



TEC

Tecnológico de Costa Rica

ESCUELA DE INGENIERÍA EN COMPUTACIÓN

PROGRAMA DE MAESTRÍA EN COMPUTACIÓN

**Evaluación del uso de Redes
Bayesianas Dinámicas para la
predicción del avance de la Sigatoka
Negra y la productividad en
cultivos agrícolas**

Propuesta de tesis
para el grado de

Magister Scientiæ en Ciencias de la computación

Autor
Sebastián Argüello

Asesor
Luis Alexander Calvo Valverde.

septiembre 2017

Resumen

La enfermedad Sigatoka negra afecta a los cultivos del banano y puede generar pérdidas en la producción. La aplicación de fungicidas para combatirla representa no sólo un costo económico significativo para la industria, sino también un problema para el medio ambiente y la salud del personal de las plantaciones.

La Corporación Bananera Nacional (CORBANA) cuenta con datos históricos de información meteorológica, la producción y el avance de la enfermedad. Con el fin de tomar decisiones de manera anticipada con respecto a la aplicación de los fungicidas y la producción, CORBANA requiere generar predicciones a partir de estos datos. Para ello se decidió evaluar la capacidad de predicción de las Redes Bayesianas Dinámicas (RBDs) y las Redes Bayesianas no dinámicas (RBs) como modelos de referencia.

En la presente investigación se utilizan los datos mencionados y se modelan tanto la productividad como el avance de la enfermedad usando las RBDs y las RBs. Se compara la capacidad de predicción de ambos tipos de redes utilizando los datos de CORBANA y se determina que no existe una diferencia significativa entre ambas.

Abstract

The Black Sigatoka affects the banana crops and can cause losses in the production. The use of fungicides to control it affects the environment and the health of the farm workers.

Corporación Bananera Nacional (CORBANA) have historical data of meteorological information, production, and the state of the disease. In order to take anticipated decisions with regards to the use fungicides and the production, CORBANA requires to produce predictions based on this data. In order to do so, the Dynamic Bayesian Networks (DBNs) and the non-dynamic Bayesian Networks (BNs) were chosen as reference models.

In the present investigation the mentioned data is used for modeling both the productivity and the state of the disease using the DBNs and the BNs. The capacity of prediction of both networks is compared using CORBANA's data and it is determined that there is not a significant different between them.

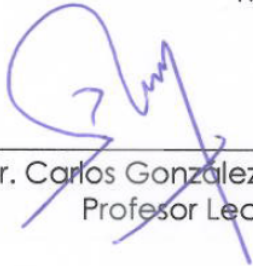
APROBACIÓN DE LA TESIS

“Evaluación del uso de Redes Bayesianas Dinámicas para la predicción del avance de la Sigatoka Negra y la productividad en cultivos agrícola”

TRIBUNAL EXAMINADOR



Máster Luis Alexander Calvo Valverde
Profesor Asesor



Dr. Carlos González Alvarado
Profesor Lector



Máster Nelson José Acuña Alpizar
Profesor Externo



Dr. Roberto Cortés Morales
Coordinador del Programa
de Maestría en Computación

TEC | Tecnológico
de Costa Rica
Maestría en Computación

Agosto, 2017

Agradecimientos

Aprovecho este espacio para agradecer a las personas que me brindaron su apoyo durante esta investigación.

A mi familia que siempre está ahí para mí cuando los necesito y quienes me han apoyado incondicionalmente en mis proyectos. A mis padres por todas las oportunidades que me dieron y enseñarme todo lo que me han enseñado. He logrado llegar a donde estoy gracias a ustedes. A mis hermanas por siempre estar ahí a pesar de las distancias y sobre todo por inspirarme a apuntar alto.

A todos mis amigos que de una u otra manera me apoyaron. Por tenerme paciencia por casi no verlos durante este tiempo. Por estar al tanto de mí y de mi investigación, y ofrecerme ayuda aún cuando no entendieran lo que estaba haciendo. Por la *compañía cibernética* durante las largas noches de trabajo que me dieron los ánimos y motivación que me dejó llegar al final.

Y un especial agradecimiento para el profesor Luis Alexander Calvo por todo el tiempo que dedicó para ayudarme y guiarme para tratar este tema. Así como por todo el apoyo, ánimo y motivación que me dio durante toda la investigación.

¡Gracias a todos!

