de gran uso en estos días. Sin embargo, las investigaciones en estas áreas no consideran

la representación de reglas obtenidas a partir de modelos predictivos como subgrafos o patrones de grafos, los cuales pueden ser buscados dentro de un repositorio de información en grafo, con el fin de encontrar las ocurrencias de las premisas descritas en cada regla.

De lo anteriormente dicho, existe un sector específico con potencial para aplicar las tecnologías mencionadas, como es el sector cafetero. Precisamente, una de las principales enfermedades que afectan a los cultivos del café a nivel mundial es la roya del cafeto, la cual en Colombia es responsable de pérdidas de hasta el 30% en variedades suceptibles a ella. Ante esta problemática, las autoridades relacionadas han desplegado estaciones de monitorización a lo largo del país y desarrollado una plataforma climática cafetera. Dicha plataforma es aprovechada como herramienta para los investigadores que intentan encontrar el efecto de las variaciones climáticas en la generación y evolución de enfermedades como la roya [3].

Teniendo en cuenta las anteriores consideraciones, con el fin de aprovechar las ventajas de los enfoques de la informática mencionados y disminuir las pérdidas en el sector agrícola causadas por diferentes enfermedades, la pregunta de investigación que motiva el desarrollo del presente trabajo de grado es la siguiente:

¿Cómo integrar el emparejamiento de grafos a un sistema experto con el propósito de prevenir enfermedades en el sector agrícola Colombiano?

1.2. Escenario de motivación

Para la agroindustria, las enfermedades en cultivos constituyen uno de sus problemas más frecuentes, generando grandes pérdidas económicas y baja calidad en los productos. En Colombia, la producción de café se encuentra entre las principales actividades agrícolas del país y, a través de los años, una de las enfermedades que más impacto ha tenido sobre estos cultivos ha sido la roya del cafeto, causada por el hongo *Hemileia Vastatrix*, originando pérdidas de hasta un 30% de la producción en variedades susceptibles. Ante este problema, el uso de fungicidas y el cambio a variedades de café resistentes a la roya se han presentado como alternativas eficaces para el control de la enfermedad. Sin embargo, estas opciones afectan la calidad del producto y elevan su costo de producción. A partir de esto, expertos en el área han estudiado los diferentes factores y desarrollo de esta enfermedad, encontrando una serie de condiciones, tanto climáticas como propiedades de los cultivos, que afectan a cada etapa del desarrollo del hongo. Adicionalmente, muchos de estos factores pueden ser medidos en un cultivo a través de instrumentos de monitorización climática y registros de propiedades agronómicas, lo que da como resultado un conjunto de datos que caracteriza las unidades de cultivo y que pueden ser analizados y asociados a los brotes de roya presentados en cada zona.

Teniendo en cuenta la problemática mencionada anteriormente, es considerado conveniente contar con un sistema experto que haga uso de una representación basada en

grafos de la información monitorizada en cultivos de café, obteniendo un modelo de datos de gran interpretabilidad y expresividad, que a su vez puede ser usado para buscar patrones de la enfermedad definidos por expertos, con el fin de identificar condiciones favorables para epidemias de roya de forma temprana y, de esta forma, poder alertar a los actores involucrados para que tomen acciones preventivas y correctivas, mejorando la productividad y competitividad de los cultivos cafeteros colombianos.

1.3. Objetivos

A continuación, son presentados el objetivo del presente trabajo de investigación, así como los objetivos específicos que buscan dar solución a el problema identificado anteriormente.

1.3.1. Objetivo General

Crear un sistema experto para la detección de condiciones favorables de aparición de enfermedades para el sector agrícola, basado en emparejamiento de grafos como método para evaluar las reglas generadas por expertos.

1.3.2. Objetivos Específicos

II. Caracterizar el sistema experto, de acuerdo a las necesidades específicas del problema y a partir del conocimiento del experto en la temática.

III. Adaptar la representación formal del sistema experto en términos de grafos.

IV. Verificar el cumplimiento de las reglas del sistema experto con base en emparejamiento de grafos (graph pattern matching).

V. Construir y evaluar experimentalmente un prototipo que implemente los anteriores objetivos en el dominio agrícola (e.g. detección de las condiciones para presencia de roya en el café).

1.4. Contribuciones

- Una evaluación del desempeño de técnicas de emparejamiento de grafos, como los algoritmos de Ullman, VF2 y DualIso.

- Un conjunto de reglas expresadas como patrones de grafos, las cuales fueron construidas a partir de la inducción de árboles de decisión, basada en el conocimiento de expertos en la roya del café.

- Una representación basada en grafos de los datos monitorizados por estaciones meteorológicas en diferentes zonas cafeteras.

- Un sistema experto que haga uso del enfoque de los árboles de decisión, representados por patrones, que a su vez serán evaluados a través de una técnica de emparejamiento de grafos.

- Validación del prototipo dentro del dominio de aplicación tratado, con el fin de contribuir a la detección temprana de condiciones favorables para el desarrollo de la roya en el café.

- Un artículo expuesto en el Workshop TIC-@gro del VII Congreso Iberoamericano de Telemática realizado entre los días 10 al 12 de junio del 2015 en Popayán – Cauca, el cual presenta la propuesta del sistema experto basado en emparejamiento de patrones en grafos. Ver Anexo B.

- Un artículo publicado en la Revista Ingenierías Universidad de Medellín, en el Número 29 de enero-abril de 2016, clasificada A2 Publindex-Colciencias | ISSN 1692-3324, el cual corresponde a la propuesta del sistema experto basado en emparejamiento de patrones en grafos, conteniendo adicionalmente avances en algunos de los módulos de la propuesta. Ver Anexo C.

- Un artículo expuesto en: MTSR 2015: 9th Metadata and Semantics Research Conference. Special track on Metadata and Semantics for Agriculture, Food & Environment (AgroSEM'15), realizado entre el 9 y 11 de septiembre de 2015 en Manchester – Reino Unido, el cual presenta la extracción a través de árboles de decisión de reglas para identificar condiciones favorables para la roya en el cafeto, y su representación como patrones de grafo. Además, este artículo fue incluido en el capítulo “Metadata and Semantic Research”, en el volumen 544 de la serie de Springer “Communications in Computer and Information Science”. ISSN 1865-0929. Ver Anexo D.

1.5. Contenido de la Monografía

La monografía se encuentra organizada en seis capítulos los cuales son presentados a continuación:

- **Capítulo II. Estado Actual del Conocimiento**

  Presenta una visión general sobre los trabajos relacionados y los conceptos que giran en torno a los sistemas expertos en entornos agrícolas, minería de datos,

representación de información basada en grafos y procesos de análisis y búsqueda en estos.

- **Capítulo III. Caracterización del Sistema Experto**

  Expone el proceso llevado a cabo para la caracterización del sistema experto para cultivos de café en Colombia y Brasil, el cual sigue la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). En este proceso son descritos la comprensión de la problemática tratada y los factores que inducen sobre ella, la comprensión de los datos asociados al entorno, la preparación de estos datos con el fin de generar un conjunto de reglas que los relacionen con la enfermedad tratada y, por último, la presentación de estas reglas y su interpretación.

- **Capítulo IV. Verificación de las reglas del Sistema Experto con base en emparejamiento de patrones en grafos**

  Explica la representación basada en grafos de los datos de monitorización y propiedades físicas de cultivos, y las reglas extraídas a través de árboles de decisión (patrones de grafo). De forma seguida, son presentados algunos algoritmos para realizar emparejamiento de patrones en grafos etiquetados (Ullman, VF2 y DualIso), para finalmente escoger uno de estos y describir su adaptación para el propósito de este trabajo de investigación.

- **Capítulo V. Experimentación y Evaluación**

  Presenta el prototipo, el proceso de evaluación y las pruebas ejecutadas al sistema experto, con el fin de analizar la calidad de los resultados y su rendimiento.

- **Capítulo VI. Conclusiones y Trabajos Futuros**

  Finalmente, los resultados del trabajo realizado son analizados, las principales contribuciones obtenidas en la ejecución del proyecto son detalladas y un conjunto de recomendaciones importantes para el desarrollo de trabajos futuros son expuestas.

2 Estado actual del conocimiento

En este capítulo se presentan las bases teóricas para comprender la temática del presente trabajo de maestría, el cual propone el desarrollo de un sistema experto basado en emparejamiento de patrones en grafos para la detección de condiciones favorables hacia epidemias de roya en cultivos de café en Colombia. De forma seguida, se exponen los trabajos relacionados en el área de estudio. Para finalizar, se presenta una discusión sobre los trabajos analizados y cómo pueden ser aprovechados distintos enfoques para llevar a cabo el objetivo de investigación del presente trabajo.

2.1. Conceptos Generales

2.1.1. Sistema Experto

Un sistema experto es una rama de la inteligencia artificial, compuesto por una serie de herramientas que hacen uso del conocimiento humano almacenado en un ordenador con el propósito de resolver un problema [4]. Precisamente, el conocimiento utilizado en estos sistemas proviene de personas que tienen un alto grado de experiencia en el área donde es necesario tomar una decisión o encontrar una solución a un problema dado. Bajo estas condiciones, los sistemas expertos han sido implementados en gran variedad de áreas de conocimiento, requiriendo que estos sean capaces de explicar los razonamientos realizados e incluir nueva información en su base de conocimiento.

De manera general según [5], la estructura de los sistemas expertos están compuestos por: un **Subsistema de adquisición de conocimiento**, encargado de la acumulación, transferencia y transformación de la base de conocimiento, a partir de la información de expertos; **base de conocimiento**, que contiene el conocimiento usado para entender, formular y resolver un problema; **motor de inferencia**, compuesto por técnicas y algoritmos que tienen la habilidad de formular conclusiones; **interfaz de usuario**, usada para la comunicación entre usuario y ordenador; Por último, un **subsistema de explicación**, utilizado para justificar la solución dada por el sistema.

Adicionalmente, los sistemas expertos hacen uso de una o más técnicas de la inteligencia artificial (árboles de decisión, redes neuronales, algoritmos genéticos, entre otras) con el fin de mejorar su razonamiento y obtener mejores resultados al inferir sobre la base de conocimiento.

2.1.2. Árboles de decisión

Los árboles de decisión son estructuras recursivas para expresar reglas de clasificación [6] y, en el área de la minería de datos, representan técnicas para explorar grandes y complejas

cantidades de datos, con el fin de encontrar patrones útiles dentro de ellos. En el mismo sentido, cumplen con la función de predecir y explicar la relación entre una nueva instancia de información y su valor objetivo [7].

Un área de aplicación de los árboles de decisión es la tarea de clasificación, la cual es un problema importante en la minería de datos. En la clasificación, los datos de entrada, llamados *conjunto de entrenamiento*, consisten en múltiples ejemplos (registros), los cuales contienen un conjunto de características o atributos. Adicionalmente, cada ejemplo está etiquetado con una *clase* especial. El objetivo de la clasificación es analizar los datos de entrada y desarrollar un modelo para cada clase usando las características presentes en los datos. Este modelo es usado para clasificar futuros datos de entrenamiento para los cuales la clase aún no esté definida. Además, el modelo desarrollado está representado por un árbol de decisión [8].

La representación en forma de árbol indica la bifurcación de sus ramas en función de los valores de las variables que intervienen en el conjunto de datos, donde cada bifurcación termina en una acción concreta. A partir de esto, el árbol tiene distintos tipos de nodos y arcos que unen estos nodos. Dichos nodos pueden representar, por un lado, la evaluación que debe ser hecha a una variable del conjunto de datos y, por otro lado, el valor final que devolverá el árbol de decisión. Por su parte, las ramas representan los distintos caminos que pueden ser tomados según la respuesta de cada nodo.

2.1.3. Grafos

Un grafo está definido como $G = (V, E)$, donde V denota un conjunto finito de nodos conectados por enlaces directos o vértices E, tales que $E \subseteq VxV$ corresponden a las relaciones existentes entre los nodos del grafo [9]. De esta forma, cada grafo puede ser asociado a su matriz característica $M = (m_{i,j})_{m*n}$, donde:

$$m_{i,j} = \begin{cases} 1 \ (v_i, v_j) \in E \\ 0 \ en \ otro \ caso \end{cases}$$

De esta forma, los grafos representan una estructura de almacenamiento, donde se puede analizar cómo el conjunto de entidades están relacionadas entre sí y las características de esas conexiones. La nueva generación de sistemas de bases de datos, que generalmente trabajan con documentos estructurados, a menudo modelada información a través de árboles y grafos.

Adicionalmente, la estructura del grafo y sus componentes tiene algunas características útiles para la fuente de información que representa. En relación a la fuerza de las conexiones entre las entidades del grafo, puede ser nominal o binaria (representa la ausencia o presencia de una conexión); polarizada (representa una conexión negativa, una

conexión positiva o sin conexión); ordinal (representa si la conexión es la más fuerte, la más débil, etc); o valorada (medido en un intervalo) [10].

2.1.4. Minería de datos en Grafos

La Minería de Datos en Grafos, a veces llamada *minería de datos basada en grafos*, es la extracción de conocimiento útil y nuevo desde una representación de información en forma de grafos. La estructura de la información que puede ser extraída a partir de grafos es, a su vez, un grafo. Por este motivo, el conocimiento es referido como patrones del grafo original, representado por subgrafos o expresiones abstractas de las tendencias reflejadas en los datos [11].

En algunos casos, los grafos pueden ser dinámicos y experimentar una evolución constante. Esto significa que la estructura del grafo puede cambiar rápidamente con el tiempo. Para tales casos, el aspecto temporal del análisis de los datos es un escenario interesante para su estudio. Con el fin de obtener un acceso a la información contenida en un grafo, son necesarios lenguajes de consulta para la gestión y manipulación de los datos estructurales. Igualmente, es necesario obtener un acceso eficiente a la información del grafo, para así resolver las consultas en el menor tiempo posible. Por esto, existen esfuerzos en el estudio del diseño de índices para las estructuras de los grafos [12].

A su vez, existe una técnica dentro de la minería de datos en grafos llamada *Emparejamiento de Patrones en Grafos*, que está definido como: *dado un grafo de datos G, y un patrón de grafo Q, encontrar todas las coincidencias en G para Q* [13]. Este tipo de búsquedas generalmente están dirigidas a encontrar entidades que tengan unas características específicas en sus atributos y en las relaciones con otros nodos del grafo. En este sentido, el patrón buscado puede ser visto como una serie de condiciones dentro de los atributos del grafo, de manera similar a la evaluación hecha por los árboles de decisión. Adicionalmente, los sistemas basados en esta técnica son divididos en sistemas de búsqueda de patrones con coincidencia exacta y sistemas de búsqueda de patrones tolerantes a errores. Aunque la coincidencia exacta ofrece una manera rigurosa de obtener un resultado en términos matemáticos, por lo general sólo es aplicado a un conjunto limitado de problemas del mundo real. En cambio, el sistema tolerante a errores es capaz de hacer frente a la distorsión en los datos del grafo, lo cual es un elemento frecuente en los problemas del mundo real. No obstante, este enfoque generalmente es menos eficiente a nivel computacional [14].

2.2. Trabajos relacionados

A continuación son presentados los trabajos de investigación relacionados con la propuesta en desarrollo, considerando los sistemas expertos y árboles de decisión en el sector

agrícola, además del emparejamiento de patrones en grafos, como las áreas de interés para el desarrollo de este proyecto.

2.2.1. Sistemas Expertos

Los Sistemas Expertos tienen como objetivo generar soluciones a problemas a través del análisis de una base de conocimiento construida a partir del conocimiento de personas con un grado de experiencia dentro de un área de conocimiento. En esta sección será presentada una serie de trabajos que implementan los sistemas expertos dentro de la agricultura y medio ambiente.

En la primera de las aproximaciones en esta área, llevada a cabo por Mansingh [2], es presentado un sistema experto para la gestión de plagas y enfermedades del café en un país en vía de desarrollo. En este sistema está considerada una base de conocimiento, que consta de reglas creadas a partir del conocimiento de expertos, un motor de inferencia y un módulo de explicación. Para la solución de problemas, el usuario debe ingresar una serie de parámetros que están relacionados con su cultivo y, a partir de estos datos, son evaluadas las reglas generadas por los expertos. Por último, el sistema es sometido a tres evaluaciones: validación, que verifica si la recomendación dada por el sistema era correcta; aceptación, por parte de los usuarios; efectividad, la cual es una medida de la disminución del impacto ambiental al aplicar las recomendaciones dadas por el sistema.

De manera similar al anterior trabajo mencionado, en [15] es construido un sistema experto para la ayuda en el diagnóstico de enfermedades del café, a partir de la exploración de las plantas por parte del agricultor. La estructura del sistema está basada en técnicas de lógica difusa y árboles de decisión, usados para representar una serie de condiciones presentes ante la existencia de alguna enfermedad, las cuales son definidas por expertos. La precisión del sistema es evaluada en 20 muestras, con un resultado del 85%.

En un tercer trabajo [16] es desarrollado un sistema experto para la identificación de insectos que representan un riesgo para la vida forestal y además propone un tratamiento relevante para el caso encontrado. El sistema hace uso de un motor de inferencia basado en reglas, que a su vez son construidas por expertos. La base de conocimiento contiene reglas y descripciones de parámetros, representados a través de ontologías de dominio. Finalmente, la evaluación a la herramienta desarrollada está basada en criterios como efectividad, usabilidad y facilidad para uso.

Con un propósito similar, son presentados sistemas expertos dirigidos al tratamiento de plagas y enfermedades en distintos sectores agrícolas (plantas frutales[5], mango[17], arroz[18], plantas solanáceas[19], cultivos de olivo[20], tomate[21] y leguminosas[22]). Estos trabajos tienen en común la recolección de información de expertos en el tema que cada una de las investigaciones aborda, con el fin de construir reglas simples del tipo *condición-conclusión (IF-THEN)*. Además, presentan interfaces de usuario orientadas a la

introducción del valor de distintos parámetros por parte de la persona que lo usa. Estos valores son utilizados como fuente de información a ser evaluada por las reglas generadas. El objetivo de estos trabajos es generar recomendaciones para el tratamiento de una plaga o enfermedad y el núcleo de los sistemas generados está compuesto por un motor de inferencia, el cual valida las reglas obtenidas de los expertos. Para la evaluación de las herramientas presentadas en cada trabajo, en la mayoría de casos hacen uso de medidas de satisfacción por parte de usuarios y expertos, tales como: validación, que verifica si la recomendación dada por el sistema es correcta; aceptación, por parte de los usuarios del sistema; claridad, en el proceso de recolección de información. Sumado a esto, existe una comprobación de casos de estudio, realizada a través de muestras que contienen parámetros controlados y que han sido definidos por los expertos colaboradores de cada proyecto.

Por último, en [23] y [24] son desarrollados sistemas expertos y de soporte a la toma de decisiones que consideran la información climática obtenida a través de sensores meteorológicos y el conocimiento de personas con cierto grado de experiencia en el dominio en el cual son aplicados. El objetivo de considerar estas dos fuentes de información es generar recomendaciones y alertas que vayan de acuerdo a la variabilidad climática de la zona de influencia de cada una de las estaciones de monitorización de clima, analizando el estado de cada una de las variables medidas y relacionándolas de acuerdo a las reglas suministradas por los expertos. En primer lugar, en [23] el sistema experto funciona de forma paralela al sistema de soporte a la toma de decisiones. En el momento en que es detectada una situación de riesgo, el sistema experto es invocado para ayudar al usuario a tomar acciones ante la problemática. En segundo lugar, en [24] el sistema experto se convierte en una guía para el usuario según el pronóstico del clima que realizan al monitorizar distintas variables meteorológicas.

2.2.2. Árboles de Decisión

En esta sección serán citadas las investigaciones más relevantes en el área de los árboles de decisión dentro de dominios de aplicación relacionados con la agricultura. Además, son presentados trabajos de investigación que abordan el uso de representaciones basadas en grafos junto a los árboles de decisión.

En primer lugar, el trabajo de investigación presentado en [25] propone el uso de árboles de decisión difusos con el fin de generar alertas de aparición de roya en el café. Los árboles usados representan umbrales de distintas variables que intervienen en esta problemática, tanto para situaciones de prevención y curación o tratamiento de la enfermedad. Además, la aproximación está basada en el clásico árbol de decisión C4.5 y es generada a partir del análisis de un conjunto de datos de aproximadamente 8 años. Para su evaluación, la herramienta es comparada con árboles de decisión tradicionales, basados en el modelo J48. El resultado de esta comparación muestra que la propuesta de este trabajo tiene una mejor precisión y mayor facilidad de interpretar el resultado dado.

Con un propósito similar, los autores de [26] y [27] desarrollan un árbol de decisión para analizar las epidemias de roya en el café. Los datos usados para generar el árbol corresponden a 364 muestras que contienen información sobre la temperatura, precipitación y humedad relativa. Estas aproximaciones hacen uso del algoritmo básico de inducción de árboles de decisión, propuesto por Han y Kamber [28]. Por último, el modelo brinda un apoyo para el entendimiento de cómo la interacción entre las variables analizadas conduce a epidemias de roya. Tras su ejecución, el modelo clasifica correctamente el 78% del conjunto de datos de entrenamiento, así como su precisión es estimada en 73% para la clasificación de nuevas muestras.

En la investigación desarrollada en [29] son utilizados dos algoritmos de árboles de decisión inductivos (ADI), basado en C4.5, y una variante del ADI combinado con Rough Set. Los ADI están basados en la búsqueda de una hipótesis o reglas dentro de un conjunto de ejemplos. Esta búsqueda es escalada y en dirección de los árboles más simples a los más complejos. De esta manera, los algoritmos descritos hacen parte de un conjunto de técnicas usadas para la generación de alertas ante la ocurrencia del oídio de mango en la India. Los resultados de la aplicación de estos algoritmos da como resultado un porcentaje de precisión de 83% para el ADI y 75% para el segundo algoritmo mencionado.

Por otra parte, existen trabajos de investigación que incorporan la representación basada en grafos a los árboles de decisión, con el fin de aprovechar las diferentes técnicas de análisis que existen para esta estructura de información. En [30], [31] los autores proponen un método llamado *Inducción de árbol de decisión basada en grafo (Decision Tree Graph-Based Induction) (DT-GBI)* [32], el cual construye un árbol de decisión para los datos de un grafo estructurado, mientras simultáneamente construye atributos para la clasificación. Al mismo tiempo, DT-GBI analiza un grafo empleando patrones discriminativos (subgrafos) como atributos del árbol de decisión generado y de forma recursiva extrae las ocurrencias de dichos patrones. Este método es usado para analizar un conjunto de datos asociados a la Hepatitis, encontrando condiciones que determinan el tipo de esta enfermedad presente en un paciente. Los resultados son satisfactorios, muchas de las muestras extraídas coinciden con la opinión de médicos expertos en la enfermedad y la taza de error obtenida es aceptable debido a los diferentes tipos de ruido contenidos en los datos.

Asimismo, en [33] son usados árboles de decisión con el fin de optimizar el indexado de una base de datos de grafos. Esta técnica construye un árbol de decisión que clasifica un grafo a partir de parámetros específicos. De esta manera, al realizar una consulta sobre el grafo, éste es recorrido haciendo uso del árbol generado, a través del cual se reduce el número de grafos candidatos a ser evaluados. El proceso mencionado es también utilizado para la aplicación de isomorfismo en grafos. Las evaluaciones realizadas muestran que este método mejora el rendimiento del proceso de emparejamiento de grafos.

2.2.3. Emparejamiento de Patrones en Grafos

Dentro de la minería de datos en grafos, el emparejamiento de patrones es una de las técnicas más exploradas por investigadores. A continuación son presentados diferentes trabajos de investigación que proponen técnicas y optimizaciones para el emparejamiento de grafos.

En primer lugar, W. Fan en sus investigaciones [34], [35] y [36] aborda el emparejamiento de patrones sobre grafos. Sus propuestas están orientadas hacia la eliminación del número de resultados excesivos ante una consulta y la generación de una puntuación para las coincidencias encontradas. De esta manera, se genera una lista ordenada de subgrafos candidatos. Adicionalmente, otras características de los algoritmos propuestos son: eficiencia en la evaluación de búsquedas en grafos de gran tamaño; soporte a la actualización de los datos y estructura del grafo; compresión del grafo para su análisis. Los datos utilizados para la evaluación de los algoritmos son simulados y reales. Por último, el soporte a actualizaciones del grafo presenta un mejor desempeño a los métodos propuestos por aproximaciones similares. A su vez, la compresión del grafo genera una reducción del 57% del tamaño original, lo cual reduce el tiempo de aplicación del emparejamiento en un 70%.

De manera similar, en [37] es presentado un software llamado GMT (Graph Matching Toolkit), que permite construir de manera gráfica patrones de grafo para su búsqueda. GMT utiliza un algoritmo para búsqueda de patrones llamado TruST [38], el cual genera un árbol binario de búsqueda para analizar los nodos del grafo. Los resultados de las pruebas realizadas no son presentados estadísticamente, sino bajo una inspección manual de la efectividad del proceso.