Contenido

Acrónimos	x
1 Introducción	1
1.1 Área temática	3
1.2 Justificación del tema	3
1.3 Problema	4
2 Marco teórico	5
2.1 Antecedentes teóricos	6
2.1.1 Sistemas de alerta temprana	6
2.1.2 Aprendizaje de Máquina	6
2.1.3 Tipos de aprendizaje	7
2.1.4 Modelos Gráficos Probabilísticos	8
2.1.5 Tipos de Modelos Gráficos Probabilísticos	10
3 Descripción general de la investigación	15
3.1 Planteamiento del problema	16
3.2 Trabajos Relacionados	16
3.3 Hipótesis	17
3.4 Objetivos	17
3.4.1 Objetivo general	17
3.4.2 Objetivos específicos	18
3.5 Resumen del Experimento	18
3.5.1 Factores	19
3.5.2 Variables de respuesta	19
3.5.3 Prueba de hipótesis	22
3.6 Detalles del Experimento	23
3.6.1 Ambiente de desarrollo	23
3.6.2 Descripción de los factores	24
3.6.3 Diseño e implementación de la Red Bayesiana Dinámica	27

3.6.4	Diseño e implementación de la Red Bayesiana	29
3.7	Alcances y limitaciones	32
4	Resultados	34
4.1	Resumen de resultados	35
4.1.1	Análisis estadístico de los resultados	37
4.1.2	F1 Micro	38
4.1.3	F1 Macro	42
4.2	Análisis de resultados	49
4.2.1	Tipo de red	49
4.2.2	Conjunto de datos	49
4.2.3	Tamaño del nodo	50
4.2.4	Tamaño del slice	50
4.3	Verificación del modelo	51
5	Conclusiones y recomendaciones	54
5.1	Conclusiones	55
5.1.1	Recomendaciones	56
5.1.2	Trabajo futuro	56
A	Resultados complementarios	58
B	Recursos en línea	62
C	Verificación del modelo	64
	Bibliografía	115

Lista de figuras

2.1	Ejemplo de un modelo gráfico probabilístico.	9
2.2	RBD para monitoreo de vehículos: a) la 2TBN; b) la red en el tiempo 0; c) la RBD resultante sobre 3 fragmentos de tiempo.	13
3.1	Diseño de la Red Bayesiana Dinámica.	28
3.2	Ejemplo del Diseño de la Red Bayesiana desenrollada para un periodo de dos semanas.	31
4.1	Gráfico cuantil-cuantil del F1 micro.	39
4.2	Gráfico de las medias del F1 micro por tipo de red.	40
4.3	Gráfico de las medias del F1 micro por conjunto de datos.	41
4.4	Gráfico de las medias del F1 micro por tamaño de nodo.	42
4.5	Gráfico de las medias del F1 micro por tamaño de slice.	43
4.6	Gráfico cuantil-cuantil del F1 macro.	44
4.7	Gráfico de las medias del F1 micro por tipo de red.	45
4.8	Gráfico de las medias del F1 macro por conjunto de datos.	46
4.9	Gráfico de las medias del F1 macro por tamaño de nodo.	47
4.10	Gráfico de las medias del F1 macro por tamaño de slice.	48
4.11	F1 micro promedio utilizando datos del Estado de Evolución. El eje X tiene la cantidad de veces que se repitieron los datos en el archivo de entrada.	52
4.12	F1 macro promedio utilizando datos del Estado de Evolución. El eje X tiene la cantidad de veces que se repitieron los datos en el archivo de entrada.	53
A.1	Gráfico del tiempo requerido para entrenar y evaluar la RBD por tamaño de slice. Datos sintéticos, nodos tamaño 3.	59
A.2	Gráfico de las medias del F1 micro y macro utilizando el conjunto de datos del Estado de la enfermedad y nodos de tamaño 3.	60
A.3	Gráfico de las medias del F1 micro y macro utilizando el conjunto de datos del Estado de la enfermedad y nodos de tamaño 5.	60

A.4	Gráfico de las medias del F1 micro y macro utilizando el conjunto de datos de la Producción y nodos de tamaño 3.	61
A.5	Gráfico de las medias del F1 micro y macro utilizando el conjunto de datos de la Producción y nodos de tamaño 5.	61

Lista de tablas

3.1	Evaluación de clasificación para la etiqueta A.	21
4.1	Resumen de los resultados de los experimentos	37
4.2	Promedio por conjunto de datos.	49
A.01	Resultados de los experimentos	59
C.01	Datos del avance de la enfermedad utilizados para la verificación del modelo.	65
C.02	Datos sintéticos utilizados para la verificación del modelo.	93

Acrónimos

2TBN Red Bayesiana de dos pasos representa la topología de una Red Bayesiana para dos instantes de tiempo.