El sistema presentado en [39], llamado G-Path, tiene como función principal la búsqueda de patrones de ruta en grafos de gran tamaño. Con el fin de demostrar el funcionamiento del sistema, los autores construyen una aplicación web la cual permite la búsqueda de entidades y relaciones en grafos de volumen considerable. Al igual que el trabajo anterior, no existen medidas estadísticas del desempeño del sistema.

En la investigación llevada a cabo por Nisar y su equipo [40] son desarrollados algoritmos distribuidos para el emparejamiento de patrones sobre grafos. Estos algoritmos presentan un comportamiento escalable conforme es incrementado el número de procesos en paralelo. Por otro lado, cada uno implementaba un tipo de particionamiento, Min-cut y Round-robin, siendo el primero el más efectivo. Además, comprueban las ventajas de procesar los datos de los grafos de manera distribuida, a través de múltiples procesadores de tareas.

Por su parte, en [41] es presentado un algoritmo para el isomorfismo de grafos de gran escala. Este algoritmo es una mejora al algoritmo VF [42], llamado VF2, donde su mayor

mejora introducida es la estructura de datos empleada durante la exploración del espacio de búsqueda. La metodología propuesta está organizada de manera que reduzca significativamente los requerimientos de memoria para su ejecución. Con esto, la evaluación del algoritmo VF2 es presentada y comparada con el algoritmo de Ullman [43], obteniendo mejores medidas de desempeño a nivel de uso de memoria y tiempo de ejecución para VF2, especialmente en grafos de gran tamaño. En el mismo sentido, en [44] es presentado un algoritmo llamado *DualIso*, el cual tiene como objetivo realizar un emparejamiento de patrones en grafos etiquetados haciendo uso de una técnica de reducción del espacio de búsqueda llamada *Simulación doble de grafo* y es comparado con los algoritmos de Ullman y VF2, obteniendo mejores resultados en el uso eficiente de memoria para diferentes tamaños de consultas y tiempo de ejecución.

Por último, las investigaciones detalladas en [45], [46] y [47] proponen técnicas que implementan el emparejamiento de grafos de manera que sean optimizados el rendimiento y precisión en su aplicación. Lo anterior demuestra que existen investigadores que mantienen el interés en el emparejamiento de grafos, debido a que el grafo, como estructura de datos, cobra gran importancia dentro de las ciencias de la computación.

Teniendo en cuenta que un patrón dentro de un grafo puede ser visto como la representación de condiciones de atributos para una clase, esta técnica y los árboles de decisión pueden complementarse entre sí con el fin de resolver un problema en específico.

A partir de la revisión de los trabajos relacionados presentada, es identificado cómo los sistemas expertos pueden hacer uso de los árboles de decisión para expresar reglas generadas por personas con cierto grado de experiencia en un dominio de aplicación. La aplicación de estos sistemas en la agricultura tiene buenos resultados y acogida. Sin embargo, las reglas sólo son representadas bajo árboles, debido a la sencillez de su modelado y análisis. En este sentido, los árboles de decisión pueden aprovechar la estructura de representación de la información que proveen los grafos y viceversa. En un caso específico, en [48] es explorado el uso de árboles de decisión para encontrar patrones y grafos isomorfos.

De igual manera, es importante tener en cuenta dos enfoques presentados años atrás y que están relacionados con las temáticas abordadas. En primer lugar, en 1982, Hoffmann [49] propone un método para encontrar patrones dentro de estructuras de datos en forma de árbol. Por su lado, una década después, Oliver [50] presenta y propone un método para la construcción de *Grafos de Decisión*, como una generalización de los árboles de decisión que implementan mejoras en cuanto a la replicación y fragmentación. Estos enfoques no han sido replicados de forma considerable hasta la actualidad. Sin embargo, para el objetivo del presente trabajo de investigación brindan elementos de utilidad hacia la implementación de la solución al problema definido.

2.3. Brechas

A continuación, son presentados los aportes y brechas más relevantes en los trabajos relacionados.

Sección - Trabajos	**Aportes**	**Brechas**
Sistemas Expertos	Hacen uso de reglas construidas a partir del conocimiento de expertos. Los sistemas son construidos para un dominio de aplicación específico igual o cercano al de este trabajo. Utilizan técnicas de inteligencia artificial como lógica difusa y árboles de decisión. En [24] y [25] es considerada la información climática obtenida a través de sensores como parte de la base de conocimiento.	Sólo consideran los árboles de decisión como forma de representación de las reglas generadas por expertos. No es considerada una representación basada en grafos.
Árboles de Decisión	Usados para la detección de condiciones favorables para una enfermedad en distintos tipos de cultivos. Facilitan el análisis de la relación entre variables intervinientes en un problema. Expresan reglas definidas por expertos en un área específica. Es explorada la integración de árboles de decisión con grafos.	En estos trabajos es llevada a cabo la clasificación de un conjunto de datos a través de la verificación de reglas. Sin embargo, sólo hacen uso de árboles de decisión para este fin. De esta manera, no es considerada dicha clasificación a partir de la representación de reglas a través de grafos.
Emparejamiento de Patrones en Grafos	Son propuestas, técnicas para el mejoramiento de la eficiencia y calidad de los algoritmos de emparejamiento. Abordan el manejo de grandes cantidades de datos en grafos de tamaño considerable. Consideran dos tipos de emparejamiento: exacto y tolerante a errores.	El enfoque de estos trabajos está en el análisis de grandes cantidades de datos para encontrar patrones en ellos. A partir de esto, son generadas recomendaciones y extraídos subgrafos con características similares. Sin embargo, no es aplicada esta técnica dentro de sistemas expertos, ni como soporte a la validación de reglas en árboles de decisión.

Tabla 1. Brechas y aportes de los trabajos relacionados

La revisión de los trabajos relacionados soporta la identificación de los aportes y brechas hacia la propuesta que este proyecto quiere abordar, como puede verse en la Tabla 1. De esta manera, es identificado cómo los sistemas expertos pueden hacer uso de los árboles de decisión para expresar reglas generadas por personas con cierto grado de experiencia en un dominio de aplicación. La aplicación de estos sistemas en la agricultura tiene buenos resultados y acogida. Sin embargo, las reglas sólo son representadas bajo árboles, debido a la sencillez de su modelado y análisis. En este sentido, los árboles de decisión pueden aprovechar la estructura de representación de la información que proveen los grafos y viceversa.

De lo anteriormente dicho, este trabajo pretende generar una integración entre las áreas de conocimiento abordadas, aprovechando la existencia de expertos en el estudio de la roya en el café. Los estudios llevados a cabo por estos investigadores les han permitido definir una serie de condiciones y reglas que relacionan variables climáticas y características del cultivo, con la aparición y desarrollo de esta enfermedad. De esta manera, el emparejamiento de patrones puede ser usado dentro de un sistema experto, con el fin de validar las reglas mencionadas y así detectar condiciones favorables para la aparición de epidemias de la enfermedad.

2.4. Resumen

En este capítulo fueron presentados el estado del arte y los trabajos relacionados alrededor de los sistemas expertos en entornos agrícolas, minería de datos y análisis de grafos. A partir de esta revisión, fueron identificados aportes y brechas de las investigaciones, tomados como punto de partida del presente trabajo. La representación basada en grafos permite una mayor expresividad e interpretabilidad de la información que estos contienen. Precisamente, las condiciones que propician el desarrollo de epidemias de enfermedades en cultivos, que a su vez están definidas por expertos en el área, pueden ser expresadas como patrones de grafos, con el fin de aprovechar las ventajas mencionadas de esta estructura de información. En este sentido, la búsqueda de estos patrones dentro de un repositorio de grafos que contenga información sobre monitorización de cultivos, puede ser implementada a través de una técnica denominada “Emparejamiento de patrones en grafo”.

3 Caracterización del Sistema Experto

Este capítulo presenta el proceso seguido para la caracterización del Sistema Experto, considerando como caso de estudio la detección de condiciones favorables para la aparición de Roya en cultivos de café. Dentro del ciclo de desarrollo de esta enfermedad existen varios factores, de diferente naturaleza, que favorecen o desfavorecen su evolución. En este sentido, los árboles de decisión son una técnica de minería de datos que pueden proporcionar una comprensión de cómo estos factores (climáticos, propiedades de cultivo y manejo agronómico) condicionan una evolución de la Roya en un cultivo de café [26].

Teniendo en cuenta que el objetivo es extraer reglas que correlacionen las variables presentes en un ambiente de cultivo de café, caracterizadas de acuerdo al estudio de expertos en la Roya, con la evolución de la enfermedad en un periodo específico, el análisis de estas variables será tomado como un proceso de descubrimiento de conocimiento en bases de datos [51], siguiendo el modelo de proceso de minería de datos CRISP-DM [52]. Las fases del proceso pueden verse en la Figura 1.

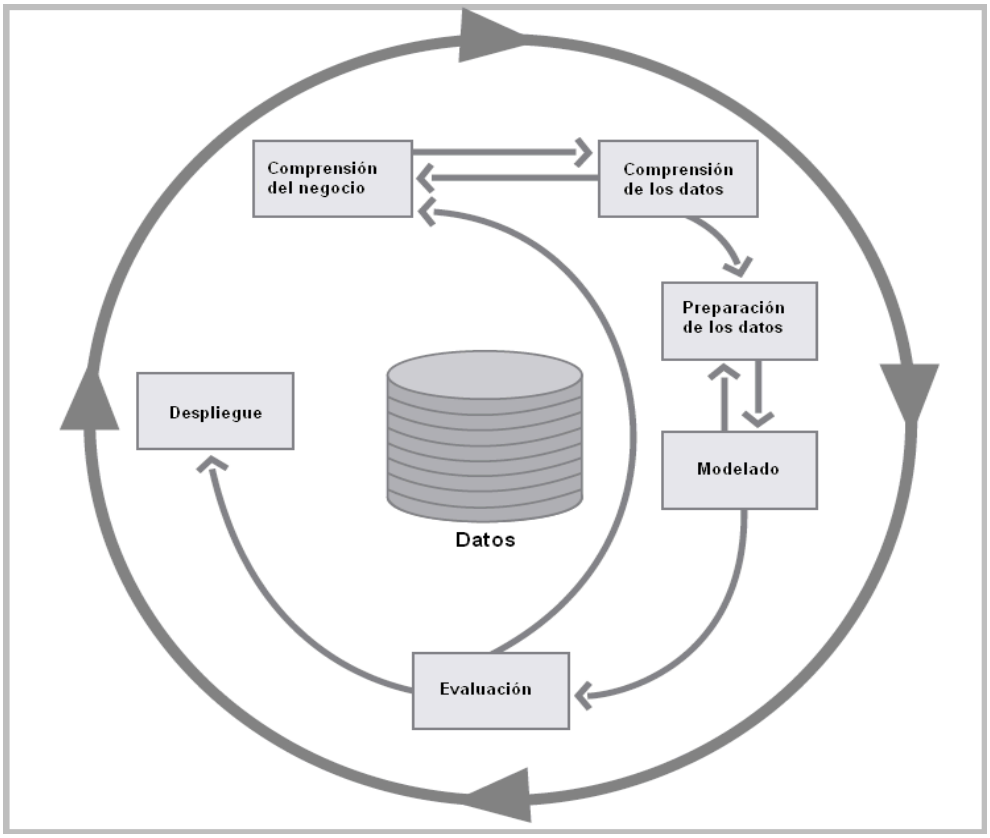


Figura 1. Fases CRISP-DM. Tomado de [52]

El proceso contiene seis fases que están relacionadas entre si por flechas que caracterizan las dependencias más importantes y frecuentes. Para la tarea de extracción de reglas (a

partir de inducción de árboles de decisión) que determinen las condiciones favorables para Roya en un cultivo de café, fueron consideradas las cinco primeras fases (Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado y Evaluación), y que son presentadas a continuación.

3.1. Comprensión del negocio

Esta fase inicial se centra en el entendimiento de los requerimientos de la solución que se quiere generar, para poder traducirlos en términos de un problema de minería de datos. A partir de esto, para el caso de estudio abordado en este trabajo, es necesario entender el ciclo de desarrollo de la Roya y los factores que inciden sobre él.

La Roya del cafeto es producida por el hongo *Hemileia Vastatrix.* Esta enfermedad ataca a cultivos de café alrededor del mundo y es responsable de pérdidas en la producción de hasta el 50%. Adicionalmente, la ocurrencia de la roya está íntimamente ligada al desarrollo fisiológico del cultivo, nivel de producción de la planta, manejo agronómico y a la distribución de algunas variables climáticas, como son la lluvia, humedad relativa y temperatura [53] [54].

El ciclo de desarrollo de esta enfermedad y los factores que lo afectan pueden verse en la Figura 2 [55].

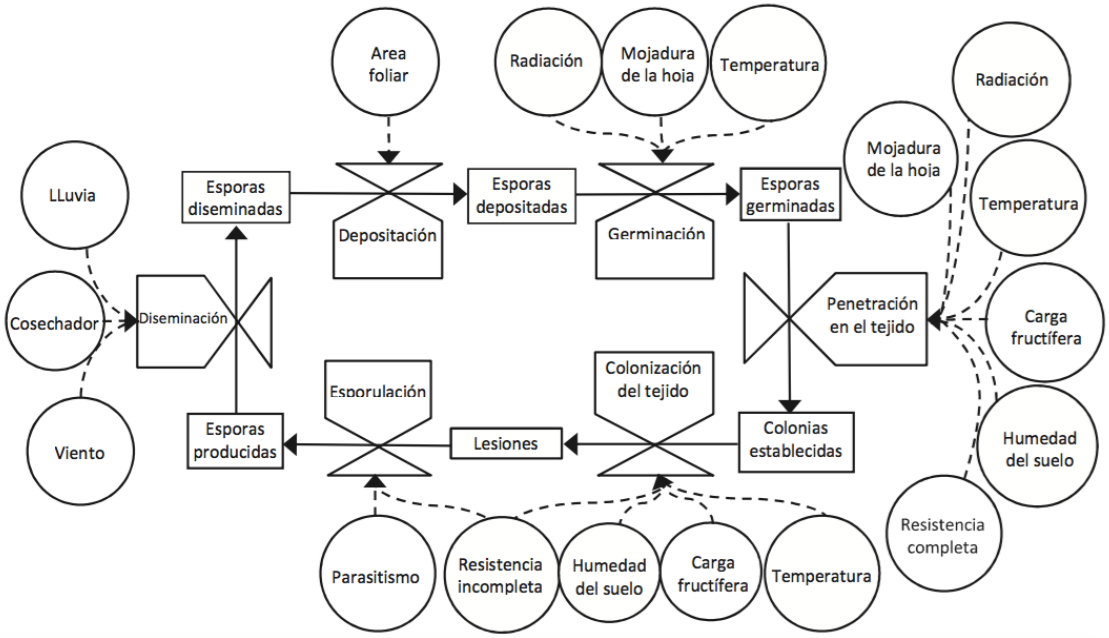


Figura 2. Diagrama de flujo del ciclo de vida de Hemileia Vastatrix (líneas continuas) y factores que lo afectan (líneas discontinuas) [55]

Este ciclo puede iniciar, arbitrariamente, a partir de la **diseminación**, en la que el hongo se libera y da paso a **depositarse** sobre la hoja. La **germinación** constituye la siguiente etapa.

Esta marca el inicio del proceso infeccioso en un sentido amplio. La **penetración** del hongo en el tejido de la hoja constituye el inicio de la tercera etapa, dando así comienzo a la infección en un sentido estricto. La **colonización del tejido** lleva a la formación de los primeros síntomas visuales. El periodo comprendido entre el inicio de la germinación y la expresión de los primeros síntomas (lesiones amarillentas) constituye el periodo de incubación. La producción y emergencia posterior de nuevas entidades infecciosas constituyen la etapa de la **esporulación**. El tiempo transcurrido entre el inicio de la germinación y la esporulación, el periodo de latencia, representa por lo mismo la variable de mayor importancia: cuanto más corto sea éste, será menor el tiempo para repetirse cada ciclo, llevando a que la epidemia presente mayor gravedad.

Adicionalmente, en la Figura 2 se puede apreciar que los factores que afectan cada parte del ciclo de la enfermedad pueden ser de diferentes tipos, como climáticos, propiedades del cultivo y manejo agronómico.

3.1.1. Factores Climáticos

Temperatura: Los investigadores de la enfermedad han generado varios valores óptimos de temperatura para el desarrollo de la roya. A continuación son presentados algunas consideraciones relacionadas a esta variable climática:

- Temperaturas entre 16 y 28°C [53].
- Temperaturas entre 21 y 25°C [54].
- Temperaturas entre 20 y 25°C, con mayor actividad en 22 °C [56]
- Temperaturas entre 22 y 24°C [57].
- Temperatura para etapa de germinación de 23.7°C [58].
- Variación de altas temperaturas (entre 22 y 28°C) y bajas temperaturas (entre 13 y 16°C) permite que la infección tenga lugar en menos de 6 horas [59].

Viento: El viento es un elemento requerido para la dispersión del hongo. Esta dispersión ocurre a inicios de la tarde cuando la humedad relativa es más baja y el viento más fuerte [60].

Lluvia: El hongo requiere de la salpicadura de la lluvia para iniciar su proceso de dispersión entre hojas y entre plantas, así como de la presencia de una capa de agua en el envés de las hojas para germinar [53] [56] [61] [54]. Sin embargo, si las lluvias son muy fuertes, las esporas pueden verse eliminadas por lavado [62]. Además, un día es considerado lluvioso cuando la precipitación acumulada del día es mayor o igual a 1mm [63].

Humedad Relativa: La humedad relativa tiene gran importancia para el estudio del ciclo del hongo de la Roya, debido a que es una medida indirecta de la **mojadura de la hoja.** La forma más utilizada para estimar el periodo de mojadura es por medio del número de horas con humedad relativa del aire por encima de un límite específico, generalmente 90% o 95%

[64] [65]. La germinación sólo ocurre si la hoja se encuentra mojada, de esta forma, una mojadura de la hoja prolongada (mínimo de 6 horas) es establecido como el tiempo mínimo necesario para que ocurra una infección [63]. Es importante considerar el cálculo del periodo de mojadura de la hoja para horas nocturnas y para todo el día. En el mismo sentido, una vez que la superficie de la hoja está mojada, la temperatura es el factor principal que determina el porcentaje de germinación de esporas y de penetración [63]. De esta forma, es necesario tener en cuenta el cálculo de temperaturas medias durante los periodos de mojadura de la hoja total y nocturno.

3.1.2. Propiedades del cultivo y manejo agronómico

Sombra: La sombra excesiva incrementa la intensidad de la infección [66] [56], ya que mantiene rangos de temperatura máxima y mínima muy estrechos y favorece una humedad relativa alta constante en las hojas del árbol [67]. Por otra parte, la sombra incide sobre otros factores involucrados en los procesos de la enfermedad, como la lluvia, viento, mojadura, carga fructífera y humedad del suelo [55].

Carga de frutos: Según varias investigaciones [68] [69] [70], la receptividad de las hojas a la Roya aumenta cuando la carga fructífera es más elevada. De esta forma, hay una asociación entre el avance de la cosecha y el avance de la epidemia.

Densidad: Cultivos con una densidad alta de árboles por sitio aumentan la competencia entre plantas por nutrientes, ofrecen una mayor interceptación de esporas y dificultan el cubrimiento de fungicidas sobre el follaje [53].

Aplicaciones de fungicidas: Estas aplicaciones deben hacerse de manera oportuna, con una correcta dosificación y demás especificaciones sugeridas para contrarrestar la enfermedad [53].

Fertilización: Una fertilización escasa o nula afecta principalmente a cafetales bajo plena exposición solar [53].

3.2. Comprensión de los datos

Esta fase comprende la recolección y entendimiento de las variables relacionadas en un conjunto de datos, identificando su relación con el problema a resolver. A partir de esta fase, fueron usados dos conjuntos de datos relacionados con monitorización y características de cultivos de café. De este modo, en cada fase existen dos subsecciones que describen el proceso llevado a cabo para cada uno de los conjuntos de datos.

3.2.1. Conjunto de datos Varginha, Brasil

El primer conjunto de datos usado fue obtenido a través de una estación meteorológica ubicada en una granja experimental de Varginha, Minas Gerais, Brasil, y la información del cultivo de café en esta zona. Estos datos fueron proporcionados por la Fundación Procafé[1] y se encuentran preparados según las consideraciones dadas por Meira et al. [26] , en escala de tiempo mensual, comprendiendo el periodo entre noviembre de 1988 y diciembre de 2014.

3.2.2. Conjunto de datos Los Naranjos, Colombia

En segundo lugar, fue usado un conjunto de datos obtenido por Corrales et al. [71] en la granja experimental *Los Naranjos*, perteneciente a la empresa Supracafé, la cual está ubicada en el municipio de Cajibío (Cauca). Este conjunto de datos, al que llamaremos *conjunto base*, contiene información de diferentes lotes de la granja, como datos climáticos (obtenidos por una estación meteorológica), estado de cultivo, propiedades de cultivo y medidas de incidencia de roya para varios meses entre el año 2011 y 2013, generando 147 instancias. Los atributos del conjunto de datos puede verse en la Tabla 2.

Factores	**Atributos**	**Tipo**
Condiciones Climáticas	Humedad relativa media en los últimos 2 meses.	Numérico
	Número de horas Humedad Relativa > 90% en el último mes	Numérico
	Promedio de Amplitud térmica en el último mes.	Numérico
	Número de días con lluvia en el último mes.	Numérico
	Lluvia acumulada últimos 2 meses.	Numérico
	Lluvia nocturna acumulada último mes.	Numérico
Estado del suelo	pH	Numérico
	Materia Orgánica	Numérico
	K	Numérico
	Ca	Numérico
	Arcilla	Numérico
Propiedades cultivo	Variedad	Nominal
	Densidad de plantas por hectárea	Numérico
	Distancia entre plantas	Numérico
	Distancia entre surco	Numérico
	Edad del cultivo	Numérico
	Sombrío	Numérico
Control del cultivo	Control de roya en el mes anterior	Nominal
	Control de roya en los tres meses anteriores	Nominal
	Fertilización en los últimos cuatro meses	Nominal
	Producción acumulada en los dos últimos dos meses	Numérico
Condición de Roya	Porcentaje de incidencia de Roya	Numérico

Tabla 2. Atributos del conjunto de datos. Granja experimental Los Naranjos [71]

[1] fundacaoprocafe.com.br/

Considerando que el objetivo es generar una serie de reglas para la ocurrencia de Roya en los cultivos, la información sobre el porcentaje de incidencia contenida en el conjunto de datos es la variable más relevante. Para obtener este valor, en Colombia, el Centro Nacional de Investigación del Café (Cenicafé) ha desarrollado la siguiente metodología basada en la exploración de un lote en un área igual o menor a una hectárea [53]:

1) Primero, la persona debe ubicarse en el centro del primer surco, allí selecciona un árbol y en él, escoge la rama con mayor follaje en cada uno de los niveles del árbol (bajo, medio y alto); en cada rama cuenta el número total de hojas y el número de éstas afectadas por la roya.

2) Posteriormente, la persona recorre el área del lote entre surcos, por el centro de los mismos, y en cada surco selecciona un árbol, hasta completar 60 árboles por lote, recorriendo todos los surcos. Es decir, si el lote tiene 60 surcos, evalúa un árbol por surco; si el lote tiene 30 surcos, evalúa dos árboles por surco; y, si el lote tiene 120 surcos, evalúa un árbol cada dos surcos.

3) Al finalizar el recorrido, se suma el total de hojas y el número de hojas afectadas por roya de los 60 árboles, y este valor se multiplica por cien, el cual corresponde al porcentaje de hojas afectadas por roya en el lote, como se puede observar en la Ecuación 1:

$$infección\ en\ el\ lote\ (\%) = \frac{Total\ de\ hojas\ con\ roya\ en\ los\ 60\ árboles}{Total\ de\ hojas\ presentes\ en\ los\ 60\ árboles} \times 100$$

Ecuación 1. Cálculo del porcentaje de incidencia de Roya

3.3. Preparación de los datos

Esta fase comprende las actividades necesarias para la construcción del conjunto de datos final, el cual será usado en la fase de modelado. Entre las actividades ejecutadas están la selección de atributos predictivos y variable dependiente, y transformación y limpieza de los datos.

3.3.1. Conjunto de datos Varginha, Brasil

Para este conjunto de datos, la variable dependiente (atributo clase) fue obtenida a partir del análisis del comportamiento de la enfermedad, según el valor de la incidencia de la Roya cada mes. De esta manera, la tasa de infección es calculada restando la incidencia del mes analizado a la incidencia del mes anterior, obteniendo tres clases o categorías:

- TX1 (>0<=5): crecimiento moderado, para tasas de infección positivas, menores o iguales a 5 puntos porcentuales (pp).
- TX2 (>5): crecimiento acelerado, para tasas de infección mayores a 5 pp.

Estas tasas fueron escogidas siguiendo las recomendaciones hechas por Campos et al. [72] para el control de la enfermedad.

Por su parte, los atributos predictivos (variables independientes) corresponden a la transformación de los valores de variables meteorológicas obtenidas por la estación a un nivel de dato por hora, en datos diarios y, finalmente, en escala mensual. Adicionalmente, fueron aplicadas varias consideraciones dadas por expertos que han estudiado el desarrollo de la enfermedad en Brasil. Por ejemplo, Moraes et al. [73] observaron que el periodo de incubación del hongo tiende a volverse más corto en los meses más calientes (28 días) y es más largo en los meses más fríos (65 días). Tras estas observaciones, fue sugerida la Ecuación 2 para la estimación del periodo de incubación, donde x_1 y x_2 representan la temperatura media máxima y mínima, respectivamente, durante un intervalo analizado y *PI* representa el periodo de incubación en días.

$$PI = 103.01 - 0.98x_1 - 2.1x_2$$

Ecuación 2. Cálculo de periodo de incubación

Con el fin de obtener la relación entre las condiciones diarias de mojadura de la hoja y temperatura (durante periodo de mojadura), con respecto a una infección de Roya, fue generada una clasificación basada en los trabajos de Montoya y Chaves [58] y Kushalappa et al. [63], la cual puede verse en la Tabla 3.

HORHR90	**Temperatura media durante horas con Humedad Relativa >= 90%**				
	T < 18	18 ≤ T < 21	21 ≤ T ≤ 24	24 < T ≤ 28	T > 28
HORHR90 < 6	Desfavorable	Desfavorable	Desfavorable	Desfavorable	Desfavorable
6 ≤ HORHR90 < 12	Desfavorable	Desfavorable	Favorable	Desfavorable	Desfavorable
HORHR90 >= 12	Desfavorable	Favorable	Favorable	Favorable	Desfavorable

Tabla 3. Matriz de favorabilidad diaria de germinación del hongo

Esta tabla permite establecer la favorabilidad de un día determinado con relación a la germinación del hongo, donde HORHR90 corresponde al número de horas en el día con Humedad Relativa >= 90% y T corresponde a la temperatura media en dichas horas.

Ahora bien, para la preparación de los datos, cada día fue tratado como un eventual día de infección y, considerando un periodo de incubación estimado, ese día fue asociado al mes correspondiente de evaluación de incidencia de Roya [26]. En la Figura 3 puede verse la representación del esquema mencionado, donde D_i representa un día de infección, E_i es la

evaluación de la incidencia de Roya, E_{i-1} la evaluación de la incidencia en el mes anterior, PI es el periodo de incubación y PINF el periodo de infección.

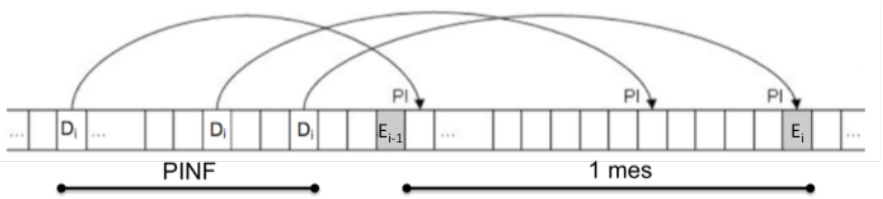


Figura 3. Esquema para preparación de los datos de entrenamiento según periodos de infección e incubación. Brasil [26]

Como se ve en la Figura 3, cada día fue asociado a una tasa de infección, para la cual posiblemente tenga una cuota de contribución. Ahora bien, el conjunto de días asociado a una tasa de infección fue denominado periodo de infección (PINF).

A partir de las consideraciones descritas anteriormente, fue obtenido el conjunto de datos de 193 instancias descrito en la Tabla 4.

Atributo	Descripción	Tipo	Unidad
TINF*	Tasa de infección de Roya. {0(TI1), 1(TI2)}	Nominal	-
TMAX_PINF	Media de temperaturas máximas diarias en PINF	Numérico	°C
TMED_PINF	Media de temperaturas medias diarias en PINF	Numérico	°C
TMIN_PINF	Media de temperaturas mínimas diarias en PINF	Numérico	°C
TMAX_PI_PINF	Media de temperaturas máximas diarias en PI para los días de PINF	Numérico	°C
TMED_PI_PINF	Media de temperaturas medias diarias en PI para los días de PINF	Numérico	°C
TMIN_PI_PINF	Media de temperaturas mínimas diarias en PI para los días de PINF	Numérico	°C
HR_PINF	Media de Humedad Relativa media diaria en PINF	Numérico	%
HORHR90_PINF	Media de número de horas diarias con Humedad Relativa >= 90% en PINF	Numérico	horas
SUMHR90_PINF	Sumatorio de número de horas con Humedad Relativa >= 90% en PINF	Numérico	horas
T_HR90_PINF	Media de temperatura media diaria durante horas con Humedad Relativa >= 90% en PINF	Numérico	°C

PRE_MED_PINF	Media de precipitaciones medias diarias en PINF	Numérico	mm
PRE_ACUM_PINF	Media de precipitación acumulada diaria en PINF	Numérico	mm
DLLUV_PINF	Número de días lluviosos (precipitación acumulada >= 1mm) durante PINF	Numérico	mm
DDI_PINF**	Número de días desfavorables para infección en PINF	Numérico	días
DFI_PINF**	Número de días favorables para infección en PINF	Numérico	días

* Variable dependiente
** Según tabla X

Tabla 4. Variables del conjunto de datos Varginha, Brasil

3.3.2. Conjunto de datos Los Naranjos, Colombia

Con base en la información mencionada anteriormente y a partir del conocimiento de expertos en el estudio de la Roya, descrito en la sección 3.1, fue construido un nuevo conjunto de datos con el fin de generar una serie de reglas que describan condiciones favorables para el desarrollo de esta enfermedad.

La variable dependiente (atributo clase) fue obtenida a partir del análisis del comportamiento de la enfermedad, según el valor de la incidencia de la Roya existente en el conjunto de datos entre meses consecutivos. De esta manera, la tasa de infección es calculada evaluando el aumento o disminución del porcentaje de incidencia entre el mes analizado y el mes siguiente, obteniendo tres clases o categorías:

- TI1 (<=0): reducción o latencia, para tasas de infección negativas o nulas.
- TI2 (>0<=2): crecimiento moderado, para tasas de infección positivas, menores o iguales a 2 puntos porcentuales (pp).
- TI3 (>2): crecimiento acelerado, para tasas de infección mayores a 2 pp.

Por otro lado, los atributos predictivos (variables independientes) basados en información meteorológica, fueron construidos a partir de los registros de la estación ubicada en la granja, la cual generaba un registro por minuto. Considerando que las instancias del conjunto base están expresadas en una escala mensual, es necesario convertir los registros de la estación, pasando desde un dato por hora, dato por día, hasta obtener un dato por mes. Los datos en cada escala son usados para analizar la temperatura, precipitación, viento y humedad relativa, conforme la relación que los expertos en Roya han definido entre estas variables climáticas y la enfermedad (Sección 3.1.1). Asimismo, las variables del conjunto base que se tuvieron en cuenta fueron densidad, sombra y porcentaje de infección.

En consecuencia, fue obtenido un conjunto de datos con 124 instancias, que define unas variables a partir de las consideraciones dadas por expertos en la enfermedad,

relacionando su valor en un mes dado, con el desarrollo de la enfermedad entre dicho mes y el mes siguiente (variable dependiente TINF). Lo anterior puede verse en la Tabla 5.

Atributo	Descripción	Tipo	Unidad
TINF*	Tasa de infección de Roya. {0(TI1), 1(TI2), 2(TI3)}	Nominal	-
DENSIDAD	Densidad del lote. {3008, 4016, 5013, 6993}	Nominal	-
SOMBRA	Porcentaje de sombrío	Numérico	%
DLLUV	Número de días lluviosos (Precipitación >= 1mm)	Numérico	días
PRE_MED	Media de precipitaciones medias diarias	Numérico	mm
PRE_ACUM	Media de precipitación acumulada diaria	Numérico	mm
HORHR90	Media de número de horas diarias con Humedad Relativa >= 90%	Numérico	horas
HORHRN90	Media de número de horas nocturnas diarias con Humedad Relativa >= 90%	Numérico	horas
SUMHR90	Sumatorio de número de horas con Humedad Relativa >= 90%	Numérico	horas
SUMHRN90	Sumatorio de número de horas nocturnas con Humedad Relativa >= 90%	Numérico	horas
HR	Media de Humedad Relativa media diaria	Numérico	%
T_HR90	Media de temperatura media diaria durante horas con Humedad Relativa >= 90%	Numérico	°C
T_HRN90	Media de temperatura media diaria durante horas nocturnas con Humedad Relativa >= 90%	Numérico	°C
TMAX	Media de temperaturas máximas diarias	Numérico	°C
TMED	Media de temperaturas medias diarias	Numérico	°C
TMIN	Media de temperaturas mínimas diarias	Numérico	°C
VVIENTO	Media de velocidad de viento media diaria	Numérico	m/s

*Variable dependiente

Tabla 5. Variables del conjunto de datos Los Naranjos, Colombia

Ahora bien, dentro de las 124 instancias, la distribución de la variable dependiente TINF está dada por 42 registros para TI1, 50 para TI2 y 32 para TI3. Estos valores muestran una diferencia entre registros de cada clase que puede afectar la etapa de modelado. Para

realizar un balanceo de clases, fue usado el algoritmo el algoritmo SMOTE [74], a través del software Weka (Waikato Environment for Knowledge Analysis - Entorno para Análisis del Conocimiento de la Universidad de Waikato) [75], el cual genera nuevas instancias sintéticas o artificiales para equilibrar la muestra de datos basado en la regla del vecino más cercano. Para esto, primero es determinado el número *k* de vecinos y una cantidad aleatoria de ejemplos guardadas en un conjunto *A*. Posteriormente, para cada elemento de *A* se encuentran los *k* vecinos más cercanos por medio del cálculo de la distancia euclidiana entre. A partir del cálculo de estos vecinos, SMOTE construye las nuevas instancias sintéticas. Después de aplicar este algoritmo, con una configuración de *k*=5 (configuración típica), el número de instancias fue 161 y el número de registros para cada clase de TINF fue 63 para TI1, 50 para TI2 y 48 para TI3.

3.4. Modelado

La generación de los modelos de desarrollo de la Roya fue abordada a través de la aplicación de un algoritmo básico para inducción de árboles de decisión. El algoritmo construye un árbol de decisión de forma recursiva, a partir de un conjunto de datos compuesto por una serie de instancias. En cada instancia se especifican valores para una colección de atributos y para una clase. De esta manera, el proceso se basa en relacionar los valores de los atributos que tengan un valor de clase similar, con el fin de generar una serie de condiciones que deben cumplir los atributos para ser clasificados en una clase específica. El rendimiento del árbol de decisión es obtenido a partir de un conjunto de prueba definido de manera aleatoria, donde un porcentaje de error en la clasificación basado en el porcentaje de instancias clasificadas erróneamente es calculado.

La inducción del árbol de decisión fue hecha por medio del software Weka, versión 3.6. Weka es una plataforma de libre distribución basada en java creada por la Universidad de Waikato bajo licencia GNU-GPL, que implementa un conjunto de algoritmos de aprendizaje automático [75]. El algoritmo escogido fue el C4.5 [76], debido a que éste construye un modelo de fácil comprensión a través de la creación una serie de nodos etiquetados que contienen las características que deben ser evaluadas en cada paso. Este enfoque puede ser tomado como una serie de reglas que caracterizan las clases definidas en el modelo. El proceso llevado a cabo por el algoritmo inicia con la selección de la variable o atributo a partir de la cual se va a dividir el conjunto de datos de entrenamiento (nodo raíz), buscando que en los subconjuntos generados haya una mínima variabilidad respecto a la clase. Este proceso es recursivo, es decir, una vez que se haya determinado la variable con la que se obtiene la mayor homogeneidad respecto a la clase en los nodos hijos, se vuelve a realizar el análisis para cada uno de los nodos hijos. Adicionalmente, son implementados métodos de pre-poda y post-poda de los árboles, basados en una evaluación de si es justificada o no la expansión de una rama o evaluación [77], para posteriormente convertir cada rama en una regla [78].

Ante las anteriores consideraciones, fue aplicado el algoritmo C4.5 a través de la implementación J48 contenida Weka, con un mínimo de instancias por hoja de 2 y un factor de confianza de 0.25. Este factor es usado para el proceso de poda, donde los valores menores implican una poda mayor.

3.5. Evaluación

Los modelos de clasificación necesitan ser evaluados, con el fin de generar estadísticas para la validación y entendimiento de los resultados obtenidos. Para esto, existen varias medidas que describen el desempeño del algoritmo de clasificación:

- **Validación cruzada:** dado un número n (orden de la validación), los datos son divididos en n partes y, por cada parte, es construído el clasificador con las n−1 partes restantes (conjunto de entrenamiento), evaluando el clasificador obtenido con la parte no contenida en el conjunto de entrenamiento(conjunto de prueba). La tasa de error resultante es derivado como la media de las tasas de error obtenidas en cada una de las iteraciones (n).

- **Estadística Kappa:** este índice representa el porcentaje de acuerdo entre observadores. Los posibles valores van desde un rango de +1 (acuerdo o acercamiento perfecto), 0 (ningún acuerdo por encima de lo esperado) y -1 (total desacuerdo) [79].

- **Error Absoluto Medio (EAM):** es la medida de la cercanía que hay entre una predicción y el valor real de un conjunto de datos [80], definido como:

$$EAM = \frac{1}{n}\sum_{i=1}^{n}|p_i - v_i|$$

Ecuación 3. Error absoluto medio

Donde, p_i es la predicción, v_i el valor real y n el número de datos.

- **Error Cuadrático Medio (ECM):** mide la diferencia entre lo pronosticado y los correspondientes valores observados, mediante la cuadrática para que luego sea promediado a lo largo de la muestra [81]. Está definido como:

$$ECM = \frac{\sqrt{\sum_{i=1}^{n}(p_i - v_i)^2}}{n}$$

Ecuación 4. Error cuadrático medio

Donde, p_i es la predicción, v_i el valor real y n el número de datos.

- **Error Absoluto Relativo (EAR):** representa el porcentaje de error de predicción de un clasificador [81]. Éste es calculado de la siguiente forma:

$$EAR = \frac{\sum_{i=1}^{n}|p_i - v_i|}{\sum_{i=1}^{n}|v_i - \bar{v}|}$$

Ecuación 5. Error absoluto relativo

Donde, p_i es la predicción, v_i el valor real, $\bar{v}$ el valor real promedio y n el número de datos.

- **Matriz de confusión:** es una herramienta de visualización usada para la evaluación del modelo con base en cada clase, donde cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Cada elemento de la matriz (M) puede ser calculando usando la Ecuación 6, donde i,j = 1,2,…,k; {C_1, C_2,…,C_k} es un conjunto de las clases para el atributo meta; (x,y) son los ejemplos del conjunto de entrenamiento T; y I(a,b) es igual a 1 si a es igual a b, o es igual a 0 en caso contrario. La diagonal principal de la matriz (elementos (i,j) donde i = j) representa los aciertos del modelo, mientras los demás elementos representan los errores relacionados a cada clase.

$$M(C_i, C_j) = \sum_{\{\forall(x,y)\in T: y=C_i\}} I(h(x), C_j)$$

Ecuación 6. Cálculo de elementos para la matriz de confusión

De esta forma, a partir de la definición de la matriz de confusión, pueden ser obtenidas las medidas que son descritas a continuación.

- **Falsos Positivos (FP):** instancias incorrectamente clasificadas en la clase C_k.

- **Falsos Negativos (FN):** instancias de la clase C_k que fueron incorrectamente clasificadas en otra clase.

- **Verdaderos Positivos (VP):** instancias correctamente clasificadas en la clase C_k.

- **Verdaderos Negativos (VN):** todas las instancias restantes correctamente clasificadas diferente a la clase C_k.

- **Tasa de Verdaderos Positivos (TVP):** es la medida encargada de calcular la proporción de verdaderos positivos predichos entre todos los positivos [82]. Esta tasa es calculada con la siguiente ecuación:

$$TVP = \frac{|VP|}{|VP + FN|}$$

Ecuación 7. Cálculo de tasa de verdaderos positivos

- **Tasa de Falsos Positivos (TFP):** es la proporción de elementos que no clasifican en la clase c_k, de entre todos los elementos que realmente son de la clase c_k .Esta tasa es calculada con la siguiente ecuación:

$$TFP = \frac{|FP|}{|VP + FP|}$$

Ecuación 8. Cálculo de tasa de falsos positivos

- **Precisión:** es la cantidad de ejemplos que realmente tienen la clase c_k entre todos los elementos clasificados dentro de la clase c_k (capacidad del clasificador para evitar el ruido) [83]. La precisión es calculada con base en la siguiente ecuación:

$$p = \frac{|VP|}{|VP + FP|}$$

Ecuación 9. Cálculo de precisión

- **Exhaustividad:** esta medida, conocida en inglés como “recall”, es calculada de la misma forma que la tasa de verdaderos positivos (ver Ecuación 7).

- **Medida F:** es una medida que relaciona la precisión con la exhaustividad, donde su mejor valor es 1 y el peor es 0 [84]. El cálculo de esta medida está dado por la Ecuación 10.

$$F = 2 \times \frac{p \times e}{p + e}$$

Ecuación 10. Cálculo de medida F

3.6. Resultados

A continuación, son resumidos los resultados para cada conjunto de datos, a partir de las consideraciones hechas para cada fase de la metodología CRISP-DM mencionadas.

3.6.1. Conjunto de datos Varginha, Brasil

Para este conjunto de datos, el árbol de decisión resultante puede verse en la Figura 4. Árbol de decisión generado para el conjunto de datos Varginha, Brasil. En esta figura, los círculos representan los nodos donde son evaluadas los atributos predictivos y los recuadros color gris representan las clases predichas. Siendo así, de los 15 atributos predictivos que se encuentran en el conjunto de datos, el algoritmo utilizado sólo relaciona

4 de estos en el modelo generado. Esto indica que los atributos que tienen mayor impacto en el aumento o disminución de la infección de la enfermedad son: DDI_PINF, TMED_PI_PINF, HR_PINF y DFI_PINF.

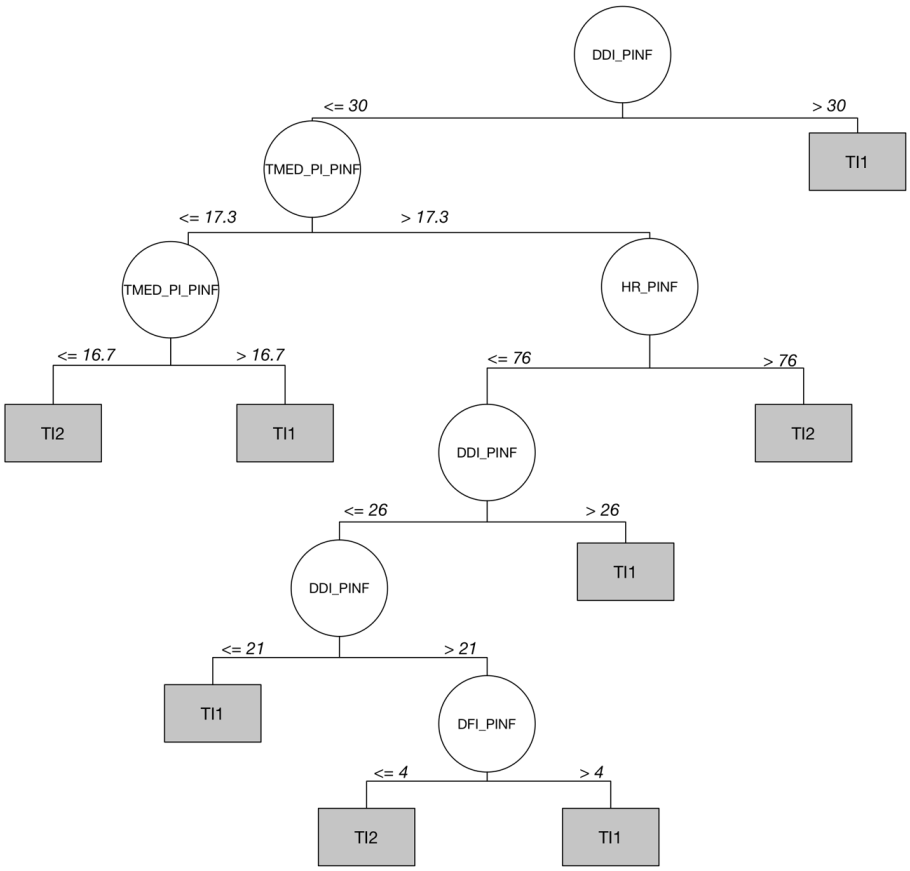


Figura 4. Árbol de decisión generado para el conjunto de datos Varginha, Brasil

Para el análisis del árbol de decisión generado, se debe partir del nodo raíz del mismo donde es evaluada la variable DDI_PINF. De esta manera, en el caso de que el número de días desfavorables para una infección en PINF es mayor a 30, la tasa predicha es T1. Para el caso de que DDI_PINF sea menor o igual a 30 días, la siguiente variable a ser evaluada es TMED_PI_PINF. Para temperaturas medias diarias en el periodo de incubación que corresponden a un PINF, si este valor es menor o igual a 17.3 °C, esta variable vuelve a ser evaluada para determinar una de las dos tasas. Teniendo en cuenta esto, puede interpretarse que si TMED_PI_PINF es menor o igual a 16.7 °C, se tiene como resultado la tasa TI2. Por el contrario, si esta variable está entre 16.7 °C y 17.3 °C, la tasa predicha es TI1. Lo anterior, permite concluir que para los días de incubación relacionados con un periodo de infección, las temperaturas medias diarias bajas generan un crecimiento acelerado de la infección de roya, mientras que para el rango entre 16.7 °C y 17.3 °C, el crecimiento es moderado. Ahora bien, para el caso que TMED_PI_PINF sea mayor a 17.3 °C, la siguiente variable a ser evaluada es HR_PINF. En el caso de que la media de

humedad relativa en PINF esté por encima de un 76%, el modelo predice un crecimiento acelerado de la infección (TI2). Cuando HR_PINF sea igual o menor a 76%, la variable a ser analizada de forma seguida es el número de días desfavorables para infección en PINF. Cuando esta variable sea mayor a los 26 días, se tiene una tasa TI1. Igualmente, para valores menores o iguales a 21 días, la misma tasa es predicha. En contraste, para valores entre 21 y 26 días, deben evaluarse los días favorables para infección en PINF, donde es obtenida una tasa TI1 para valores mayores a 4 días y una tasa TI2 para valores menores o iguales a 4 días.

De esta forma, el árbol de decisión es recorrido siguiendo las distintas evaluaciones de las variables en cada nodo, con el fin de concluir un valor de la tasa de infección (TI1 o TI2).

Adicionalmente, las medidas de desempeño del algoritmo de clasificación, con una validación cruzada de orden 10, son:

- Instancias clasificadas correctamente: 153 (79.2746 %)
- Instancias clasificadas incorrectamente: 40 (20.7254 %)
- Estadística Kappa: 0.589
- Error absoluto medio: 0.2618
- Error cuadrático medio: 0.4067
- Error absoluto relativo: 52.5299 %

La matriz de confusión que evalúa el modelo generado es la siguiente (Tabla 6):

Clases	Clasificadas como	
	TI1	TI2
TI1 (102 instancias)	72	30
TI2 (91 instancias)	10	81

Tabla 6. Matriz de confusión. Árbol de decisión Varginha, Brasil

A partir de esta matriz son extraídas medidas de evaluación para cada clase, como son (Tabla 7):

Clase	Medida				
	TVP	TFP	Precisión	Exhaustividad	Medida F
TI1	0.71	0.11	0.88	0.71	0.79
TI2	0.89	0.29	0.73	0.89	0.79

Tabla 7. Medidas de evaluación. Árbol de decisión Varginha, Brasil

Los resultados obtenidos se encuentran en un grado aceptable de instancias clasificadas correctamente por el modelo generado. Como puede verse en la matriz de confusión, la clase que presenta mayores instancias clasificadas incorrectamente es TI2. De esta manera, la precisión de la clase TI1 es más confiable que la de TI2. La menor precisión presente en TI2 puede deberse al número mayor de instancias contenidas en el conjunto de entrenamiento. Por otro lado, en la Figura 4, puede verse como es predicha la clase TI1 basada únicamente en el análisis de la variable DDI_PINF (mayor a 30), lo cual puede generar un mayor número de instancias clasificadas incorrectamente, al tener una sola restricción en los valores dentro de la instancia.

3.6.2. Conjunto de datos Los Naranjos, Colombia

El árbol de decisión generado para este conjunto de datos puede verse en la Figura 5.