MOM Modelos Ocultos de Márkov es un modelo que representa variables aleatorias y las relaciones entre ellas utilizando únicamente una variable estado que resume todos los posibles estados del sistema.

MGP Modelos Gráficos Probabilísticos modelos probabilísticos construidos utilizando un grafo.

RB Red Bayesiana es un modelo que representa variables aleatorias y las relaciones entre ellas utilizando un grafo acíclico dirigido

RBD Red Bayesiana Dinámica es la red resultante de desenrollar una Red Bayesiana en el tiempo, utilizando una 2TBN

RM Red de Márkov es un modelo que representa variables aleatorias y las relaciones entre ellas utilizando un grafo no dirigido

Capítulo 1

Introducción

La enfermedad denominada Sigatoka negra es ocasionada por el hongo *Mycosphaerella fijiensis*, el cual crece en las hojas del banano. En los años sesentas, esta enfermedad comenzó a esparcirse en América, y llegó a Costa Rica para finales de los setenta. Esta enfermedad se caracteriza al inicio como unas manchas negras sobre las hojas y, conforme se desarrolla, el hongo se extiende hasta formar líneas y finalmente oscurecer áreas completas de las hojas [16].

El oscurecimiento de las hojas reduce la capacidad de la planta para poder realizar la fotosíntesis con normalidad. Esto impacta no sólo el desarrollo de la planta, sino que además hace que los frutos se maduren prematuramente, generando pérdidas en la producción [16].

Esta enfermedad es controlada aplicando fungicidas, en algunos casos inclusive hasta 50 aplicaciones en un año. El costo de las aplicaciones del fungicida puede llegar a ser hasta un 40% del costo total de producción, y además puede afectar la salud del personal de la plantación [21].

Un mecanismo que pueda dar información de cómo se va a comportar la enfermedad tendría mucho valor para la industria bananera. Esta podría utilizarse para tomar decisiones y programar la aplicación de los fungicidas.

En el área de la Inteligencia artificial, existen diversos algoritmos que se utilizan para la predicción de fenómenos del mundo real. Entre ellos están los Modelo Gráficos Probabilísticos (MGPs). Estos algoritmos se caracterizan por codificar la información utilizando grafos. Esto permite modelar las relaciones que tienen los distintos componentes que conforman un fenómeno particular.

En esta tesis se estudia la aplicación del MGP llamado Redes Bayesianas Dinámicas (RBD), para la predicción del avance de la enfermedad en los cultivos. Así como también para la predicción de la producción de los mismos. Este modelo fue seleccionado ya que con él se pueden representar las relaciones entre las variables que influyen en la productividad y el avance de la enfermedad, y además permite modelar la evolución del

modelo a través del tiempo.

1.1 Área temática

Ciencias de la Computación → Inteligencia Artificial → Aprendizaje de Máquina → Modelos Gráficos Probabilísticos → Redes Bayesianas Dinámicas.

Esta tesis se enmarca dentro del área de Aprendizaje de Máquina. Se aplicarán técnicas específicas de Modelos Gráficos Probabilísticos sobre datos de producción y el avance de la enfermedad en cultivos agrícolas.

1.2 Justificación del tema

La Corporación Bananera Nacional (CORBANA)[6] es una entidad pública no estatal costarricense que desarrolla programas de investigación en torno al cultivo de banano. CORBANA[6] cuenta con los datos históricos de distintas variables meteorológicas, de la producción y del nivel de avance de la enfermedad Sigatoka Negra. Esta organización requiere obtener una predicción tanto de la productividad como del avance de la enfermedad.

Combatir esta enfermedad tiene un alto costo para la industria bananera. En el año 2013 el Departamento de Agricultura de los Estados Unidos estimó el costo de las aplicaciones del fungicida en un valor cercano al 40% del costo total de producción [21]. Además, estos fungicidas pueden afectar la salud del personal de la plantación [21]. Es por esto que contar con una herramienta que permita predecir el avance de la enfermedad sería de gran ayuda para el sector productor de este cultivo. La misma podría ser utilizada tanto para la toma de decisiones sobre la frecuencia y selección del área de fumigación, como para la estimación de la producción.