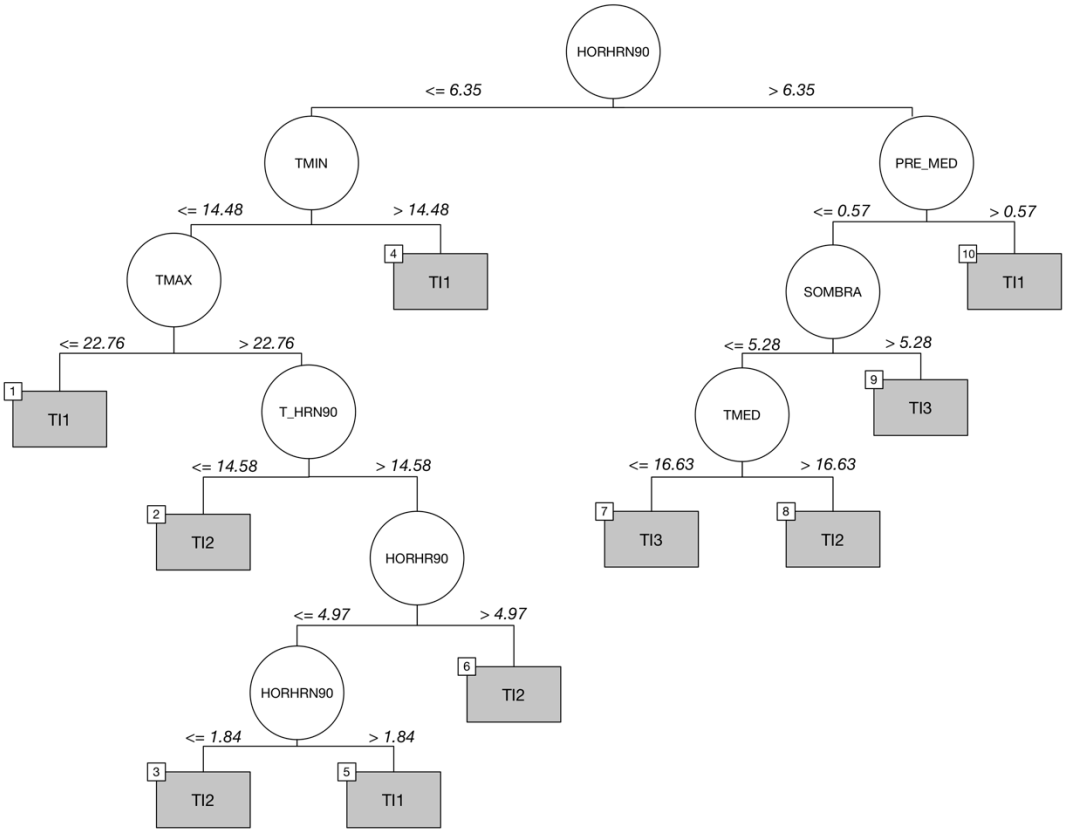


Figura 5. Árbol de decisión generado para el conjunto de datos Los Naranjos, Colombia

Este modelo representa una ayuda para comprender cómo las variables contenidas en el conjunto de datos se relacionan con las tasas de infección de roya en el café. Los círculos representan los nodos donde son evaluadas los atributos predictivos y los recuadros color gris representan las clases predichas. Siendo así, de los 16 atributos predictivos que se

encuentran en el conjunto de datos, el algoritmo utilizado sólo relaciona 8 de estos en el modelo generado.

Para el análisis del árbol de decisión generado, se debe partir del nodo raíz del mismo donde es evaluada la variable HORHRN90. Los horas en las cuales la humedad relativa es mayor a 90%(horas de mojadura de la hoja) propician un ambiente favorable para la germinación del hongo. En este caso, el modelo le da una prioridad a la media mensual de las horas nocturnas con esta condición, estableciendo dos rangos: media de número de horas nocturnas de mojadura de la hoja mayor a 6.35 y, por otro lado, menor o igual a 6.35; lo cual está relacionado con el criterio de número de horas con mojadura de la hoja mínimas, necesarias para una infección [63]. Para el caso en que HORHRN90 esté por debajo o sea igual a 6.35, la variable siguiente a ser analizada es TMIN(media de temperaturas mínimas), para la cual, si se encuentra por encima de 14.48 °C, la clase predicha es TI1: y, por otro lado, si es menor o igual a 14.48 °C se debe seguir con la evaluación de TMAX(media de temperaturas máximas). En caso de que TMAX sea menor o igual a 22.76 °C, la clase predicha es TI1, y si es mayor a ese valor, se analiza T_HRN90(media de temperaturas en horas nocturnas de humedad relativa mayor a 90%). La variable T_HRN90, sugerida por [63], lleva a la predicción de la clase TI2 cuando ésta es igual o menor a 14.58 °C y, por otro lado, cuando es mayor a este valor se debe pasar a un análisis de HORHR90 (media de horas diarias de humedad relativa mayor a 90 %). En la evaluación de HORHR90, si su valor está por encima de 4.97, la clase predicha es TI2; y, si el valor es menor o igual a 4.97 debe volverse a presentar un análisis de la variable ubicada en el nodo raíz del árbol, HORHRN90. En este caso, si es mayor a 1.84, la clase predicha es TI1, que podría expresarse como TI1 mayor a 1.84 y menor a igual a 6.35 (recordando la evaluación del nodo raíz). En contraste, si es menor o igual a 1.84, la clase predicha es TI2.

Ahora bien, volviendo al nodo raíz, en caso de que HORHRN90 sea mayor a 6.35, la variable siguiente a evaluar es PRE_MED (media de precipitaciones medias diarias). Si PRE_MED es mayor a 0.57 mm, la clase predicha es TI1, en caso contrario, si es menor o igual a ese valor, debe analizarse la variable SOMBRA. En caso de que el cultivo tenga un porcentaje de sombra mayor a 5.28 %, la clase predicha es TI3 y, para cultivos con sombra menor, debe ser evaluada la variable TMED (media de temperaturas medias diarias). Para TMED mayor a 16.63 °C, la clase predicha es TI2 y para TMED igual o menor a ese valor, la clase predicha es TI3.

En este sentido, el recorrido desde el nodo raíz hasta cada una de las clases predichas puede tomarse como un conjunto de condiciones o patrones que deben cumplir las variables analizadas, con el fin de poder predecir una clase en particular.

Adicionalmente, las medidas de desempeño del algoritmo de clasificación, con una validación cruzada de orden 10, son:

- Instancias clasificadas correctamente: 131 (81.3665 %)

- Instancias clasificadas incorrectamente: 30 (18.6335 %)
- Estadística Kappa: 0.7201
- Error absoluto medio: 0.1638
- Error cuadrático medio: 0.3232
- Error absoluto relativo: 37.1271 %

La matriz de confusión que evalúa el modelo generado es la siguiente (Tabla 8):

Clases	Clasificadas como		
	TI1	TI2	TI3
TI1 (63 instancias)	53	6	4
TI2 (50 instancias)	3	33	14
TI3 (48 instancias)	1	2	45

Tabla 8. Matriz de confusión. Árbol de decisión Los Naranjos, Colombia

A partir de esta matriz son extraídas medidas de evaluación para cada clase, como son (Tabla 9):

Clase	Medida				
	TVP	TFP	Precisión	Exhaustividad	Medida F
TI1	0.841	0.041	0.93	0.841	0.883
TI2	0.66	0.072	0.805	0.66	0.725
TI3	0.938	0.159	0.714	0.938	0.811

Tabla 9. Medidas de evaluación. Árbol de decisión Los Naranjos, Colombia

Los resultados de la evaluación del modelo generado presentan un buen porcentaje de instancias clasificadas correctamente. Sin embargo, las medidas de evaluación discriminadas por cada clase permiten determinar que la clase TI3 presenta una mayor proporción de errores que las otras clases. En la matriz de confusión puede verse cómo existe un gran número de instancias pertenecientes a TI2(50) fueron clasificadas como TI3(14), lo cual podría ser ocasionado por una proximidad fuerte entre las reglas que predicen cada una de estas clases. Asimismo, la clase TI1 es la que presenta menos instancias clasificadas incorrectamente dentro de otra clase, teniendo en cuenta que esta clase es la que menor número de instancias posee.

3.7. Comparación de resultados de los conjuntos de datos

Las medidas de evaluación, extraídas para el algoritmo de clasificación aplicado en los dos conjuntos de datos, permiten realizar una comparación de los modelos generados. Los resultados de estas pruebas generaron un porcentaje aceptable de instancias clasificadas correctamente. Sin embargo, se debe tener en cuenta que éstas medidas son obtenidas a partir de un proceso de validación cruzada, donde son tomadas de forma aleatoria algunas instancias de prueba a las cuales es aplicado el modelo generado, con el fin de garantizar que los resultados son independientes de la partición entre datos de entrenamiento y prueba. Esto puede conducir a estimaciones de rendimiento irrealmente optimistas, pero no representa el rendimiento de los nuevos datos que se pueden encontrar cuando el modelo se aplica en la práctica.

De esta manera, el modelo obtenido para el conjunto de datos de Los Naranjos, Colombia, presenta unas mejores medidas de desempeño. Esto puede deberse a que en el modelo para Varginha (Brasil) existen solo dos clases, con lo cual la posibilidad de clasificar a una instancia dentro de una tasa de infección es limitada a dos opciones. Por otro lado, las variables predictivas tomadas en cuenta para el conjunto de datos de Brasil dependen de los cálculos de periodos de incubación e infección y las condiciones climáticas en estos periodos, por lo cual las imprecisiones del proceso de definición de estos periodos pueden afectar en gran medida el valor de cada variable.

3.8. Resumen

Este capitulo presentó la caracterización del Sistema Experto para la detección de condiciones favorables para algunas tasas de infección de Roya en cultivos de café en Brasil y Colombia. Para este propósito, fueron definidas variables predictivas para esta enfermedad, a partir del conocimiento de expertos en la Roya y sus estudios de cada etapa de la enfermedad y los factores que la afectan. Las variables predictivas están relacionadas con información de monitorización climática de cultivos de café y propiedades físicas de los mismos, y es el insumo principal para la generación de clasificadores que relacionen condiciones existentes en el conjunto de datos con algunas tasas de infección de Roya registradas en los cultivos monitorizados. Adicionalmente, para la generación de los clasificadores fue usado un algoritmo de inducción de árboles de decisión, debido a que el modelo generado es fácilmente interpretable y cada rama del árbol puede ser tomada como una regla para una clase determinada. De esta forma, cada regla puede expresarse como un patrón de grafo.

4 Verificación de las reglas del Sistema Experto con base en emparejamiento de patrones en grafos

En el capítulo anterior fue expuesto el proceso para la generación de modelos de clasificación que permitan identificar condiciones favorables para tres tasas de infección de roya en cultivos de café. A partir de estos modelos, pueden ser identificadas una serie de reglas que determinan la predicción de cada tasa de infección. Teniendo en cuenta la cantidad de reglas que pueden ser extraídas, en este capítulo es abordada su representación como patrones de grafos. Asimismo, es presentado el proceso para la generación de un repositorio de grafos que contenga la información de monitorización y propiedades físicas de los cultivos. Por último, es definida la adaptación de un algoritmo de emparejamiento de patrones en grafos, con el fin de buscar los subgrafos que coincidan con los patrones generados, dentro del repositorio. De este modo, las coincidencias encontradas representan instancias donde las condiciones (climáticas y agronómicas) del cultivo presentan un estado favorable a una tasa de infección.

4.1. Representación basada en grafos

La representación basada en grafos es aprovechada en muchas áreas de conocimiento, debido a que pueden ser usados para modelar las relaciones y procesos dinámicos pertenecientes a cada área. Con el fin de incluir la variedad de semántica contenida en problemas de la vida real, es definido un *Grafo de Datos* como *G(V,E,L)* [13], donde:

- *V* es un conjunto de nodos.
- $E \subseteq VxV$ es un conjunto de aristas o relaciones entre los nodos, en donde *(v,v')* representa una arista del nodo *v* al nodo *v'.*
- L es una función definida en *V* de modo que para cada *v* en *V*, *L(v)* es la etiqueta de *v* y *L(v,v')* es la etiqueta de la arista entre el nodo *v* y el nodo *v'*.

Precisamente, en la práctica, *L(v)* puede indicar la variedad semántica mencionada, como tipos de relaciones, propiedades de nodos, entre otros.

A partir de la definición de *Grafo de Datos* expuesta, fue definida la representación basada en grafos de los datos de monitorización de clima y propiedades físicas de cultivos de café. Adicionalmente, las reglas para la predicción de la tasa de infección de roya fueron expresadas como patrones de grafo. A continuación, es presentado el proceso y las consideraciones para la generación de cada una de las representaciones mencionadas.

4.1.1. Repositorio de información como grafos

Los repositorios de información basados en grafos permiten modelar la información sobre un dominio de aplicación a través de entidades y sus relaciones. Esto genera un alto grado de interpretación, conservando la complejidad y dinámica, tanto de sus nodos, como de las relaciones entre estos.

Ahora bien, la información producida dentro de un entorno de producción agrícola es de naturaleza dinámica, donde se necesita modelar las variables (climáticas, agronómicas, de control, entre otras) que intervienen en el desarrollo del fruto. Estas variables son monitorizadas de forma constante, con el fin de relacionar sus variaciones con los niveles de producción, la toma de decisiones (control químico, fertilización, cosechas, etc), el surgimiento de enfermedades y plagas, entre otros. En consecuencia, la representación de datos basada en grafos brinda un modelo que satisface las necesidades para la persistencia de la información producida dentro de este entorno.

Tal como se ha expresado en los capítulos anteriores, en la presente tesis se ha tomado como dominio de aplicación la ocurrencia de la roya en el cafeto, razón por la cual fueron definidas en la sección 3.3 algunas variables que están relacionadas con esta enfermedad (atributos predictivos). Estas variables son obtenidas a partir de los datos de monitorización de clima brindados por una estación meteorológica. Siendo así, algunas de las entidades relevantes que corresponden a los nodos del grafo son:

- Cultivo: entidad principal que representa un cultivo determinado, donde sus etiquetas corresponden a propiedades como identificador único, nombre y localización.
- Instancia: entidad relacionada con el registro de los atributos predictivos para una escala de tiempo determinado, que en este caso es un mes. Esta entidad está conectada con otras entidades que determinan los tipos de atributos medidos y sus valores, como atributos de propiedades de cultivo, condiciones climáticas y control.
- Persona: entidad que representa a un ser humano, sus afiliaciones, datos personales y relación con el cultivo. Además, puede estar relacionado con otras entidades del mismo tipo.
- Entidades de línea de tiempo: los nodos etiquetados como Mes y Año corresponden a un grafo de soporte para la línea de tiempo, que caracteriza de forma jerárquica las propiedades del momento en que fue obtenida una instancia (año, mes, día). De esta forma, facilita la consulta de otros tipos de entidades bajo un criterio basado en fechas.

Los tipos de entidades identificados, indican el rol de cada nodo dentro del grafo que contiene la información de los cultivos. Por otro lado, las relaciones entre los nodos también pueden ser etiquetadas, lo cual permite establecer el contexto semántico existente en el ambiente de producción de café. A partir de estas consideraciones, fue propuesta una

estructura para el *Grafo de Datos* mostrada en la Figura 6, donde pueden apreciarse las entidades mencionadas y los diferentes tipos de relaciones existentes entre ellas. Además, los nodos contienen una serie de etiquetas que expresan los valores de las variables que estos representan.

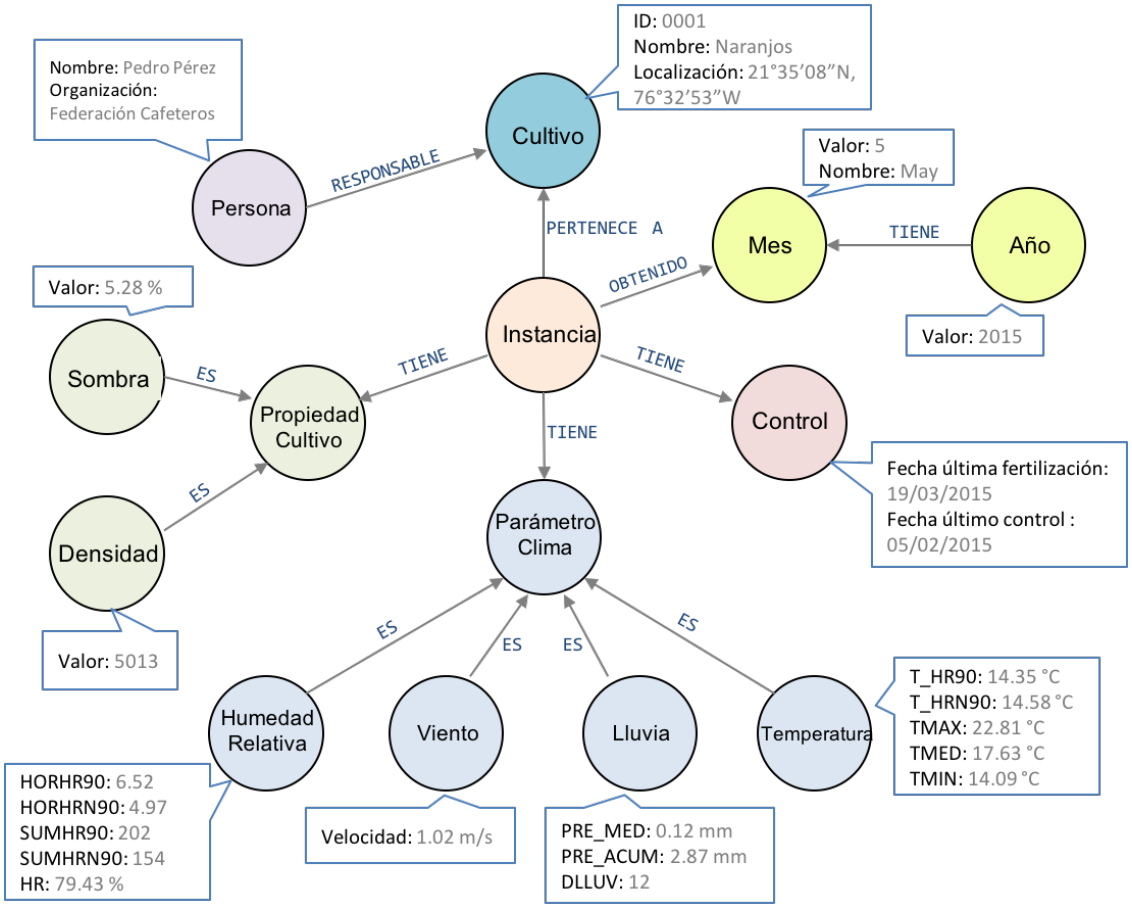


Figura 6. Ejemplo de un subgrafo del *Grafo de Datos*

Las representaciones basadas en grafos llevan consigo ventajas para los sistemas de administración de información, considerando que, como usuarios, somos capaces de inferir las dependencias semánticas entre las entidades, pero los modelos de datos (incluso las bases de datos) no siempre pueden identificar esas conexiones. Precisamente, la habilidad de añadir propiedades a nodos y relaciones es de gran utilidad para brindar metadatos enriquecidos semánticamente, los cuales pueden ser aprovechados por los algoritmos de minería de datos en grafos [85].

En este sentido, para obtener un *Grafo de Datos* con la estructura propuesta es necesario el desarrollo de un módulo mapeo de la información contenida en el conjunto de datos de mediciones de condiciones climáticas y propiedades de cultivos, a una representación

basada en grafos. Además, los grafos generados necesitan ser almacenados en una base de datos orientada a este tipo de representación de la información.

En la Tabla 10 son presentadas las variables que se encuentran en las bases de datos de monitorización y propiedades de cultivo, y cómo estas se relacionan con las variables predictivas empleadas para la extracción de reglas y las entidades modeladas en los nodos del *Grafo de Datos*.

Variable - información de cultivo		**Variable predictiva**	**Entidad modelada en nodo**
Monitorización del cultivo	Temperatura	TMAX TMED TMIN	Temperatura
	Humedad Relativa	HORHR90 HORHRN90 SUMHR90 SUMHRN90 HR	Humedad Relativa
	Humedad Relativa y Temperatura	T_HR90 T_HRN90	Temperatura
	Pluviosidad	DLLUV PRE_MED PRE_ACUM	Lluvia
	Velocidad de viento	VVIENTO	Viento
Propiedades del cultivo	Densidad de cultivo	DENSIDAD	Densidad
	Sombra	SOMBRA	Sombra

Tabla 10. Variables del repositorio de información, predictivas y entidades modeladas en nodos

Con base en la estructura para el *Grafo de Datos* propuesta y la relación entre variables del repositorio de información, variables predictivas y entidades en nodos, fue desarrollado un módulo de mapeo que sigue el proceso descrito en la Figura 7 y es descrito a continuación.

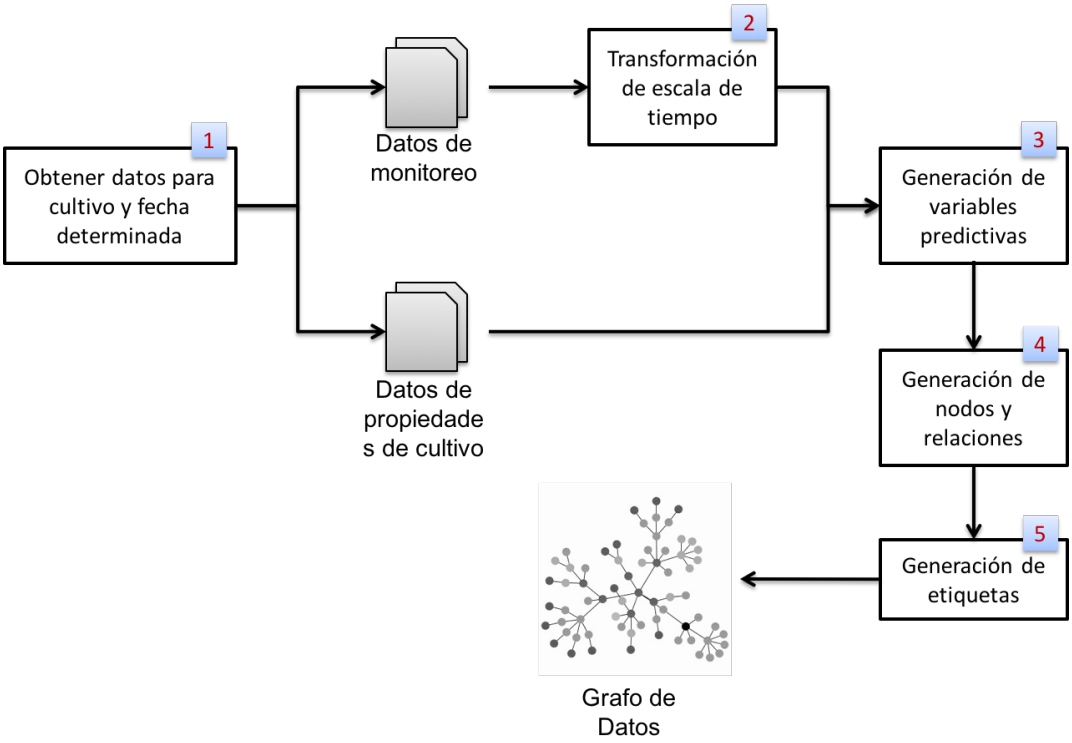


Figura 7. Mapeo del repositorio de información al *grafo de Datos*

El proceso inicia con la consulta de los datos contenidos en las bases de datos de monitorización de clima y propiedades de cultivos (1). Esta consulta se hace para cada cultivo, de modo recursivo, en intervalos de tiempo diarios, y da como resultado dos conjuntos de datos. El primer conjunto de datos contiene información del monitorización de variables agroclimáticas por parte de estaciones meteorológicas asociadas a cada cultivo, el cual pasa por una etapa de transformación de escala de tiempo (2), con el fin de obtener datos por hora, ya que las estaciones miden cada variable en intervalos de uno a cinco minutos. Por su parte, el segundo conjunto de datos contiene la información de propiedades físicas y manejo agronómico del cultivo. La siguiente etapa del proceso es la generación de las variables predictivas (3) para la Roya en el café, teniendo en cuenta los criterios de los expertos en la enfermedad, relatados en la sección 3.1, y la relación entre variables descrita en la Tabla 10. Con la generación de la información descrita anteriormente, son formados los nodos, según las entidades que representan, y relaciones que caracterizan sus vínculos con la entidad *Instancia,* que representa el estado del cultivo para una fecha determinada (4). Por último, los valores de las variables son almacenados como etiquetas de los nodos y son asociados con la entidad respectiva (5), como puede verse en el ejemplo de la Figura 6.

En el mismo sentido, la estructura del *Grafo de Datos* (Figura 6) propuesta considera el uso de etiquetas en el nodo que modela el cultivo (entidad *Cultivo*) con el fin de almacenar sus características como identificador, nombre, localización, entre otros. Además, los nodos de

tipo *Instancia* están conectados a nodos de "línea de tiempo", los cuales permiten que se realice una búsqueda en el *Grafo de Datos* de instancias según una fecha determinada con mayor facilidad.

De este modo, fue descrito el proceso de generación del *Grafo de Datos*, el cual es el repositorio de información donde será aplicado el emparejamiento de patrones en grafo, con el fin de encontrar instancias que coincidan con los patrones para la detección de condiciones favorables para Roya en el café, que serán definidos a continuación.

4.1.2. Reglas y clasificadores como patrones de grafo

Las reglas obtenidas a partir del conocimiento de expertos y aplicación de inducción de árboles de decisión pueden ser expresadas usando patrones de grafo. Por lo tanto, los patrones generados deben ser modelados de manera similar al grafo que representa el repositorio de información de cultivos, a nivel de la estructura y caracterización de las instancias involucradas en las reglas definidas.

Con base en la definición de Fan et al. [86], definimos un patrón de grafo como $Q = (V_p, E_p, f_v, f_e)$, donde:

- V_p es un conjunto de nodos y E_p es un conjunto de aristas dirigidas, tal como fueron definidos para un Grafo de Datos.
- $f_v()$ es una función definida en V_p, de modo que para cada nodo *u*, $f_v(u)$ es una etiqueta de *u.*
- $f_e()$ es una función definida en E_p, de modo que para cada arista (u, u') en E_p, $f_e(u, u')$ es una etiqueta de la relación entre los nodos (u, u').

De esta forma, estas funciones pueden ser utilizadas para especificar condiciones semánticas en la búsqueda o rangos de variables definidas por etiquetas, en términos de predicados Booleanos.

Ante las anteriores consideraciones, fueron generados 10 patrones de grafos que corresponden a cada una de las ramas del árbol de decisión obtenido en la sección 3.6.2 (Figura 5), los cuales pueden encontrarse en el Anexo A. Cada clase al final de las ramas (cuadro color gris) se relaciona con una tasa de infección predicha, por lo tanto, los patrones de grafo generados son divididos en tres grupos: 4 patrones para TI1, 4 patrones para TI2 y 2 patrones para TI3. En la siguiente tabla (Tabla 11), se encuentra un resumen de las propiedades de cada uno de los 10 patrones definidos.

Patrón	Entidad (nodo)	Valor etiqueta de variables predictivas	Tasa de infección
Patrón_TI1_1	Humedad Relativa	HORHRN90 <= 6.35	TI1: Desciende o se mantiene estable
	Temperatura	TMIN > 14.48	
Patrón_TI1_2	Humedad Relativa	1.84 < HORHRN90 <= 6.35 HORHR90 <= 4.97	TI1: Desciende o se mantiene estable
	Temperatura	TMIN <= 14.48 TMAX > 22.76 T_HRN90 > 14.58	
Patrón_TI1_3	Humedad Relativa	HORHRN90 <= 6.35	TI1: Desciende o se mantiene estable
	Temperatura	TMIN <= 14.48 TMAX <= 22.76	
Patrón_TI1_4	Humedad Relativa	HORHRN90 > 6.35	TI1: Desciende o se mantiene estable
	Lluvia	PRE_MED > 0.57	
Patrón_TI2_1	Humedad Relativa	HORHRN90 <= 6.35	TI2: Aumenta entre 0 y 2 puntos porcentuales
	Temperatura	TMIN <= 14.48 TMAX > 22.76 T_HRN90 <= 14.58	
Patrón_TI2_2	Humedad Relativa	HORHRN90 <= 6.35 HORHR90 > 4.97	TI2: Aumenta entre 0 y 2 puntos porcentuales
	Temperatura	TMIN <= 14.48 TMAX > 22.76 T_HRN90 > 14.58	
Patrón_TI2_3	Humedad Relativa	HORHRN90 <= 1.84 HORHR90 <= 4.97	TI2: Aumenta entre 0 y 2 puntos porcentuales
	Temperatura	TMIN <= 14.48 TMAX > 22.76 T_HRN90 > 14.58	
Patrón_TI2_4	Humedad Relativa	HORHRN90 > 6.35	TI2: Aumenta entre 0 y 2 puntos porcentuales
	Temperatura	TMED > 16.63	
	Lluvia	PRE_MED <= 0.57	
	Sombra	SOMBRA <= 5.28	
Patrón_TI3_1	Humedad Relativa	HORHRN90 > 6.35	TI3: Aumenta más de 2 puntos porcentuales
	Lluvia	PRE_MED <= 0.57	
	Sombra	SOMBRA > 5.28	
Patrón_TI3_2	Humedad Relativa	HORHRN90 > 6.35	TI3: Aumenta más de 2 puntos porcentuales
	Lluvia	PRE_MED <= 0.57	
	Sombra	SOMBRA <= 5.28	

	Temperatura	TMED <= 16.63	

Tabla 11. Características de los patrones generados para cada tasa

Por ejemplo, para la hoja etiquetada como "7" en el árbol de decisión mostrado en la Figura 5. Árbol de decisión generado para el conjunto de datos Los Naranjos, Colombia, el camino para llegar a TI3 comienza cuando HORHRN90 es mayor que 6,35 horas. En este caso, la siguiente variable a evaluar es PRE-MED (media precipitación media diaria). Si PRE-MED es menor o igual a 0,57 mm, se debe analizar la variable SOMBRA. Si el cultivo tiene un porcentaje de sombra de menos de 5.28%, la variable TMED debe ser evaluada (media de las temperaturas medias diarias). Por último, para la evaluación de TMED igual a o menor que 16,63 ºC, la clase predicha es TI3. Como resultado, en la Figura 8 se puede ver uno de los patrones de grafo generados para TI3 (Patrón TI3_2), obtenido a partir de la regla descrita por la rama del árbol de decisión mencionado anteriormente.

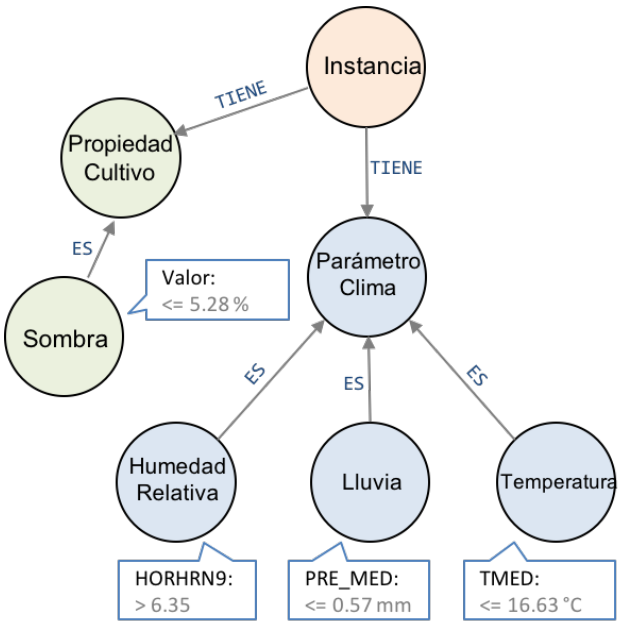


Figura 8. Patrón de grafo para TI3 (Patrón TI3_2)

En el patrón de grafo mostrado, cada nodo está asociado con una unidad de clasificación o entidad (instancia, parámetro climático, propiedad de cultivo), lo cual indica su rol en la dinámica del Grafo de Datos y la enfermedad que es deseada detectar. Además, esta representación permite agrupar las variables predictivas según su naturaleza, los tipos de relaciones establecen el contexto semántico del ambiente modelado. Asimismo, las etiquetas establecen el rango que deben cumplir los valores de cada variable para que exista un emparejamiento del patrón.

4.2. Algoritmo para el emparejamiento de patrones en grafo

El uso de técnicas de emparejamiento de patrones en grafos tiene como objetivo encontrar subgrafos o estructuras, dentro de un grafo de datos, que cumplan con ciertas condiciones en sus nodos y relaciones, definidas por un patrón de grafo. Para el objetivo del presente trabajo, el emparejamiento debe ser aplicado a un repositorio de grafos con las características descritas anteriormente y los patrones usados corresponden a las reglas identificadas a partir del árbol de decisión extraído para detección de condiciones favorables para la roya en el café (sección 3). De esta manera, es necesario usar un algoritmo para el emparejamiento de patrones que considere el análisis de grafos etiquetados, donde los patrones buscados, como es de esperar, también contienen etiquetas en sus nodos y relaciones.

4.2.1. Enfoques aproximados

A partir de la revisión del estado del arte relacionado con esta área (sección 2.2.3), pueden ser identificados los algoritmos más relevantes para el emparejamiento de patrones en grafos, como son el algoritmo de Ullman y el algoritmo VF2, tomados por la mayoría de investigaciones en esta técnica como fundamentos teóricos para el desarrollo de nuevos algoritmos y realizar modificaciones de estos que mejoren su desempeño. Por un lado, el algoritmo de Ullman hace uso de árboles de búsqueda para llevar la tarea de isomorfismo de subgrafo. Cada nodo en un grafo de datos que sea candidato a emparejamiento con un nodo del patrón debe tener un nodo hijo que concuerde con alguno de los nodos hijos del patrón. Este proceso es repetido para cada uno de los nodos del patrón, aumentando sucesivamente la profundidad de su búsqueda, con base en un grado definido. Adicionalmente, las coincidencias encontradas son almacenadas en un conjunto de nodos. El algoritmo luego da marcha atrás, hasta que no haya una ruta de búsqueda sin explorar. Por otro lado, el algoritmo VF2 puede ser descrito a partir de la representación de estados. Un estado es generado cada vez que es encontrada una coincidencia entre el patrón y el grafo de datos, el cual es evaluado recorriendo cada nodo y sus conexiones. Estos estados se generan utilizando reglas de viabilidad que quitan pares de nodos que no pueden ser isomorfos. No obstante, estos dos algoritmos fueron diseñados originalmente para una búsqueda topológica en la estructura del grafo, sin tener en cuenta el establecimiento de etiquetas en nodos y relaciones. En consecuencia, diversas investigaciones han propuesto el análisis de grafos etiquetados dentro de los procesos de emparejamiento de patrones. Una de ellas [87] [44], presenta una comparación del desempeño de los algoritmos de Ullman y VF2, considerando grafos etiquetados y, además, propone un nuevo algoritmo llamado “DualIso” que lleva a cabo un emparejamiento de patrones en grafos etiquetados.

4.2.2. Algoritmo DualIso

Este algoritmo tiene como punto inicial la obtención de los emparejamientos factibles de un patrón dentro de un grafo de datos. El tipo de emparejamiento considerado es de tipo exacto, donde no es considerado el grado de similitud del patrón con un subgrafo, sino su coincidencia total. De esta forma, dado un nodo u correspondiente al patrón buscado, es creado un conjunto $\Phi(u)$ que contiene todos los nodos del grafo de datos que son del mismo tipo que u. Este proceso es similar al primer paso llevado a cabo por el algoritmo de Ullman, excepto que no tiene en cuenta restricciones en el grado de la profundidad de búsqueda. Como segundo paso es llevada a cabo la poda del grafo consultado, lo cual tiene como objetivo reducir el espacio de búsqueda. Para este proceso, DualIso hace uso del concepto de *simulación doble de grafo* [36], que reduce el espacio de búsqueda al comprobar, dado un nodo candidato para el emparejamiento, sus nodos hijos, así como sus nodos padre. El pseudocódigo del algoritmo que lleva a cabo la simulación doble es presentado a continuación:

Algoritmo 1. DualSim – Simulación doble. [44]

Procedimiento $DualSim(G, Q, \Phi)$:
  $modificado \leftarrow verdadero$
  Mientras $modificado$ **hacer**
    $modificado \leftarrow falso$
    Para $u \leftarrow V_q$ **hacer**
      Para $u' \leftarrow Q.ady(u)$ **hacer**
        $\Phi'(u') \leftarrow \emptyset$
        Para $v \leftarrow \Phi(u)$ **hacer**
          $\Phi_v(u') \leftarrow G.ady(v) \cap \Phi(u')$
          Si $\Phi_v(u') = \emptyset$ **entonces**
            eliminar v de $\Phi(u)$
            Si $\Phi(u) = \emptyset$ **entonces**
              retorna Φ vacío
            Fin Si
            $modificado \leftarrow verdadero$
          Fin Si
          $\Phi'(u') \leftarrow \Phi'(u') \cup \Phi_v(u')$
        Fin Para
        Si tamaño de $\Phi'(u')$ es menor que $\Phi(u')$ **entonces**
          $modificado \leftarrow verdadero$
        Fin Si
        $\Phi(u') \leftarrow \Phi(u') \cap \Phi'(u')$
      Fin Para
    Fin Para
  Fin Mientras
  retorna Φ

27: Fin Procedimiento

Para entender el procedimiento del algoritmo anterior, partimos de un grafo de datos $G(V, E, l)$ (definiendo el conjunto de nodos, aristas y etiquetas) y un patrón de grafo $Q(V_q, E_q, l_q)$. Las operaciones principales del algoritmo se encuentran entre la línea 5 y 9. De esta forma, el conjunto $\Phi(u')$ contiene los nodos que tienen un nodo hijo en $\Phi(u)$ y, dado a que al menos alguno de los nodos $\Phi(u')$ debe contener alguno de sus nodos padres en $\Phi(u)$, cualquier nodo válido en $\Phi(u')$ debe estar contenido en la intersección descrita después del ciclo **Para** (línea 8). Por lo tanto, es construido un nuevo conjunto $\Phi'(u')$ que contiene sólo los nodos que tienen un padre en $\Phi(u)$ (línea 17). Una vez que todos los nodos en $\Phi(u)$ han sido recorridos, el nuevo $\Phi(u')$ es tomado para realizar la intersección $\Phi(u') \leftarrow \Phi(u') \cap \Phi'(u')$ (línea 22). De esta forma, es obtenido el conjunto de nodos que cumplen con las condiciones del emparejamiento a través de un proceso que es efectivo en grafos de gran tamaño.

Ahora bien, teniendo en cuenta el procedimiento de la *simulación doble*, DualIso hace uso de esta reducción del espacio de búsqueda para definir su procedimiento de emparejamiento de patrones en grafos etiquetados, el cual es llamado *Isomorfismo basado en doble simulación (Dual-based Isomorphism)*. Este algoritmo está descrito en el siguiente pseudocódigo:

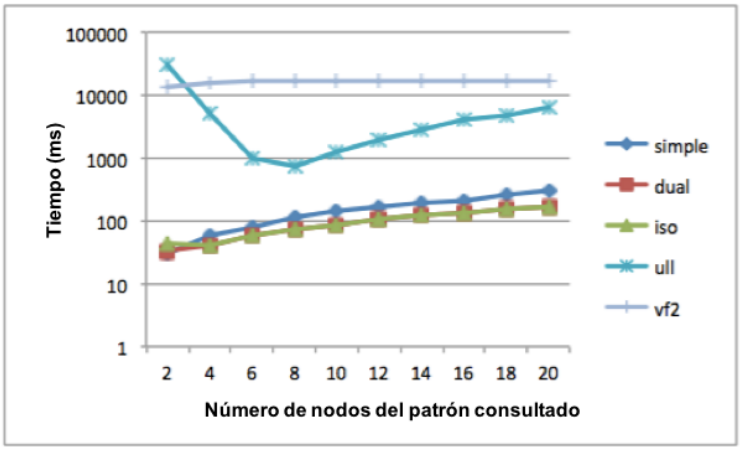
Algoritmo 2. Dual-based Isomorphism – Isomorfismo basado en simulación doble. [44]

```
1:  Procedimiento EmparejamientoPatron(G, Q)
2:      coincidencias ← ∅
3:      Φ0 ← CoincidenciasFactibles(G, Q)
4:      Φ0 ← DualSim(G, Q, Φ)
5:      Búsqueda(G, Q, Φ0, 0)
6:
7:      Procedimiento Búsqueda(G, Q, Φ0, profundidad)
8:          Si profundidad = Q.tamaño entonces
9:              coincidencias ← coincidencias ∪ Φ
10:         Si no
11:             Para v ← Φ(profundidad) hacer
12:                 Si v ∉ Φ(0) ∪…∪ Φ(profundidad − 1) entonces
13:                     Φ' ← copia de Φ
14                      Φ'(profundidad) ← {v}
15:                     Φ' ← DualSim(G, Q, Φ)
16:                     Si Φ' no está vacío
17:                         Búsqueda(G, Q, Φ', profundidad + 1)
18:                     Fin Si
19:                 Fin Si
20:             Fin Para
```

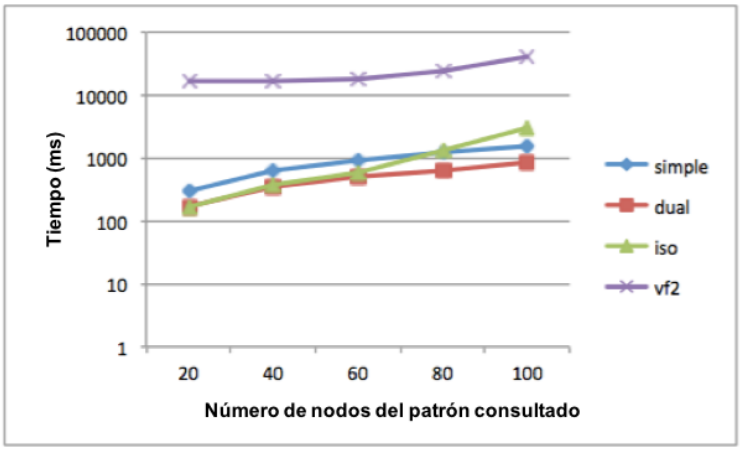
21: **Fin Si**
22: **Fin Procedimiento**
23:
24: **retorna** $coincidencias$
25: **Fin Procedimiento**

El procedimiento para el emparejamiento de patrones propuesto en DualIso comienza con la búsqueda de coincidencias factibles para cada nodo u del patrón de grafo a través de la recuperación de todos los nodos v en el grafo de datos que cumplan con la condición $l(v) = l_q(u)$; es decir, que tengan la misma etiqueta. Ahora bien, el proceso de simulación doble es aplicado al grafo obtenido en el paso anterior, que contiene los nodos con etiquetas similares a las del patrón. Como resultado, el espacio de búsqueda es limitado a los nodos candidatos para el emparejamiento. En el paso siguiente, es aplicado el proceso *Búsqueda*, que recorre los resultados a partir de una búsqueda en profundidad. Primero, es clonado Φ a Φ' (línea 13) y luego $\Phi'(\text{profundidad})$ es igualado a $\{v\}$, tratándolo como el único nodo que coincide con el nodo del patrón de búsqueda indexado en la posición dada por el valor de profundidad del ciclo de búsqueda. La simulación doble es llevada a cabo para Φ', que necesariamente remueve los nodos en $\Phi(1), \ldots \Phi(|V_q| - 1)$ que no están contenidos en el emparejamiento a partir de isomorfismo (coincidencia en estructura). Si no hay coincidencias viables para todos los nodos en Φ' en este punto, significa que no existe ninguna coincidencia isomorfa para la búsqueda actual y, por lo tanto, el algoritmo retrocede en el índice de búsqueda. De lo contrario, el procedimiento de búsqueda continúa de forma recursiva hacia el siguiente nivel de profundidad hasta la profundidad máxima.

De esta manera, DualIso presenta un método para llevar a cabo el emparejamiento de patrones de tipo exacto en grafos etiquetados, bajo una eficiencia en el uso de memoria y tiempo de ejecución, tal como sus autores presentan en algunas pruebas de desempeño donde es comparado con otros algoritmos. En la Figura 9, puede apreciarse la comparación para diferentes configuraciones de tamaño de patrón de grafo, para búsqueda en anchura, entre DualIso (dual), simulación simple (simple), isomorfismo simple (iso), algoritmo de Ullman (ull) y VF2. Para estas pruebas, fue usado un grafo de datos de 10^6 nodos y patrones sin considerar etiquetas, donde el tamaño de los patrones de grafo buscados fue incrementado sucesivamente. En los resultados puede verse cómo el proceso de reducción de espacio de búsqueda llevado a cabo por DualIso y simulación simple, reduce el tiempo necesario para llevar a cabo la tarea de búsqueda de coincidencias a un patrón dado. Para patrones de gran tamaño (b), los resultados de Ullman no son considerados, dado a sus altos valores en tiempo de ejecución. Además, tanto para tamaños de patrón pequeños (a) y grandes (b), DualIso presenta una tendencia a la estabilización del tiempo que toma su procesamiento, a medida que el número de nodos del patrón consultado crece.



(a) Tamaños de patrón pequeños, búsqueda en anchura, $|V| = 10^6$



(b) Tamaños de patrón grandes, búsqueda en anchura, $|V| = 10^6$

Figura 9. Comparación de diferentes algoritmos y técnicas de emparejamiento de patrones en un grafo de datos, bajo distintas configuraciones de patrones [87]

Ahora bien, para el caso del desempeño del emparejamiento en grafos y patrones de grafos etiquetados, los resultados de la comparación hecha entre los algoritmos es presentada en la Figura 10. Debido a que el número de etiquetas afecta en gran medida las coincidencias posibles para cada nodo en una consulta, es de esperar que tenga un impacto de consideración en el tiempo de ejecución para todos los algoritmos probados. Para estos experimentos, los autores generaron grafos con un número de etiquetas especificado, distribuidas aleatoriamente a través de los nodos. Esto significa que para un grafo con n etiquetas, el poder de filtrado de una etiqueta es de aproximadamente $\frac{1}{n}$. Por lo tanto, mientras más etiquetas existan en el patrón buscado, menos nodos hay en el espacio de búsqueda antes del proceso de simulación doble inicial.

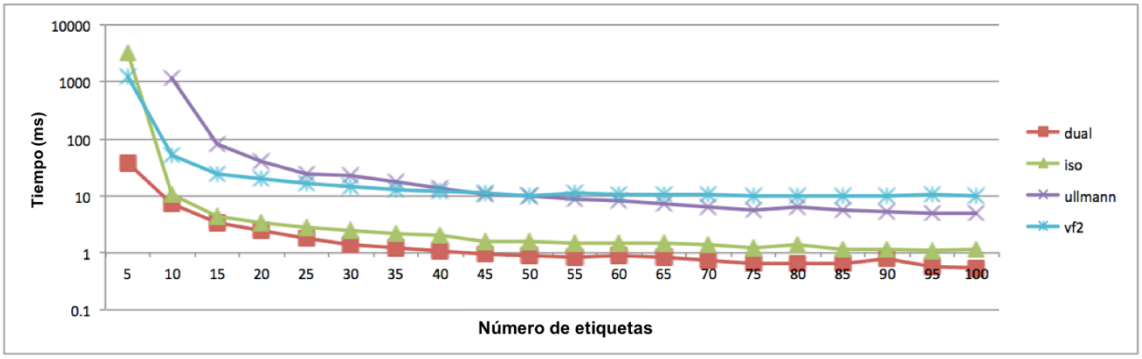


Figura 10. Efecto de la consideración de etiquetas en el desempeño de diferentes algoritmos para el emparejamiento de patrones en grafos [87]

De esta forma, es claro que el proceso de emparejamiento de patrones en grafo toma mayor tiempo para el caso de grafos sin etiquetas, debido a que todo nodo en el grafo de datos es inicialmente un candidato para todos los nodos en el patrón buscado.

4.3. Adaptación del algoritmo DualIso

En la sección pasada fueron explorados algunos algoritmos y técnicas para llevar a cabo el emparejamiento de patrones en grafos. De forma específica y, siguiendo el objetivo y características del grafo de datos a ser usados en este trabajo de investigación, fue considerado el emparejamiento en grafos y patrones de grafo que contengan etiquetas en sus nodos y relaciones. Esta búsqueda dio como principal candidato al algoritmo DualIso, que se basa en conceptos del algoritmo de Ullman y una técnica de reducción de espacio de búsqueda llamada *simulación doble de grafo*, para generar un procedimiento de búsqueda de patrones en grafos etiquetados, con mejoras en tiempos de ejecución y desempeño.

Ahora bien, a partir de la definición de la estructura del grafo de datos y patrones para el problema tratado en este trabajo, es necesario definir una adaptación de DualIso, con el fin de encontrar las ocurrencias de los patrones para las distintas tasas de incidencia de Roya generadas en el capítulo 3. Esta adaptación debe considerar el acceso a una base de datos de grafos, en lugar de la carga del grafo en memoria, donde estará almacenado el grafo de datos que contiene la información sobre la monitorización de variables climáticas y propiedades físicas de los cultivos. De manera adicional, es posible aprovechar algunas funciones que brindan los sistemas de administración de bases de datos en grafos, para llevar a cabo algunas tareas del emparejamiento de patrones.