Las RBD se caracterizan por ser un sistema estacionario homogéneo. Lo que implica que sus variables solamente dependen del estado de la variable en el estado anterior, y

además que el modelo no cambia a través del tiempo [15]. Estas características permiten que con ellas se pueden representar, de una manera compacta, grandes cantidades de información de muchas variables. Por esta razón se eligió este modelo para representar los datos de CORBANA[6], y para desarrollar la herramienta que permita predecir la producción y el avance de la enfermedad. Esta herramienta será utilizada para la creación de un sistema de alerta temprana, esto con el fin facilitar la toma de decisiones en temas afines al cultivo y al manejo de la enfermedad. Por ejemplo se podría decidir cuándo o en qué medida aplicar fungicidas.

1.3 Problema

Los costos de combatir la enfermedad son altos para la industria. Dado que se cuenta con los datos históricos de las distintas variables meteorológicas, de la producción y del nivel de avance de la enfermedad Sigatoka Negra, se quiere crear una herramienta que permita la predicción de la Sigatoka Negra y la productividad del cultivo.

Debido a su capacidad de representar datos temporales de manera compacta, es decir utilizando un número de variables menor que modelos equivalente como los MOM, se utilizarán las RBDs para resolver este problema.

Se quiere utilizar los datos históricos con los que se cuenta para crear una herramienta que permite advertir sobre los cambios en la producción del cultivo y en el avance de la enfermedad. Esto con el fin de poder reducir riesgos y tomar decisiones acerca del cultivo y los mecanismos que se utilizan para detener el avance de la enfermedad.

Marco teórico

2.1 Antecedentes teóricos

En este proyecto de tesis se incluyen antecedentes de temas de aprendizaje de máquina, específicamente el tema de Modelos Gráficos Probabilísticos. Dentro de este tema se estudiarán las Redes Bayesianas y las Redes Bayesianas Dinámicas.

2.1.1 Sistemas de alerta temprana

Los sistemas de alerta temprana son herramientas técnicas que se utilizan para la reducción de riesgos; esto con el objetivo de proteger a las personas y sus medios de vida expuestas a peligros y en preparación ante desastres (Cruz Roja Paraguaya, citado en [8]).

Utilizando las técnicas de aprendizaje de máquina que serán presentados en el presente documento, se pretende crear una herramienta que facilite la toma de decisiones y reducción de riesgos en temas relacionados a la productividad del cultivo y al avance de la enfermedad.

2.1.2 Aprendizaje de Máquina

El Aprendizaje de Máquina es una rama de la Inteligencia Artificial. Puede definirse como métodos computacionales que, utilizando la experiencia, permiten mejorar el rendimiento o hacer predicciones acertadas [17].

2.1.2.1 Tipos de problema

Dependiendo de las características particulares de los datos, así como del resultado que se pretende alcanzar, es posible agrupar los problemas en tipos. A continuación, se resumen brevemente los tipos de problemas que se pueden resolver utilizando métodos del área de Aprendizaje de Máquinas.

2.1.2.2 Clasificación

Los problemas de clasificación son aquellos en los que se necesita asignar una categoría a cada elemento. En esta clase de problemas, las categorías son definidas con anterioridad y suelen ser un número relativamente pequeño [17].

2.1.2.3 Regresión

En los problemas de regresión se quiere predecir un valor para cada elemento. En estos problemas se cuenta con sus valores reales, por eso es posible penalizar las predicciones incorrectas dependiendo de la magnitud de la diferencia entre los valores reales y los valores predichos [17].

2.1.2.4 Ranking

Los problemas de ranking son aquellos en los que se cuenta con un conjunto de elementos y un criterio por el cual los elementos deben ser ordenados [17].

2.1.2.5 Clustering

Consiste en los problemas en los que se busca separar los elementos en regiones o grupos homogéneos. Esta técnica es usada comúnmente para el análisis de conjuntos de datos muy grandes [17].

2.1.2.6 Reducción de dimensionalidad

En este tipo de problemas se transforman los elementos desde su representación inicial a una dimensión menor, conservando ciertas cualidades [17].

2.1.3 Tipos de aprendizaje

El tipo de aprendizaje depende de: el tipo de datos disponibles, el orden y el método por el cual el conjunto de entrenamiento es recibido, y los datos de prueba utilizados

para evaluar el algoritmo [17]. Usualmente los tipos de aprendizaje se pueden agrupar en tres categorías: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por reforzamiento.