De lo anteriormente dicho, fue generada una adaptación del algoritmo *DualIso*, llamada *CRD-GPM*, la cual lleva a cabo el proceso de emparejamiento de patrones en grafos para encontrar condiciones favorables para el desarrollo de tres tasas de infección de Roya en el café colombiano. Adicionalmente, esta adaptación considera la consulta y

almacenamiento del grafo de datos, que contiene la información sobre la monitorización de cultivos, a través de un sistema de gestión de base de datos en grafos, aprovechando algunas funciones de estos sistemas en cuanto a métodos de búsqueda e indexación. A continuación, es mostrado el pseudocódigo de la adaptación mencionada.

Algoritmo 3. CRD-GPM. Adaptación de Duallso para Emparejamiento de Patrones en Grafos

```
1:  Procedimiento EmparejamientoPatron(Q)
2:      coincidencias ← ∅
3:      Φ ← CoincidenciasFactibles( tipo nodo)
4:      σ_o ← CoincidenciasREntrada(Q, Φ)
5:      σ_i ← CoincidenciasRSalida(Q, Φ)
6:      Φ ← σ_o ∩ σ_i
7:      Para i = 0 hasta Φ.tamaño hacer
8:          GPM ← Búsqueda(Q, Φ_i)
9:          Si GPM entonces
10:             coincidencias ← coincidencias ∪ Φ_i
11:         Fin Si
12:     Fin Para
13:     retorna coincidencias
14  Fin Procedimiento
15:
16: Procedimiento Búsqueda(Q, Φ)
17:     emparejamiento ← falso
18:     nodoInicioQ ← Encontrar(n: Instancia) en Q
19:     nodoInicioΦ ← Encontrar(n: Instancia) en Φ
20:     Mientras emparejamiento hacer
21:         Para j = 0 hasta Q.tamaño hacer
22:             Si recorrido(Q, j, nodoInicioQ) ∈ recorrido(Φ, j, nodoInicioΦ)
23:                 λ ← pareja nodos de cada recorrido
24:                 Para k = 0 hasta λ.tamaño hacer
25:                     Si λ_k(Q).etiquetas = λ_k(Φ).etiquetas entonces
26:                         emparejamiento ← verdadero
27:                     Si no
28:                         emparejamiento ← falso
29:                     Fin Si
30:                 Fin Para
31:             Fin Si
32:         Fin Para
33:     Fin Mientras
34:     retorna emparejamiento
35: Fin Procedimiento
```

El procedimiento para el emparejamiento de patrón en grafos comienza con la generación de un conjunto Φ a partir de la función $CoincidenciasFactibles(\ tipo\ nodo)$, que realiza una búsqueda de los nodos que sean del mismo tipo que el parámetro de consulta "tipo nodo". Para el propósito de este trabajo, el tipo de nodo corresponde a los nodo *Instancia*, que representa los registros de variables climáticas y propiedades de cultivo, para una fecha y cultivo determinado, de acuerdo a la estructura del grafo de datos y patrones definidos anteriormente. De este modo, se realiza una primera reducción del espacio de búsqueda. De forma seguida, son generados dos conjuntos σ_o y σ_i con un fin similar a la simulación doble llevada a cabo en DualIso. σ_o contiene los nodos *Instancia* en Φ con relaciones de salida similares a las del nodo *Instancia* del patrón de grafo, mientras σ_i contiene los nodos *Instancia* en Φ con relaciones de entrada similares al patrón. Como paso siguiente, Φ es redefinido como la intersección de σ_o , y σ_i y a cada elemento de Φ le es aplicado un procedimiento llamado *Búsqueda*, que comprueba la coincidencia del patrón de grafo alrededor de dicho nodo. Este procedimiento inicia con la identificación del nodo de inicio de búsqueda tanto en el patrón como en el elemento de Φ que va a ser analizado, que en este caso son los nodos *Instancia*. El ciclo **Para** (línea 21) recorre el patrón desde el nodo de inicio, incrementando la profundidad de búsqueda, con el fin de que la coincidencia sea evaluada en todos los nodos del patrón. Para esto, es utilizada la función *recorrido*, que extrae los nodos y relaciones asociadas a un nodo de inicio, para diferentes profundidades de conexiones en el grafo, a través del marco *Traversal*, brindado por el sistema de administración de base de datos en grafos. Siendo así, son comparados los resultados de la función *recorrido* para el patrón y el nodo del grafo en diferentes profundidades. En las iteraciones donde los resultados coinciden, es creado un conjunto que contiene las parejas de nodos que corresponden a la coincidencia (nodo del patrón y nodo del grafo) y, posteriormente se comparan sus etiquetas. En el caso de que las etiquetas coincidan para todas las profundidades, es obtenida una coincidencia total del patrón (coincidencia en estructura y etiquetas) y el elemento de Φ que se encontraba siendo analizado es añadido al conjunto de coincidencias totales que representan las instancias donde existen condiciones favorables para la tasa de infección de Roya relacionada al patrón buscado.

4.4. Caso de estudio: Detección de tasas de infección de roya a partir del Emparejamiento de Patrones en Grafos

Para entender la aplicación del algoritmo propuesto con el objetivo de detectar condiciones favorables para distintas tasas de infección de roya, a continuación es explicado brevemente un ejemplo de su uso.

Los 10 patrones de grafo definidos anteriormente (Tabla 11) corresponden cada uno a una tasa de infección de roya. Tomando en particular el patrón TI3_2, que predice una tasa de infección mayor a 2 puntos porcentuales, el algoritmo comienza la búsqueda de los nodos

tipo *Instancia*. Estos nodos contienen la información de monitorización climática y agronómico de un cultivo determinado, razón por la cual es tomado como nodo principal a ser buscado. A continuación, se determina si las relaciones de salida y entrada en cada nodo del grupo de instancias es igual a estas relaciones en el patrón. Para este caso, el nodo *Instancia* del patrón solo contiene relaciones de salida. A los nodos que cumplieron con la anterior condición, el procedimiento de ***Búsqueda*** les es aplicado. En este proceso, son evaluados los recorridos de las relaciones salientes y entrantes del nodo *Instancia.* En el patrón TI3_2 se tienen 4 recorridos, que representan las variables implicadas en la predicción de TI3: Instancia-Propiedad Cultivo-Sombra, Instancia-Parámetro Clima-Humedad Relativa, Instancia-Parámetro Clima-Lluvia, Instancia-Parámetro Clima-Temperatura. Como paso siguiente, se comparan los valores de las etiquetas de los nodos, las cuales contienen los valores de las variables predictivas. Dado que los valores en los patrones están en términos de operadores lógicos, la comparación reconoce el tipo de operador implicado en el patrón y determina si el valor está dentro del rango establecido. En el caso de que uno de los nodos *Instancia* del *Grafo de Datos* cumpla con todas las condiciones del patrón, este nodo representa una condición de un cultivo que es propicio a tener una tasa de infección TI3.

De esta manera, la búsqueda de cada patrón a través del algoritmo propuesto tendrá como resultado un conjunto de nodos *Instancia* que registraron condiciones favorables para la tasa de infección que cada patrón representa.

4.5. Resumen

En este capítulo fueron presentados los elementos necesarios para llevar a cabo la verificación de las reglas del Sistema Experto, haciendo uso del emparejamiento de patrones en grafos. Para lograr este objetivo, fue considerada una representación, basada en grafos, de los datos de monitorización y propiedades físicas de los cultivos, de acuerdo a las variables predictivas definidas en el capítulo 3, generando un grafo de datos. Este grafo se encuentra almacenado en una base de datos orientada a grafos, donde puede ser consultado para encontrar las coincidencias de cada patrón. Adicionalmente, fue definida la estructura de los patrones de grafo para las diferentes tasas de infección de Roya, obtenidas a partir del árbol de decisión para el conjunto de datos de Los Naranjos, teniendo en cuenta que debe ser similar a la estructura del grafo de datos que representa el repositorio de información de condiciones presentes en los cultivos. Por último, fueron presentados algunos algoritmos para el emparejamiento de patrones en grafos, escogiendo *DualIso* como el punto de partida para este trabajo de investigación, debido a su mejor desempeño frente sus similares y los requerimientos del problema a resolver.

5 Prototipo y experimentación

Este capítulo contiene la descripción del prototipo del Sistema Experto basado en Emparejamiento de Patrones en Grafos para la detección de condiciones favorables para distintas tasas de infección de Roya en el café colombiano. Además, presenta el diseño del experimento para la validación del prototipo, planificación de las pruebas realizadas y la presentación de los resultados obtenidos.

5.1. Prototipo

Los capítulos 3 y 4 describieron el uso de una representación basada en grafos, aplicada a una serie de reglas para la detección de condiciones favorables para tres tasas de infección de Roya en el café, extraídas a través de la inducción de árboles de decisión en un conjunto de datos de monitorización y propiedades de cultivos cafeteros. Adicionalmente, fue presentada la estructura del *Grafo de Datos* que contiene la información sobre las unidades de cultivo, la cual agrupa, en nodos llamados *Instancia*, el estado de variables en cada cultivo (parámetros climáticos, propiedades, información de control y manejo) para una fecha determinada. Por último, fue definido un algoritmo para la búsqueda de coincidencias de los patrones mencionados en el *Grafo de Datos,* con el fin de encontrar las *Instancias* donde existan condiciones propicias para una tasa determinada de infección de Roya. En consecuencia, fue desarrollado un prototipo que implementa un Sistema Experto basado en el Emparejamiento de Patrones en Grafos para la identificación de la enfermedad en cultivos de café mencionada.

Por su parte, el dominio de aplicación del prototipo fue la granja experimental *Los Naranjos*, perteneciente a la empresa Supracafé, la cual se encuentra ubicada en el municipio de Cajibío (Cauca). En este entorno, el objetivo es identificar condiciones favorables para la ocurrencia de tres tasas de infección de Roya en los cultivos de café.

En las siguientes secciones son descritas las funcionalidades, arquitectura e interfaz del sistema propuesto.

5.1.1. Funcionalidades del sistema

Teniendo en cuenta las consideraciones para el desarrollo del Sistema Experto presentadas anteriormente, es necesario el uso de bases de datos orientadas a grafos, que son un sistema de administración de información con métodos para Crear, Leer, Actualizar y Eliminar (CRUD por sus siglas en inglés), exponiendo un modelo de datos en grafos. En consecuencia, normalmente están optimizadas para un alto rendimiento e integridad en las transacciones y una buena disponibilidad operativa [85]. En estas bases de datos, las relaciones son elementos principales del modelo de datos en grafo, diferenciándose de otros sistemas de gestión de base de datos, donde las conexiones entre las entidades

deben ser inferidas haciendo uso de claves externas o reducción de mapeo. Precisamente, para el desarrollo del presente trabajo de grado, fue usado el motor de bases de datos orientada a grafos basado en software libre, Neo4j [88, p. 4]. Este motor ofrece esquemas flexibles, almacenamiento y procesamiento de grafos de forma nativa, transacciones de tipo ACID (por las siglas en inglés de Atomicidad, Consistencia, Aislamiento y Durabilidad), un lenguaje de consulta, interfaces de programación de diferentes lenguajes, acceso a través de clientes http, entre otras ventajas.

De esta manera, el algoritmo propuesto (CRD-GPM) toma ventaja de *Cypher,* una de las características más relevantes de Neo4j, el cual es un lenguaje declarativo inspirado en SQL para la consulta dentro de grafos. Esto significa que al usar *Cypher,* es especificado lo que se quiere buscar dentro de la base de datos de grafos, sin tener en cuenta la manera de encontrarlo. Adicionalmente, las sentencias pueden ser enviadas a través de un cliente http haciendo uso del controlador JDBC para Neo4j, habilitando el acceso remoto a la base de datos. El uso de este lenguaje permite que los pasos dentro del algoritmo no requieran un almacenamiento de todo el grafo consultado en memoria, lo cual lleva a un menor uso de memoria, aunque incremente el tiempo de ejecución, al requerir de una apertura de conexión http para cada sentencia de *Cypher* a ser ejecutada.

En este sentido, la arquitectura lógica, que contiene los diferentes componentes del sistema y su organización, es mostrada en la Figura 11.

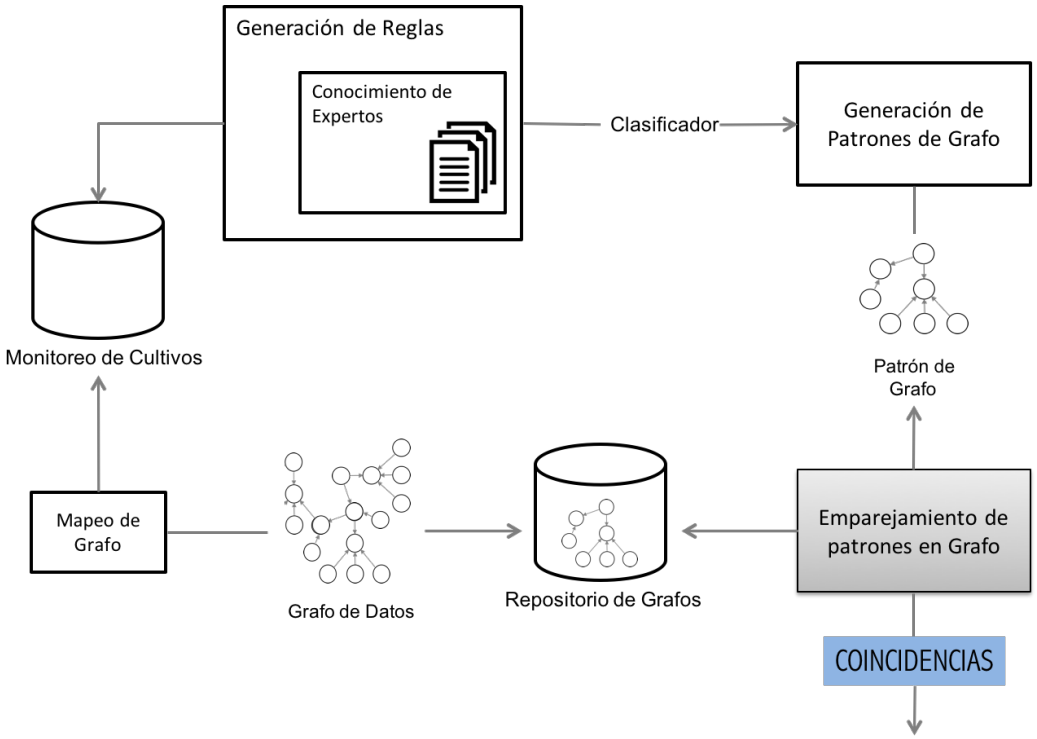


Figura 11. Arquitectura lógica del sistema

A continuación, son descritos sus componentes:

- **Monitorización de Cultivos:** corresponde a una base de datos relacional, la cual contiene los datos monitorizados por estaciones meteorológicas ubicadas en los cultivos de café, sus propiedades agronómicas y la información de control de los mismos.

- **Mapeo de Grafo:** este componente es responsable de la transformación a una representación basada en grafos de la información contenida en la base de datos *Monitorización de Cultivos*, conforme al proceso descrito en la sección 4.1.1 (Figura 7). Como resultado, es obtenido un *Grafo de Datos* que se almacena en el *Repositorio de Grafos.*

- **Repositorio de Grafos:** representa la base de datos orientada a grafos, almacenada y creada en Neo4j. Esta repositorio contiene el *Grafo de Datos* creado en el componente descrito anteriormente.

- **Generación de reglas:** compuesta por el algoritmo para inducción de árboles de decisión descrito en la sección 3.4, el cual hace uso de las variables predictivas para el desarrollo de la Roya (sección 3.1), definidas a partir del conocimiento de expertos en esta enfermedad. El resultado de este módulo está dado por un *clasificador* que representa un conjunto de reglas. Estas reglas expresan las condiciones que deben darse en los valores de las variables predictivas, para conducir de esta manera a una de las tasas de infección establecidas.

- **Generación de Patrones de Grafo:** su función es la generación de patrones de grafo a partir del clasificador representado por el árbol de decisión que fue obtenido en el módulo de *Generación de Reglas.* Para esto, cada rama del árbol de decisión es tomada como una regla del tipo “Si…entonces” (premisas y conclusión). Cada conclusión corresponde a una tasa de infección, lo que lleva a tener una serie de reglas para cada tasa. Por último, estas reglas son expresadas con base en una representación basada en grafos, como describe la sección 4.1.2. El resultado de este componente es una serie de patrones de grafos para cada tasa de infección definida.

- **Emparejamiento de Patrones en Grafo:** este componente contiene el algoritmo adaptado para la búsqueda de patrones en el *Grafo de Datos*, descrito en la sección 4.2. Para llevar a cabo esta tarea, toma un patrón para su búsqueda e implementa sus procesos a través de consultas a la base de datos orientada a grafos. Como resultado, se obtiene un conjunto de coincidencias encontradas, que representan las instancias donde las condiciones del cultivo (climáticas y agronómicas) son favorables para la tasa de infección que representa el patrón buscado.

A partir de los componentes descritos anteriormente, la secuencia lógica de la arquitectura presentada está divida en dos procesos: La **generación de patrones de grafo**, ejecutada de manera asistida en intervalos de tiempo irregulares, ya que su iteración tiene como objetivo el refinamiento del modelo generado para extraer las reglas que determinan las condiciones para cada tasa infección. Por este motivo, es necesario que exista una gran cantidad de registros nuevos en la base de datos de monitorización de cultivos y nuevos registros del seguimiento de la enfermedad en los mismos. Por otro lado, el proceso de **búsqueda de patrones en grafo** es ejecutado diariamente de manera automática, con el fin de identificar el cambio de condiciones en las variables predictivas de los cultivos y su relación con cada tasa de infección.

Generación de Patrones de Grafo:

1) El componente *Generación de Reglas* realiza una consulta a la base de *Monitorización de Cultivos* y extrae la información necesaria para realizar la preparación del conjunto de datos a partir de las variables predictivas para la enfermedad, como es descrito en la sección 3.4.
2) Los clasificadores obtenidos son tomados por el componente *Generación de Patrones de Grafo* para generar una serie de patrones de grafo, divididos en tres grupos que representan cada tasa de infección definidas.

Búsqueda de Patrones en Grafo:

1) El componente *Mapeo de Grafo* realiza una consulta diaria que extrae los datos de *Monitorización de Cultivo* de un mes atrás, a partir de la fecha en curso. Esto es debido a que las variables predictivas están dadas en escala mensual, como fue mencionado en el capítulo 3.
2) El nodo de tipo *Instancia* generado es asociado a la fecha actual, a partir de la generación de una relación con un nodo tipo día del grafo de *Línea de Tiempo*. Este proceso es realizado para todas las unidades de cultivo monitorizadas y es actualizado el *Grafo de Datos.*
3) El componente de *Emparejamiento de Patrones en Grafo* consulta los nodos *Instancia* que hayan sido generados en la fecha actual y lo toma como el grafo *G* de entrada para el algoritmo *CRD-GPM*.
4) El proceso de búsqueda es llevado a cabo para los tres grupos de patrones. Como resultado, son obtenidas las instancias donde existen condiciones favorables para: disminución o estabilidad en infección de Roya (TI1), aumento del porcentaje de incidencia de Roya entre 0 y 2 puntos porcentuales (p.p.), y aumento del porcentaje de incidencia de Roya por más de 2 p.p.

5.1.2. Vista lógica del sistema

La vista lógica organiza el Sistema Experto en paquetes, subsistemas y capas. Esta es implementada en tres capas diferentes: aplicación, mediación y capa de almacenamiento y servicio [89]. La Figura 12 muestra las capas de la arquitectura y la interacción entre ellas, así como los paquetes más importantes que componen cada capa.

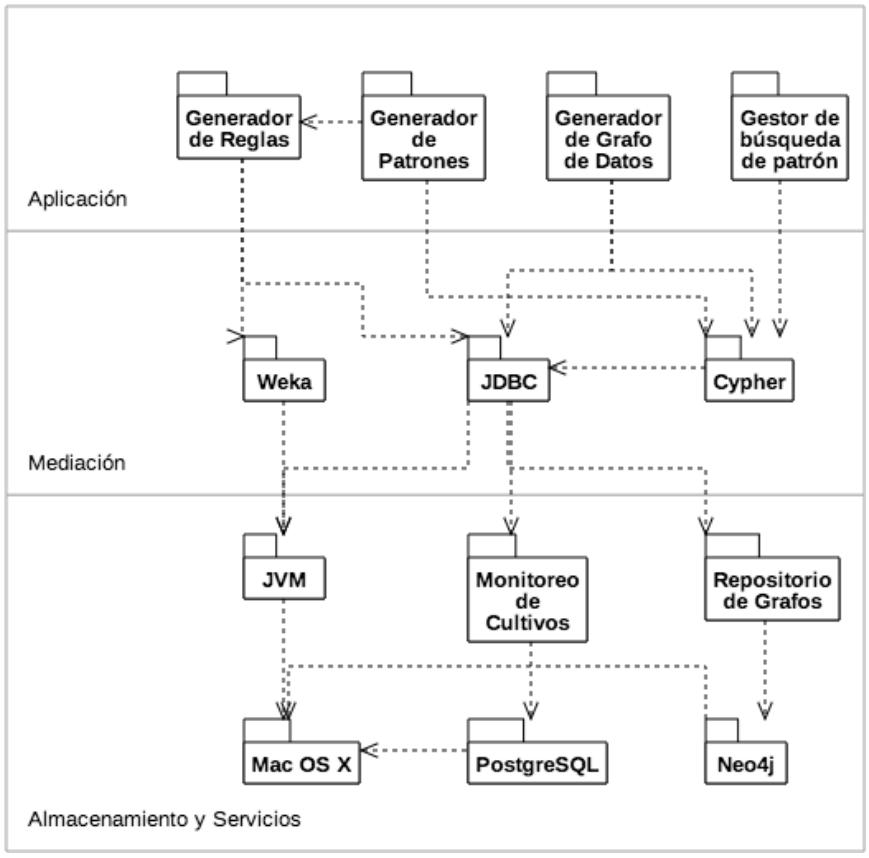


Figura 12. Vista lógica del Sistema Experto

- **Capa de Aplicación:** contiene los paquetes que implementan las funcionalidades del sistema experto. Está compuesta por:

    - Generador de reglas: tiene como función aplicar el algoritmo de inducción del árbol de decisión, con el fin de extraer las reglas que pueden ser interpretadas a partir de las ramas de este.

    - Generador de patrones: su función es obtener las reglas generadas y representarlas como patrones de grafo.

- Generador de Grafo de Datos: es el encargado de extraer la información de monitorización de cultivos para generar las variables predictivas de la enfermedad y posteriormente representarlas en el grafo de datos.

- Gestor de Búsqueda de Patrón: contiene el algoritmo para el emparejamiento de patrones en grafos. Para llevar a cabo la tarea de este algoritmo, debe contarse con los patrones de grafo y el grafo de datos creados en los paquetes anteriores. Adicionalmente, este paquete está encargado de la aplicación del algoritmo para cada patrón relacionado a las tasas de infección definidas, realizando su ejecución de forma periódica en escala diaria.

- **Capa de Mediación:** compuesta por los elementos que permiten la comunicación entre las capas de aplicación y almacenamiento. Sus componentes son:

  - Weka: es una plataforma de libre distribución, basada en Java, que implementa un conjunto de algoritmos de aprendizaje automático y minería de datos.

  - JDBC (Java Database Connectivity): es la API que permite la gestión de las operaciones de conexión a bases de datos del sistema experto desde Java.

  - Cypher: es un lenguaje declarativo para consulta sobre una base de datos basada en grafos. Las sentencias en este lenguaje pueden ser ejecutadas a través de JDBC.

- **Capa de Almacenamiento y Servicios:** incluye las aplicaciones y elementos básicos que sirven de plataforma para el prototipo. Esta capa está integrada por los siguientes paquetes:

  - JVM (Java Virtual Machine): Es el entorno en el que son ejecutadas las aplicaciones Java. Define un ordenador abstracto y especifica las instrucciones que este ordenador puede ejecutar.

  - Monitorización de Cultivos: representa el almacenamiento centralizado de la información de monitorización de cultivos (condiciones climáticas y propiedades agronómicas).

  - Repositorio de Grafos: representa el almacenamiento centralizado del Grafo de Datos. Además, soporta su consulta y modificación.

  - Neo4j: es un sistema de administración de bases de datos orientadas a grafos, implementado en Java. Este entorno implementa una persistencia

basada en disco, que almacena datos estructurados en grafos en lugar de tablas. Adicionalmente, contiene un lenguaje propietario para la consulta y modificación de grafos, llamado Cypher.

- PostgreSQL: es un sistema de gestión de bases de datos multiusuario, multiplataforma y de código abierto. Emplea lenguaje SQL para consultas y modificación de registros.

- Debian: es el sistema operativo que soporta el prototipo.

5.2. Experimentación

Para la evaluación experimental del prototipo, fue tomada como punto de referencia la metodología propuesta por Wohlin [90], quien recomienda seguir un conjunto de fases para la planeación del experimento, descritas a continuación.

5.2.1. Definición

El objetivo de la evaluación experimental es comprobar que el sistema experto construido detecte condiciones favorables en los cultivos de café para el desarrollo de una tasa de infección de Roya. Para esto, se hace uso del emparejamiento de patrones en grafo para validar una serie de reglas obtenidas con base en el conocimiento de expertos.

Por otro lado, el contexto de aplicación del prototipo corresponde al cultivo de café en Colombia y una de las enfermedades que más lo afectan, como es la Roya. Esta enfermedad genera pérdidas anuales cercanas al 30% de producción en variedades susceptibles. Como caso específico de validación, fueron usados los datos de monitorización y propiedades de cultivo obtenidos entre 2011 y 2014 en la granja experimental *Los Naranjos*, perteneciente a la empresa Supracafé, la cual se encuentra ubicada en el municipio de Cajibío (Cauca). Adicionalmente, en esta granja se encuentran registros de infección de Roya para algunos meses del periodo mencionado anteriormente, permitiendo la validación de las respuestas dadas por el sistema experto construido.

5.2.2. Criterios de Evaluación

Para la aplicación de las pruebas sobre el prototipo desarrollado, de ser tomada en cuenta la hipótesis que motivó a la construcción de la solución, la cual es: *El emparejamiento de patrones en grafos puede ser usado como método de evaluación de reglas generadas a partir del conocimiento de expertos, dentro de un sistema experto para la detección de condiciones favorables de aparición de enfermedades para el sector agrícola.*

En este sentido, varias medidas de evaluación y desempeño de sistemas expertos han sido definidas y aplicadas en diversas investigaciones [91][92]. Precisamente, para el presente trabajo de grado fueron seleccionadas las siguientes métricas:

- **Validación predictiva:** En esta prueba son usados estudios de caso históricos, a través de registros de la enfermedad en cultivos donde hayan sido monitorizadas variables climáticas y se tenga registro de sus propiedades agronómicas. El sistema es puesto a prueba suministrando la información mencionada anteriormente y comparando su respuesta con el registro histórico existente, en compañía de un experto que supervise el proceso.

- **Oportunidades de mejora:** A medida que los usuarios hagan uso del sistema, cada sección o módulo de este puede ser realimentado con sugerencias hacia la información tratada en cada uno y su interacción con el usuario. Este tipo de recomendaciones por parte del usuario son a menudo capturadas a través de un buzón de sugerencias en cada sección o módulo. De esta manera, esta retroalimentación será tenida en cuenta dentro de la evaluación de la base de conocimiento y la validación predictiva.

- **Evaluación de base de conocimiento**: Este proceso está dirigido a las reglas(patrones) de la base de conocimiento, con el fin de realizar una evaluación, por parte de experto, para comprobar su coherencia y alcance. Cada regla o patrón existente contiene una serie de premisas que deben darse para llegar a una conclusión. En consecuencia, el experto debe partir de cada conclusión y verificar si las premisas asociadas a esta son coherentes y corresponden al alcance que pueda tener la regla; es decir, que la evaluación de las premisas sea suficiente para llegar a dicha conclusión.

Para el caso del modelo generado a través de la metodología CRISP-DM (capítulo 3), sus medidas de evaluación fueron presentadas en la sección 3.6, ya que la revisión de dichas medidas hace parte de la metodología seguida.

Por otro lado, para la evaluación del algoritmo que lleva a cabo el emparejamiento de patrones en grafo propuesto, existen dos criterios a tener en cuenta, siguiendo los experimentos llevados a cabo en [44]:

- Efecto del tamaño de grafo consultado en el tiempo de ejecución.
- Efecto del tamaño del patrón de grafo consultado en el tiempo de ejecución.

La implementación del algoritmo fue hecha en Java, haciendo uso de clientes http para el acceso al Grafo de Datos, almacenado en Neo4j.

5.3. Planificación

Las pruebas de calidad y rendimiento, se realizan sobre el sistema completo, así como en algunos de sus componentes que cumplen las funcionalidades más importantes, como son: Generación de Reglas y Emparejamiento de Patrones en Grafo.

A continuación, en la Tabla 12, se describen las pruebas realizadas según los criterios de evaluación definidos anteriormente:

Ref.	Módulo	Prueba
P-001	Generación de Reglas	Evaluación de base de conocimiento
P-002	Emparejamiento de Patrones en Grafo	Tiempo de ejecución del algoritmo CRD-GPM para diferentes tamaños de grafo consultado
P-003	Emparejamiento de Patrones en Grafo	Tiempo de ejecución del algoritmo CRD-GPM para diferentes tamaños de patrón de grafo buscado
P-004	Sistema Experto	Validación predictiva

Tabla 12. Plan de pruebas

Para las pruebas de tiempos de ejecución del algoritmo, fue usado un ordenador con un procesador Intel Core i5 de doble núcleo a una velocidad de 1,3 GHz y 4GB de memoria RAM DDR3 de 1600 MHz.

5.4. Resultados

A continuación, son presentados los resultados de las pruebas definidas en la Tabla 12.

5.4.1. P-001: Evaluación de la base de conocimiento

La realización de esta prueba contó con el acompañamiento de Álvaro Gaitán Bustamante[2], quien es un investigador científico de Cenicafé (Centro Nacional de Investigación del Café), dentro de la disciplina de fitopatología, y un experto en el estudio de la roya en Colombia.

En primer lugar, fueron presentadas al experto las variables predictivas tomadas en cuenta para la generación del árbol de decisión, las cuales corresponden a factores que afectan el desarrollo y evolución de la roya según diversas investigaciones. Para el experto, la selección de las variables corresponde de manera acertada a factores que juegan un rol importante en las distintas etapas de la enfermedad. Adicionalmente, recomienda la sustitución de la variable SOMBRA, ya que su valor corresponde a un porcentaje estimado y establecido empíricamente. En su lugar, podría tenerse un estimado de las horas durante el día en las que el cultivo tuvo incidencia de brillo solar, lo cual puede ser calculado a partir del dato de radiación que capturan las estaciones meteorológicas. Por otro lado, la

[2] http://scienti.colciencias.gov.co:8081/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000004260

precipitación en horas nocturnas debería ser tomada en cuenta como una variable adicional, ya que lluvias ocurridas después de las 5 de la tarde generan capas de agua en las hojas, generando condiciones óptimas para el desarrollo del hongo acompañado de las condiciones de temperatura, brillo y humedad de las noches. Asimismo, el experto señala que sería importante contar con la fecha exacta en la que se hizo la medición del porcentaje de incidencia de roya en los cultivos. Lo anterior se debe a que el análisis de 30 días antes del día en el que se hizo la medición de la enfermedad determina de manera más exacta las condiciones que llevaron a un porcentaje de infección calculado, en lugar de tomar el mes del calendario gregoriano en el que se hizo dicha medición.

Como paso siguiente, el experto revisó el árbol de decisión generado y los patrones extraídos a partir de éste. Para el experto, el modelo acierta al ubicar como primer variable a evaluar el número de horas nocturnas donde se presentó una humedad relativa mayor a 90% (HORHRN90). El rango de evaluación (menor o igual a 6.35 horas o mayor a este valor) se acerca a los estudios de las horas de mojadura mínimas requeridas para una infección. Además, el experto resalta que la ocurrencia de este fenómeno en horas nocturnas es aún más favorable para la enfermedad, teniendo en cuenta las demás condiciones climáticas en la ausencia de luz solar. Con relación a las demás variables, los rangos de su evaluación están dentro de valores aceptables para la generación de las distintas tasas. Para el caso de la variable SOMBRA, la recomendación es igual a la mencionada anteriormente, donde recomienda su cambio por una estimación de las horas de brillo solar. De todas maneras, el modelo tiene coherencia al predecir la tasa de aumento acelerado de la infección (TI3) con una estrecha relación al porcentaje de sombrío del cultivo, ya que la sombra produce una alteración en temperaturas y humedad del cultivo ubicado debajo de ésta.

Por último, el experto recomienda validar los patrones teniendo en cuenta los comportamientos de la enfermedad en zonas ya diferenciadas del país, como lo ha estudiado Cenicafé. De esta forma, para la zona donde está la hacienda Los Naranjos, lugar donde fue obtenido el conjunto de datos, las epidemias de roya suelen empezar a finales de año y terminan a inicios del año siguiente. Por otro lado, el experto propone que cada día sean obtenidas las variables predictivas de 30 días hacia atrás, a partir del día en curso. Lo anterior, con el propósito de obtener una alerta sobre la posible tasa de infección de la enfermedad de manera diaria, conforme las condiciones climáticas de los últimos 30 días hayan cambiado.

5.4.2. P-002: Tiempo de ejecución del emparejamiento según tamaño de grafo

La medida del tiempo necesario para que el proceso de emparejamiento se lleve a cabo es de importancia para conocer la escalabilidad del mismo. Teniendo en cuenta que el objetivo de esta evaluación es obtener el tiempo de ejecución del algoritmo, fueron creados nodos de tipo *Cultivo* adicionales en el *Grafo de Datos*, con el mismo número de instancias relacionadas al cultivo original para el cual existen los datos de monitorización y

propiedades (Los Naranjos). Lo anterior tiene el propósito de comprobar la escalabilidad del algoritmo conforme el grafo consultado aumenta su tamaño. Adicionalmente, las etiquetas relacionadas a la información contenida en las instancias adicionales creadas fueron generadas de forma aleatoria, con variaciones de hasta el 10% del valor original contenido en las instancias de los Naranjos. De esta manera, en la Figura 13 pueden verse los resultados en tiempo de ejecución para distintos tamaños del Grafo de Datos.

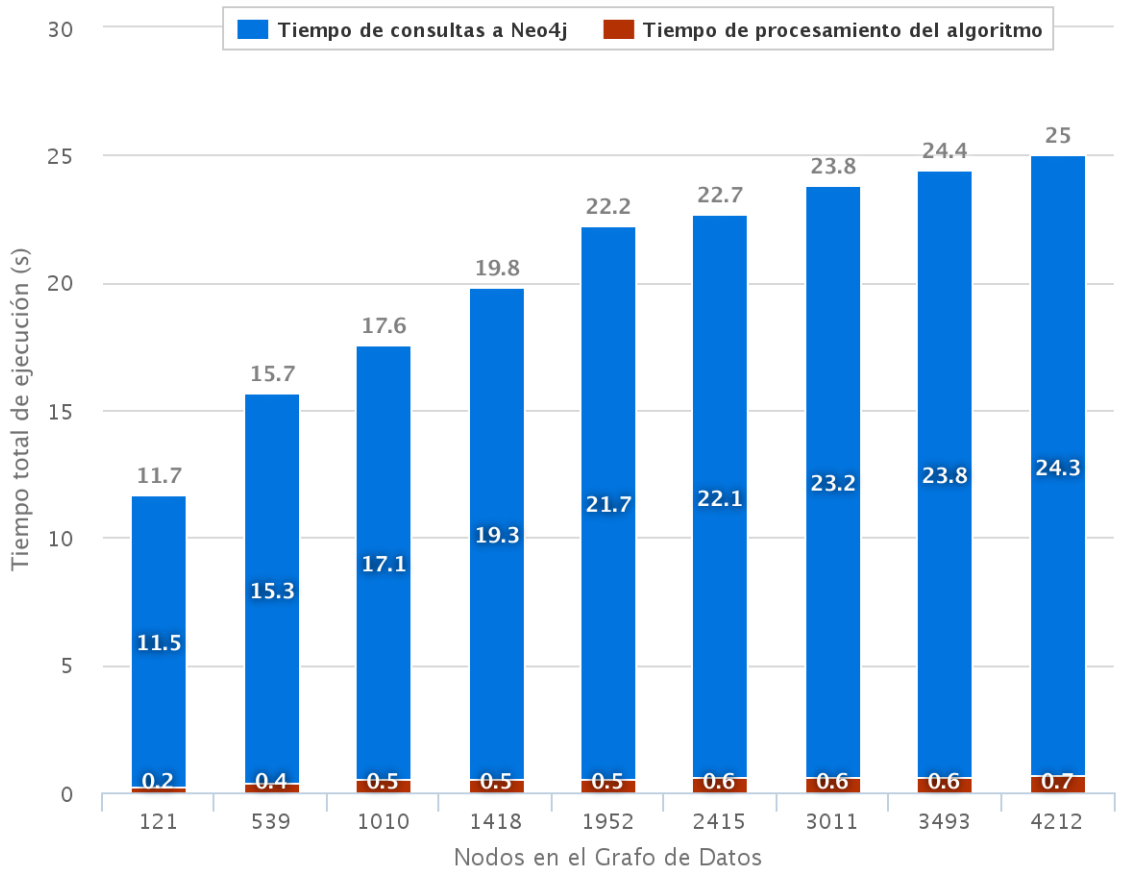


Figura 13. Efecto del tamaño del Grafo de Datos en el tiempo de ejecución del emparejamiento de patrones

El tiempo total de ejecución del algoritmo fue divido en el tiempo de procesamiento de las tareas descritas en el pseudocódigo del mismo (Algoritmo 3, sección 4.3) y el tiempo que toman las consultas a la base de datos implementada en Neo4j. Debido a que el algoritmo fue implementado de forma que no mantenga el grafo consultado en memoria, sino que sea consultado desde una base de datos de forma local o remota, cada ejecución de una sentencia de consulta Cypher requiere el inicio de un cliente http. Ahora bien, dentro de las tareas del algoritmo, existen varios procesos repetitivos, como por ejemplo la evaluación de cada nodo que ha sido seleccionado como candidato para el emparejamiento. Esto hace que los clientes http sean iniciados muchas veces y como resultado el tiempo total de ejecución crece en gran medida. Si bien los tiempos obtenidos presentan un gran valor con

relación a los tamaños del grafo consultado, las tasas de crecimiento tienden a ser menores al aumentar el tamaño del grafo. Lo anterior puede ser visto al analizar el crecimiento de tiempo entre la ejecución para 539 y 2415 nodos (diferencia de 1876 nodos), el cual es 7 segundos de diferencia. Por otro lado, entre 2415 y 4212 nodos (diferencia 1797 nodos), el crecimiento de tiempo de ejecución fue solo de 2.3.

5.4.3. P-003: Tiempo de ejecución del emparejamiento según tamaño de patrón de grafo

Para esta evaluación, fue tomado el grafo de mayor tamaño generado para el anterior criterio presentado. Este grafo contiene 4212 nodos, 7368 etiquetas y 404 relaciones. De esta manera, el algoritmo fue aplicado para patrones de 2 a 8 nodos.

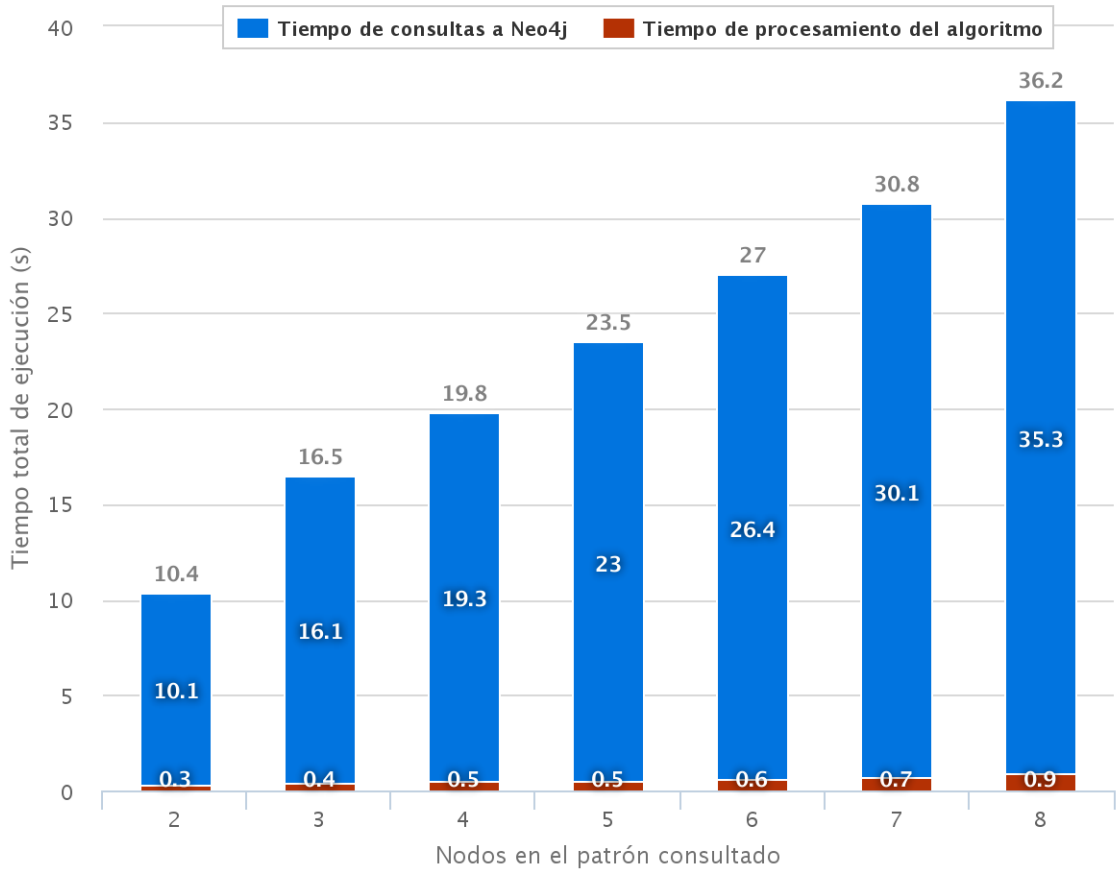


Figura 14. Efecto del tamaño del patrón de grafo en el tiempo de ejecución del emparejamiento de patrones

La Figura 14 muestra los resultados obtenidos para esta evaluación. En ella puede verse la diferencia entre los tiempos de consulta y procesamiento, conforme fue explicado en el criterio anterior. Por otro lado, el tiempo total de ejecución presenta un comportamiento exponencial. Lo anterior puede explicarse debido a que, al incrementar el número de nodos del patrón consultado, se incrementa el número de repeticiones del ciclo entre la línea 21 y

32 del algoritmo propuesto (Algoritmo 3, sección 4.3). Precisamente, dentro de este ciclo se realizan bastantes consultas a la base de datos, ya que se comprueban las relaciones del nodo inicio en profundidad (DFS) y las etiquetas de los nodos candidatos. Estos procesos requieren de un gran número de conexiones del cliente http para enviar las sentencias de Cypher hacia el servidor de Neo4j y, por lo explicado en el anterior criterio, el tiempo se incrementa de forma acelerada.

5.4.4. P-004: Validación predictiva

Esta prueba fue aplicada a los registros históricos contenidos en el conjunto de datos de la hacienda Los Naranjos. Este conjunto de datos pasó a ser representado como un repositorio de información en grafo (Grafo de Datos), siguiendo el proceso descrito en la sección 4.1.1. De esta manera, fueron consultados los 10 patrones de grafo obtenidos mediante la aplicación del algoritmo propuesto (Algoritmo 3 – CRD-GPM). Adicionalmente, tras la aplicación del emparejamiento de patrones, fueron consideradas únicamente las instancias que corresponden a meses donde existe un registro de la evolución de la infección de roya. Lo anterior es debido a que para meses donde no está este registro, no existe la posibilidad de validar si las coincidencias encontradas son correctas. Con base en las anteriores consideraciones, en la Tabla 13 se presentan los resultados del emparejamiento de cada patrón, aplicado al conjunto de datos presentado en la sección 3.3.2 (conjunto de datos sin aplicar el algoritmo de balanceo de clases).

Patrón	Tasa de infección	Coincidencias	Correctas	Incorrectas
1	TI1	10	8	2
2	TI2	2	2	0
3	TI2	8	8	0
4	TI1	42	38	4
5	TI1	12	8	4
6	TI2	10	10	0
7	TI3	4	2	2
8	TI2	12	10	2
9	TI3	16	12	4
10	TI1	8	6	2
	Total	**124**	**104**	**20**

Tabla 13. Resultados de la aplicación del algoritmo CRD-GPM para cada patrón

A partir de los resultados presentados, puede verse que el número de coincidencias encontradas que corresponden correctamente a la tasa de infección que representaba el patrón buscado es 124 (83.87%), mientras el número de coincidencias incorrectas es 20 (16.13%). Estos valores son similares a las medidas de evaluación del modelo presentadas en la sección 3.6.2. Lo anterior se debe a que, precisamente, los patrones fueron extraídos

del árbol de decisión generado, por lo cual, su precisión y acierto van a estar directamente relacionados a las medidas de desempeño del algoritmo de clasificación usado. Adicionalmente, el tipo de emparejamiento de patrones considerado en el algoritmo propuesto corresponde a un emparejamiento exacto, donde no son considerados subgrafos con un grado de proximidad cercano al patrón buscado.

Ahora bien, en los valores obtenidos, la tasa que contiene menos instancias incorrectas en la búsqueda de los patrones que la representan es TI2. Esto puede deberse a que el número de instancias de la clase TI2 en el conjunto de datos es mayor a las de TI1 y TI3, por lo cual, el modelo generado predice de mejor manera esta clase. Lo anterior es conocido como "desbalanceo de clase" y es un problema común en la aplicación de algoritmos de clasificación en conjuntos de datos donde se tienen mayores registros para una clase con relación a las demás [93].

Por otro lado, si bien el porcentaje de coincidencias correctas es alto, es ideal contar con información sobre diferentes cultivos en las zonas cafeteras del país, con el fin de obtener un número mayor de instancias en el Grafo de Datos. De esta manera, la validación de los patrones tendría una mayor generalidad, en lugar de ser específica para un cultivo.

5.5. Resumen

Este capítulo presentó el prototipo desarrollado que consta de un Sistema Experto para la detección de condiciones favorables para la roya en el café, a partir de la búsqueda de patrones de grafo, extraídos a través de la inducción de un árbol de decisión que contiene variables predictivas para la enfermedad, definidas a partir del conocimiento de expertos en el área. De esta manera, el sistema hace uso de bases de datos orientadas a grafos para almacenar un *Grafo de Datos* que contiene la información de monitorización de variables climáticas y propiedades físicas de cultivos de café. Esta base de datos es consultada por un algoritmo de emparejamiento de patrones de grafo, que busca las coincidencias de dichos patrones dentro del *Grafo de Datos.*

Posteriormente, fue presentado el diseño del experimento para validar el prototipo mencionado y la planificación de las pruebas a ser realizadas al sistema y a algunos de sus componentes, como son la evaluación de la base de conocimiento, tiempos de ejecución del algoritmo con relación a tamaño de grafo y tamaño de patrón consultado, y validación predictiva del sistema. Por último, fueron expuestos los resultados de cada prueba realizada, su interpretación e identificando las oportunidades de mejora en las mismas.

6 Conclusiones y trabajos futuros

Este capítulo expone las conclusiones de la presente investigación, así como también los trabajos futuros identificados a través del desarrollo de la misma.

6.1. Conclusiones

Una vez construido el Sistema Experto basado en Emparejamiento de Patrones en Grafos, con el fin de detectar condiciones favorables para la ocurrencia de roya en cultivos de café en Colombia, las siguientes conclusiones fueron obtenidas:

- La consideración del conocimiento de expertos en el estudio de la roya en el café permitió construir variables predictivas más específicas para el problema tratado. Estas variables conforman el conjunto de entrenamiento del algoritmo de clasificación que generó el árbol de decisión y, por ende, las reglas que más adelante fueron expresadas como patrones de grafo. De esta manera, las clases predichas, que en este caso corresponden a las tres tasas de infección definidas, están correlacionadas con características determinantes en los valores de las variables climáticas y agronómicas que presentan los cultivos.

- El conjunto de datos de entrenamiento utilizado para generar el árbol de decisión que permitió extraer las reglas para las distintas tasas de infección de roya, fue construido con 124 instancias de manera inicial. Posteriormente, al aplicar el algoritmo SMOTE para balanceo de clases, el número de instancias fue 161. Esta cantidad de instancias es baja para el entrenamiento de un clasificador, por lo cual el modelo generado no presenta una gran precisión en la predicción de una o más clases. Por lo tanto, en la medida que exista un mayor número de registros y cultivos monitorizados, el modelo puede ser refinado y tener más confiabilidad. Asimismo, la baja cantidad de instancias con las que se contó en el experimento se debe a que la recolección y monitorización de datos en cultivos requieren de un gran esfuerzo en tiempo y dinero.

- El algoritmo de inducción de árboles de decisión usado (C4.5) genera un modelo fácilmente interpretable, donde cada rama del árbol puede ser tomada como una regla para llegar a una clase determinada. De esta forma, cada regla puede expresarse como un patrón de grafo, que contenga en sus nodos una serie de etiquetas, como representación de las premisas de dicha regla.

- En conjuntos de reglas de gran tamaño existen algunos problemas como integridad, reglas conflictivas, duplicación, entre otros; que han sido abordados en algunas investigaciones a partir de representaciones basadas en grafos. Precisamente, los patrones de grafo permiten la representación de una regla y su validación

directamente sobre la estructura y contenido de un grafo de datos. Además, las etiquetas y relaciones entre los nodos generan una integridad en la búsqueda y coherencia en los resultados obtenidos.