El aprendizaje supervisado tiene como objetivo aprender a asociar las entradas a las respectivas salidas. Para esto utiliza un conjunto de entrada especial llamado el conjunto de entrenamiento. En este conjunto, cada elemento está relacionado a su respectiva etiqueta [19]. Dado que se cuenta con un conjunto con los resultados esperados, es posible medir qué tan acertado es el modelo.

En el aprendizaje no supervisado el objetivo es descubrir “patrones interesantes” en los datos. También es conocido como descubrimiento de conocimiento. En este caso no se cuenta con un conjunto de aprendizaje, por lo que no es posible determinar qué tan lejos se está de alcanzar un resultado [19].

El tercer tipo es el aprendizaje por reforzamiento. Consiste en aprender cómo actuar o comportarse cuando ocasionalmente se recibe una recompensa o un castigo [19]. Si el comportamiento se considera positivo se intentará reforzar el mismo al dar un estímulo positivo o recompensa. Por el otro lado si el comportamiento se considera negativo se generará un estímulo negativo o un castigo.

2.1.4 Modelos Gráficos Probabilísticos

Los modelos gráficos probabilísticos utilizan una representación basada en grafos para codificar, de manera compacta, distribuciones complejas en espacios de alta dimensionalidad [15].

La representación gráfica puede ser interpretada de dos formas: como un conjunto de las independencias que se mantienen en la distribución, o cómo la agrupación de las probabilidades de factores más pequeños. Ambas son inducidas por la estructura del grafo [15].

La figura 2.1 [15] presenta dos ejemplos de un modelo gráfico probabilístico. La

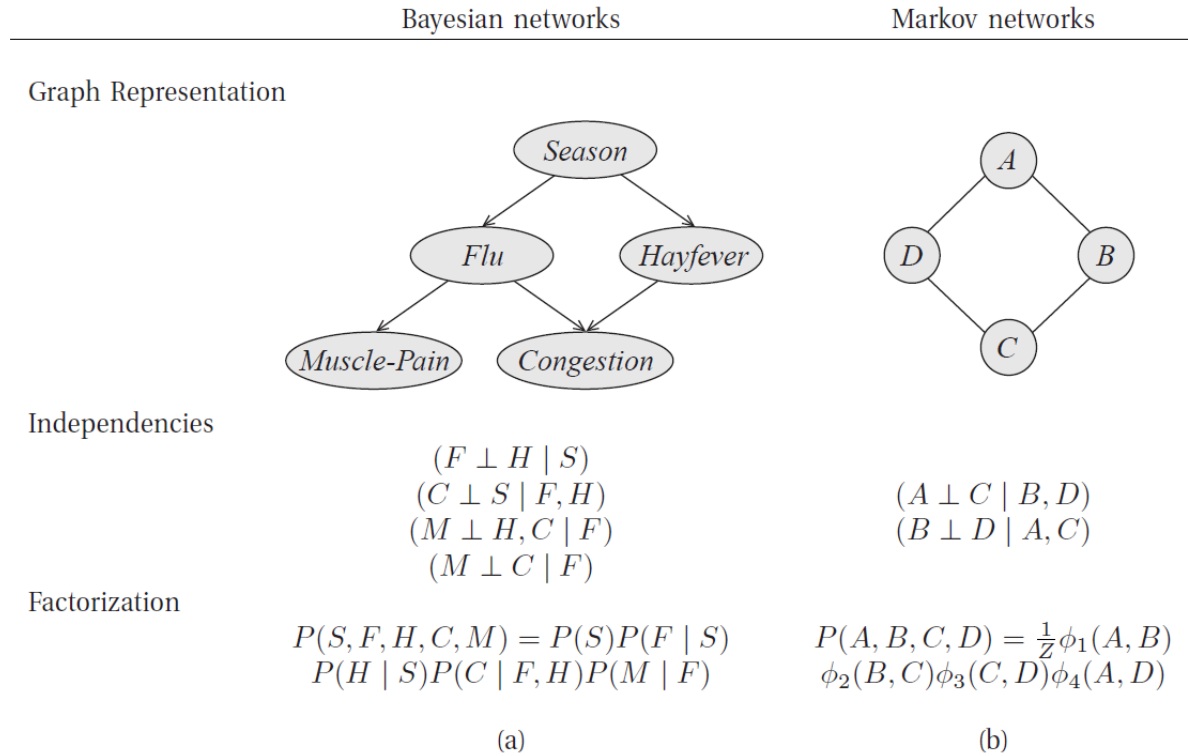


Figura 2.1: Ejemplo de un modelo gráfico probabilístico.

columna (a) presenta un ejemplo de una Red Bayesiana. Cada uno de los nodos expone un síntoma, excepto el primer nodo, que representa *Season* [estación]. La columna (b) tiene un ejemplo de una Red de Márkov. Cada uno de los nodos representa una variable y las aristas representan las influencias entre cada uno de los nodos. En la parte baja de la imagen se pueden observar las dos perspectivas mencionadas anteriormente.