- La representación basada en grafos de repositorios de información, permite una gran expresividad e interpretabilidad. Esto ofrece una ventaja para los modeladores de datos, ya que la estructura y organización de las entidades de la base de conocimiento modelada se encuentra explícita en los nodos y sus relaciones. Lo anteriormente mencionado, facilita el entendimiento del esquema de la base de datos y la construcción de consultas para extraer o modificar su información.

- La tarea del emparejamiento de patrones en grafos ha sido una técnica ampliamente abordada a través de los últimos años. El enfoque dirigido a aplicar algoritmos de emparejamiento directamente sobre consultas a la base de datos de grafos sin cargar todo el grafo consultado en memoria, reduce en gran medida los costos computacionales. Sin embargo, este enfoque involucra tiempos adicionales a las instrucciones del algoritmo, generados por el establecimiento de conexión a la base de datos, el envío de la consulta y posterior recibimiento de la respuesta. Estos son procesos repetitivos en diferentes instrucciones del algoritmo de emparejamiento de patrones.

- La validación de las soluciones tecnológicas, aplicadas en dominios como la agricultura, presentan algunas dificultades, como obtención de cantidades suficientes de datos del entorno, calidad en estos datos y colaboración de expertos agrónomos. Si bien, los resultados obtenidos al aplicar el prototipo son buenos, la implementación de este sistema en un entorno de producción necesita de una validación más rígida, a partir de una mayor cantidad de datos para realizar las pruebas y la opinión de más expertos en el problema atacado.

- El tipo de emparejamiento de patrones de grafos abordado fue un emparejamiento exacto. En este escenario, las coincidencias encontradas corresponden de manera rígida a los rangos definidos. De igual manera, estos rangos, que se encuentran expresados como etiquetas en los nodos, pueden cambiar la predicción de una clase u otra con una mínima desviación en el valor que debe ser analizado. Siendo así, el emparejamiento de patrones de tipo inexacto podría considerar un grado de acercamiento de una instancia hacia uno o más patrones consultados, con el fin de determinar porcentajes de acercamiento de un patrón hacia varias clases. Lo anterior, aplicado al sistema desarrollado en este trabajo de grado, permitiría generar y caracterizar de mejor manera alertas para la ocurrencia de roya en el café.

6.2. Trabajos futuros

En el presente trabajo de maestría fue aportada una solución dirigida hacia la detección de condiciones que favorezcan la ocurrencia de una epidemia de roya en cultivos de café en Colombia, a través del uso de un algoritmo de emparejamiento de patrones en grafos. Precisamente, estos patrones fueron extraídos con base en el conocimiento de expertos en la enfermedad y un algoritmo de clasificación. De esta manera, con relación al campo de estudio de esta investigación, son propuestos los siguientes trabajos futuros:

- **Proponer una estructura basada en grafos para el almacenamiento del ambiente de producción agrícola:** Con el propósito de modelar la gran cantidad de datos existentes en los ambientes de producción agrícola, como: datos de monitorización climática, control de cultivos, propiedades agronómicas, entidades responsables, registros de epidemias, propiedades fonológicas, entre otras; es propuesto crear un esquema para el almacenamiento de la información mencionada, basada en grafos. El objetivo es obtener una base de conocimiento de fácil acceso e interpretación, que pueda ser usada para modelar diferentes clases de cultivos.

- **Implementar mejoras en el algoritmo de emparejamiento de patrones en grafos:** A partir de los resultados en las pruebas realizadas, el tiempo de ejecución del algoritmo es el factor que presenta la mayor necesidad de mejora. En particular, el tiempo requerido para realizar las consultas a la base de datos, ya que este representa alrededor del 95% del tiempo total de la tarea de emparejamiento. Para lograr su optimización, podría ser explorado de manera más profunda el lenguaje de consulta del motor para base de datos en grafos Neo4j (Cypher). De esta manera, al mejorar la definición de las consultas en Cypher, que condicionen la búsqueda de manera más estrecha a la necesidad de información, los procesos iterativos en el algoritmo van a verse reducidos.

- **Expansión del prototipo para diferentes cultivos:** Tras el experimento llevado a cabo en la producción de café y una de sus enfermedades, como es la roya, la conceptualización del proceso llevado a cabo para la construcción del prototipo podría ser utilizada en otras enfermedades del café u otros cultivos. Los distintos sectores productivos agrícolas han tomado conciencia de la importancia de contar con plataformas de seguimiento de variables agroclimatológicas, lo cual está reflejado en implementaciones de redes de sensores de monitorización climática y seguimientos continuos de las enfermedades que afectan cada cultivo. De esta manera, en la medida que existan mayores registros de enfermedades en cultivos de diferente naturaleza, la metodología seguida y el sistema propuesto en el presente trabajo de investigación, pueden ser usados como punto de partida para implementar procesos de seguimiento a enfermedades que afectan a cultivos y reducir las pérdidas en su producción.

- **Uso del emparejamiento de patrones en grafos inexacto:** La consideración de un grado de proximidad en la tarea de búsqueda de patrones en grafos para una enfermedad de un cultivo, permitiría generar y caracterizar de mejor manera los niveles de alertas para la ocurrencia de dicha enfermedad. Por otro lado, al incrementar el número de enfermedades analizadas, el emparejamiento de patrones inexacto podría determinar la cercanía entre las condiciones que las definen, lo cual llevaría a una caracterización de las probabilidades de una instancia hacia una u otra enfermedad.

7 Bibliografía

[1] Ministerio de Tecnologías de la Información y las Comunicaciones, “PLAN NACIONAL DE CIENCIA, TECNOLOGÍA E INNOVACIÓN PARA EL DESARROLLO DE LOS SECTORES ELECTRÓNICA, TECNOLOGIAS DE LA INFORMACIÓN Y LAS COMUNICACIONES (ETIC) EN COLOMBIA”, Resumen ejecutivo, 2013.

[2] G. Mansingh, H. Reichgelt, y K.-M. O. Bryson, “CPEST: An expert system for the management of pests and diseases in the Jamaican coffee industry”, *Expert Syst. Appl.*, vol. 32, núm. 1, pp. 184–192, 2007.

[3] D. C. Corrales, I. D. López, F. Campo, S. A. Ordoñez, J. C. Corrales, A. F. Casas, y C. L. Roa, “Plataforma para el seguimiento de variables meteorológicas y ambientales para el sector agropecuario”, *VIII Congr. Ibérico Agroingeniería Cienc. Hortícolas*.

[4] E. Turban y L. E. Frenzel, *Expert systems and applied artificial intelligence*. Prentice Hall Professional Technical Reference, 1992.

[5] S. Dewanto y J. Lukas, “Expert System For Diagnosis Pest And Disease In Fruit Plants”, en *EPJ Web of Conferences*, 2014, vol. 68, p. 00024.

[6] J. R. Quinlan, “Decision trees and decision-making”, *Syst. Man Cybern. IEEE Trans. On*, vol. 20, núm. 2, pp. 339–346, 1990.

[7] L. Rokach, *Data mining with decision trees: theory and applications*, vol. 69. World scientific, 2008.

[8] M. Mehta, R. Agrawal, y J. Rissanen, “SLIQ: A fast scalable classifier for data mining”, en *Advances in Database Technology—EDBT’96*, Springer, 1996, pp. 18–32.

[9] J. A. Bondy y U. S. R. Murty, *Graph theory with applications*, vol. 290. Macmillan London, 1976.

[10] R. A. Hanneman y M. Riddle, *Introduction to social network methods*. University of California Riverside, 2005.

[11] D. J. Cook y L. B. Holder, *Mining graph data*. John Wiley & Sons, 2006.

[12] C. C. Aggarwal y H. Wang, “An Introduction to Graph Data”, en *Managing and Mining Graph Data*, Springer, 2010, pp. 1–11.

[13] X. Wang, “Graph pattern matching on social network analysis”, University of Edinburgh, 2013.

[14] H. Bunke y M. Neuhaus, “Graph matching. exact and error-tolerant methods and the automatic learning of edit costs”, *Min. Graph Data*, pp. 17–32, 2007.

[15] W. A. Derwin Suhartono, M. Lestari, y M. Yasin, “Expert System in Detecting Coffee Plant Diseases”, *Int. J. Electr. Energy*, vol. 1, núm. 3, pp. 156–162, 2013.

[16] S. Kaloudis, D. Anastopoulos, C. P. Yialouris, N. A. Lorentzos, y A. B. Sideridis, “Insect identification expert system for forest protection”, *Expert Syst. Appl.*, vol. 28, núm. 3, pp. 445–452, 2005.

[17] R. Prasad, K. R. Ranjan, y A. K. Sinha, “AMRAPALIKA: An expert system for the diagnosis of pests, diseases, and disorders in Indian mango”, *Knowl.-Based Syst.*, vol. 19, núm. 1, pp. 9–21, 2006.

[18] S. K. Sarma, K. R. Singh, y A. Singh, “An Expert System for diagnosis of diseases in Rice Plant”, *Int. J. Artif. Intell.*, vol. 1, núm. 1, pp. 26–31, 2010.

[19] B. D. Mahaman, H. C. Passam, A. B. Sideridis, y C. P. Yialouris, “DIARES-IPM: a diagnostic advisory rule-based expert system for integrated pest management in< i> Solanaceous</i> crop systems”, *Agric. Syst.*, vol. 76, núm. 3, pp. 1119–1135, 2003.

[20] J. L. Gonzalez-Andujar, “Expert system for pests, diseases and weeds identification in olive

crops", *Expert Syst. Appl.*, vol. 36, núm. 2, pp. 3278–3283, 2009.
[21] V. López-Morales, O. López-Ortega, J. Ramos-Fernández, y L. B. Muñoz, "JAPIEST: An integral intelligent system for the diagnosis and control of tomatoes diseases and pests in hydroponic greenhouses", *Expert Syst. Appl.*, vol. 35, núm. 4, pp. 1506–1512, 2008.
[22] R. Jain y others, "PulsExpert: An expert system for the diagnosis and control of diseases in pulse crops", *Expert Syst. Appl.*, vol. 38, núm. 9, pp. 11463–11471, 2011.
[23] V. Rossi, P. Meriggi, T. Caffi, S. Giosué, y T. Bettati, "A Web-based Decision Support System for Managing Durum Wheat Crops", 2010.
[24] R. F. Chevalier, G. Hoogenboom, R. W. McClendon, y J. O. Paz, "A web-based fuzzy expert system for frost warnings in horticultural crops", *Environ. Model. Softw.*, vol. 35, pp. 84–91, 2012.
[25] M. E. Cintra, C. A. A. Meira, M. C. Monard, H. A. Camargo, y L. H. A. Rodrigues, "The use of fuzzy decision trees for coffee rust warning in Brazilian crops", en *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, 2011, pp. 1347–1352.
[26] C. A. Meira, L. H. Rodrigues, y S. A. Moraes, "Análise da epidemia da ferrugem do cafeeiro com árvore de decisão", *Trop. Plant Pathol.*, vol. 33, núm. 2, pp. 114–124, 2008.
[27] C. A. A. Meira y L. H. A. Rodrigues, "ÁRVORE DE DECISÃO NA ANÁLISE DE EPIDEMIAS DA FERRUGEM DO CAFEEIRO", 2009.
[28] H. Jiawei y M. Kamber, "Data mining: concepts and techniques", *San Franc. CA Itd Morgan Kaufmann*, vol. 5, 2001.
[29] R. Jain, S. Minz, V. Ramasubramanian, y others, "Machine learning for forewarning crop diseases", *J Ind Soc Agril Stat.*, vol. 63, núm. 1, pp. 97–107, 2009.
[30] W. Geamsakul, T. Yoshida, K. Ohara, H. Motoda, T. Washio, H. Yokoi, y K. Takabayashi, "Extracting diagnostic knowledge from hepatitis dataset by decision tree graph-based induction", en *Active Mining*, Springer, 2005, pp. 126–151.
[31] K. Ohara, T. Yoshida, W. Geamsakul, H. Motoda, T. Washio, H. Yokoi, y K. Takabayashi, "Analysis of Hepatitis Dataset by Decision Tree Graph-Based Induction", *Proc. Discov. Chall.*, pp. 173–184, 2004.
[32] W. Geamsakul, T. Yoshida, K. Ohara, H. Motoda, H. Yokoi, y K. Takabayashi, "Constructing a decision tree for graph-structured data and its applications", *Fundam. Informaticae*, vol. 66, núm. 1, pp. 131–160, 2005.
[33] C. Irniger y H. Bunke, "Graph database filtering using decision trees", en *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, vol. 3, pp. 383–388.
[34] W. Fan, X. Wang, y Y. Wu, "ExpFinder: Finding experts by graph pattern matching", en *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, 2013, pp. 1316–1319.
[35] W. Fan, X. Wang, y Y. Wu, "Diversified top-k graph pattern matching", *Proc. VLDB Endow.*, vol. 6, núm. 13, pp. 1510–1521, 2013.
[36] S. Ma, Y. Cao, W. Fan, J. Huai, y T. Wo, "Capturing topology in graph pattern matching", *Proc. VLDB Endow.*, vol. 5, núm. 4, pp. 310–321, 2011.
[37] K. Ogaard, H. Roy, S. Kase, R. Nagi, K. Sambhoos, y M. Sudit, "Discovering patterns in social networks with graph matching algorithms", en *Social Computing, Behavioral-Cultural Modeling and Prediction*, Springer, 2013, pp. 341–349.
[38] K. P. Sambhoos, *Graph matching applications in high level information fusion*. ProQuest, 2007.
[39] Y. Bai, C. Wang, Y. Ning, H. Wu, y H. Wang, "G-path: flexible path pattern query on large graphs", en *Proceedings of the 22nd international conference on World Wide Web companion*, 2013, pp. 333–336.
[40] M. U. Nisar, A. Fard, y J. A. Miller, "Techniques for graph analytics on big data", en *Big Data (BigData Congress), 2013 IEEE International Congress on*, 2013, pp. 255–262.

[41] L. P. Cordella, P. Foggia, C. Sansone, y M. Vento, “A (sub) graph isomorphism algorithm for matching large graphs”, *Pattern Anal. Mach. Intell. IEEE Trans. On*, vol. 26, núm. 10, pp. 1367–1372, 2004.
[42] L. P. Cordella, P. Foggia, C. Sansone, y M. Vento, “Performance evaluation of the VF graph matching algorithm”, en *Image Analysis and Processing, 1999. Proceedings. International Conference on*, 1999, pp. 1172–1177.
[43] J. R. Ullmann, “An algorithm for subgraph isomorphism”, *J. ACM JACM*, vol. 23, núm. 1, pp. 31–42, 1976.
[44] M. Saltz, A. Jain, A. Kothari, A. Fard, J. A. Miller, y L. Ramaswamy, “Dualiso: An algorithm for subgraph pattern matching on very large labeled graphs”, en *Big Data (BigData Congress), 2014 IEEE International Congress on*, 2014, pp. 498–505.
[45] P. Barceló, L. Libkin, y J. L. Reutter, “Querying regular graph patterns”, *J. ACM JACM*, vol. 61, núm. 1, p. 8, 2014.
[46] X. Yan y J. Han, “CloseGraph: mining closed frequent graph patterns”, en *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 286–295.
[47] G. Kollias, M. Sathe, O. Schenk, y A. Grama, “Fast parallel algorithms for graph similarity and matching”, *J. Parallel Distrib. Comput.*, vol. 74, núm. 5, pp. 2400–2410, 2014.
[48] B. T. Messmer y H. Bunke, “A decision tree approach to graph and subgraph isomorphism detection”, *Pattern Recognit.*, vol. 32, núm. 12, pp. 1979–1998, dic. 1999.
[49] C. M. Hoffmann y M. J. o’DONNELL, “Pattern matching in trees”, *J. ACM JACM*, vol. 29, núm. 1, pp. 68–95, 1982.
[50] J. Oliver, *Decision graphs: an extension of decision trees*. Citeseer, 1992.
[51] U. Fayyad, G. Piatetsky-Shapiro, y P. Smyth, “From data mining to knowledge discovery in databases”, *AI Mag.*, vol. 17, núm. 3, p. 37, 1996.
[52] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, y R. Wirth, “CRISP-DM 1.0 Step-by-step data mining guide”, 2000.
[53] C. Rivillas, C. Serna, M. Cristancho, y A. Gaitán, “Roya del Cafeto en Colombia: Impacto, Manejo y Costos del Control”, *Chinchiná Bol. Téc.*, núm. 36, 2011.
[54] J. M. Waller, M. Bigger, y R. J. Hillocks, *Coffee pests, diseases and their management*. CABI, 2007.
[55] J. Avelino, L. Willocquet, y S. Savary, “Effects of crop management patterns on coffee rust epidemics”, *Plant Pathol.*, vol. 53, núm. 5, pp. 541–547, 2004.
[56] F. J. Nutman, F. M. Roberts, y R. T. Clarke, “Studies on the biology of Hemileia vastatrix Berk. & Br”, *Trans. Br. Mycol. Soc.*, vol. 46, núm. 1, pp. 27–44, 1963.
[57] L. Zambolim, F. do Vale, H. Costa, A. A. Pereira, G. M. Chaves, y L. Zambolim, “Epidemiologia e controle integrado da ferrugem-do-cafeeiro”, *ZAMBOLIM O Estado Arte Tecnol. Na Produção Café Viçosa Suprema Gráfica E Ed.*, pp. 369–449, 2002.
[58] R. H. Montoya y G. M. Chaves, “Influencia da temperatura e da luz na germinacao, infectividade e periodo de geracao de Hemileia vastatrix Berk. y Br.”, *Experientiae*, 1974.
[59] E. J. De Jong, A. B. Eskes, J. G. J. Hoogstraten, y J. C. Zadoks, “Temperature requirements for germination, germ tube growth and appressorium formation of urediospores of Hemileia vastatrix”, *Neth. J. Plant Pathol.*, vol. 93, núm. 2, pp. 61–71, 1987.
[60] S. Becker, “Diurnal periodicity in spore dispersal of Hemileia vastatrix in relation to weather factors”, *Z. Pflanzenkrankh. Pflanzenschutz*, vol. 84, núm. 10, pp. 577–591, 1977.
[61] K. R. Bock, “Dispersal of uredospores of Hemileia vastatrix under field conditions”, *Trans. Br. Mycol. Soc.*, vol. 45, núm. 1, pp. 63–74, 1962.
[62] A. C. Kushalappa y A. B. Eskes, *Coffee rust: epidemiology, resistance, and management*.

CRC Press, 1989.
[63] A. C. Kushalappa, M. Akutsu, y A. Ludwig, “Application of survival ratio for monocyclic process of Hemileia vastatrix in predicting coffee rust infection rates”, *Phytopathology*, vol. 73, núm. 1, pp. 96–103, 1983.
[64] J. C. Sutton, T. J. Gillespie, y P. D. Hildebrand, “Monitoring weather factors in relation to plant disease [Crop microclimate, electrical sensors, temperature and wetness gauges, sources of error].”, *Plant Dis.*, 1984.
[65] R. E. Jensen y L. W. Boyle, “A technique for forecasting leafspot on peanuts”, *Plant Rep*, vol. 50, núm. 11, pp. 810–814, 1966.
[66] R. A. Muller, D. Berry, J. Avelino, y D. Bieysse, “Coffee diseases”, *Coffee Grow. Process. Sustain. Prod. Guideb. Grow. Process. Traders Res.*, pp. 491–545, 2004.
[67] P. J. Arcila, V. F. Farfán, A. B. Moreno, L. F. Salazar, y E. Hincapié, “Sistemas de producción de café en Colombia”, *Blanocolor Chinchiná Colomb.*, 2007.
[68] A. C. Kushalappa, M. Akutsu, S. H. Oseguera, G. M. Chaves, C. A. Melles, J. M. Miranda, y G. F. Bartolo, “Equations for predicting the rate of coffee rust development based on net survival ratio for monocyclic process of Hemileia vastatrix [Coffea arabica]”, *Fitopatol. Bras. Braz.*, 1984.
[69] J. Avelino, H. Zelaya, A. Merlo, A. Pineda, M. Ordoñez, y S. Savary, “The intensity of a coffee rust epidemic is dependent on production situations”, *Ecol. Model.*, vol. 197, núm. 3, pp. 431–447, 2006.
[70] C. A. A. Meira, L. H. A. Rodrigues, y S. A. Moraes, “Analysis of coffee leaf rust epidemics with decision tree”, *Trop. Plant Pathol.*, vol. 33, núm. 2, pp. 114–124, abr. 2008.
[71] D. C. Corrales, A. Ledezma, A. J. Peña, J. Hoyos, A. Figueroa, y J. C. Corrales, “Un nuevo conjunto de datos para la detección de roya en cultivos de café Colombianos basado en clasificadores”, *Sist. Telemática*, vol. 12, núm. 29, pp. 9–23, 2014.
[72] V. P. Campos, F. X. R. Vale, y L. Zambolim, “Café (Coffea arabica L.) Controle de doenças. Doenças causadas por nematóides”, *Controle Doenças Plantas Gd. Cult. Viçosa UFV*, vol. 1, pp. 141–180, 1997.
[73] S. de Moraes, M. H. Sugimori, I. J. A. Ribeiro, A. A. Ortolani, y M. J. Pedro Junior, “Período de incubação de Hemileia vastatrix Berk. et Br. em três regiões do Estado de São Paulo”, *Summa Phytopathol.*, vol. 2, núm. 1, pp. 32–38, 1976.
[74] N. V. Chawla, K. W. Bowyer, L. O. Hall, y W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique”, *J. Artif. Intell. Res.*, vol. 16, núm. 1, pp. 321–357, 2002.
[75] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, y I. H. Witten, “The WEKA data mining software: an update”, *ACM SIGKDD Explor. Newsl.*, vol. 11, núm. 1, pp. 10–18, 2009.
[76] J. R. Quinlan, “C4. 5: Programming for machine learning”, *Morgan Kauffmann*, 1993.
[77] S. Ruggieri, “Efficient C4. 5 [classification algorithm]”, *Knowl. Data Eng. IEEE Trans. On*, vol. 14, núm. 2, pp. 438–444, 2002.
[78] S. J. Cunningham y G. Holmes, “Developing innovative applications in agriculture using data mining”, en *The proceedings of the southeast asia regional computer confederation conference*, 1999.
[79] J. Chuang, “Agreement between categorical measurements: Kappa statistics”, *Retirado Em*, vol. 15, núm. 11, p. 2004, 2001.
[80] R. J. Hyndman y A. B. Koehler, “Another look at measures of forecast accuracy”, *Int. J. Forecast.*, vol. 22, núm. 4, pp. 679–688, 2006.
[81] J. S. Armstrong y F. Collopy, “Error measures for generalizing about forecasting methods: Empirical comparisons”, *Int. J. Forecast.*, vol. 8, núm. 1, pp. 69–80, 1992.
[82] S. V. Stehman, “Selecting and interpreting measures of thematic classification accuracy”, *Remote Sens. Environ.*, vol. 62, núm. 1, pp. 77–89, 1997.

[83] G. Salton y M. J. McGill, “Introduction to modern information retrieval”, 1986.
[84] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Newton, MA, USA: Butterworth-Heinemann, 1979.
[85] I. Robinson, J. Webber, y E. Eifrem, *Graph databases*. O’Reilly Media, Inc., 2013.
[86] W. Fan, J. Li, S. Ma, N. Tang, Y. Wu, y Y. Wu, “Graph pattern matching: from intractable to polynomial time”, *Proc. VLDB Endow.*, vol. 3, núm. 1–2, pp. 264–275, 2010.
[87] M. W. Saltz, “A fast algorithm for subgraph pattern matching on large labeled graphs”, University of Georgia, 2013.
[88] “Neo4j, the World’s Leading Graph Database”, *Neo4j Graph Database*. [En línea]. Disponible en: http://neo4j.com/. [Consultado: 09-jul-2015].
[89] I. Jacobson, G. Booch, J. Rumbaugh, J. Rumbaugh, y G. Booch, *The unified software development process*, vol. 1. Addison-wesley Reading, 1999.
[90] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, y A. Wesslén, *Experimentation in software engineering*. Springer, 2012.
[91] P. D. Grogono, A. D. Preece, R. Shinghal, y C. Y. Suen, “A review of expert systems evaluation techniques”, en *Workshop on Validation and Verification of Knowledge-Based Systems*, 1993, pp. 120–125.
[92] R. M. O’Keefe y D. E. O’Leary, “Expert system verification and validation: a survey and tutorial”, *Artif. Intell. Rev.*, vol. 7, núm. 1, pp. 3–42, 1993.
[93] N. V. Chawla, N. Japkowicz, y A. Kotcz, “Editorial: special issue on learning from imbalanced data sets”, *ACM Sigkdd Explor. Newsl.*, vol. 6, núm. 1, pp. 1–6, 2004.