En la sección *Independencies* (independencias), se resume la información que está en el grafo desde la perspectiva de las independencias. Para tres variables aleatorias X , Y y Z , la notación $P(X \perp Y | Z)$ indica que X es independiente de Y dado Z . Siguiendo el ejemplo de la figura anterior, para una distribución P entonces: $P(C | F, H, S) = P(C | F, H)$, es decir que $(C \perp S | F, H)$. Esto no implica que S es independiente de C , sino que toda la información que se podría obtener de cómo S influye en C ya se conoce dado F y H [15].

En la parte final de la figura, en la sección *Factorization* (factorización), se puede

ver la otra perspectiva del modelo: la factorización. Esto permite que en lugar de codificar todos los posibles valores para todas las variables del dominio, se pueda separar la distribución en factores, cada uno con un espacio de probabilidades más pequeño. Es posible definir la probabilidad conjunta total como el producto de estos factores. Siguiendo el ejemplo de los síntomas, la probabilidad del evento (primavera, no gripe, congestión, dolor muscular) se puede obtener al multiplicar: $P(S = primavera), P(F = falso | S = primavera), P(H = verdadero | S = primavera), P(C = verdadero | H = verdadero, F = falso), P(M = verdadero | F = falso)$. Esta representación es más compacta ya que requiere de menos parámetros en comparación con la distribución conjunta original [15].

2.1.5 Tipos de Modelos Gráficos Probabilísticos

Los modelos gráficos probabilísticos se pueden dividir en dos grandes grupos: las Redes Bayesianas, que se representan utilizando un grafo acíclico dirigido, y las Redes de Márkov, que utilizan un grafo acíclico no dirigido.

2.1.5.1 Redes Bayesianas

Las Redes Bayesianas (BN) corresponden al ejemplo en la columna (a) de la figura 2.1. Este modelo utiliza un grafo acíclico dirigido para codificar la información. En estos grafos, los nodos son variables aleatorias y las aristas representan la influencia directa entre un nodo y otro [15].

Dada una BN con un estructura de grafo G , sus nodos representan las variables aleatorias X_1, \dots, X_n . Si se denotan los padres de X_i como: Pa_X^G , y a todas las variables en el grafo que no son descendientes de X_i como: $NoDescendientesX_i$,. Entonces G contiene el conjunto de independencias locales $L_i(G) = X_i : (X_i \perp NoDescendientesX_i | Pa_X^G)$. Es decir que toda variable aleatoria en el modelo, es condicionalmente independiente de cualquiera de sus no descendientes dados sus padres

[15].

Los elementos básicos que deben ser definidos para poder modelar una BN son: las variables, la estructura y las probabilidades iniciales.

2.1.5.2 Redes de Márkov

Las Redes de Márkov (RM) utilizan grafos no dirigidos para codificar la información. Al igual que en las Redes Bayesianas, los nodos del grafo representan las variables aleatorias, y las aristas representan una noción de interacción directa entre dos variables [15].

Al codificar la información utilizando un grafo no dirigido, las RM son útiles para modelar problemas en los cuales, por su naturaleza, no es posible representar la dirección de las interacciones entre las variables [15].

En la columna (b) de la figura 1, se presentan las independencias y la factorización para el caso de las MN. En este caso, al ser un grafo no dirigido, se puede resumir las independencias descritas en el grafo a: $(A \perp C \mid B, D)$ y $(B \perp C \mid A, D)$ [15].

Para expresar la misma información como el producto de factores, primero se divide el grafo en subgrafos que tengan todos sus nodos interconectados o cliques. Luego se tiene un factor, representado con la letra ϕ , que mide la relación entre un grupo de variables. Para este ejemplo se pueden tomar los cliques: (A, B) , (B, C) , (C, D) y (A, D) . Este valor tiene que ser finalmente normalizado, por lo que obtenemos: $P(a, b, c, d) = \frac{1}{Z} \phi_1(a, b) \phi_2(b, c) \phi_3(c, d) \phi_4(a, d)$, donde Z es conocido como la función de partición tal que $Z = \sum_{a,b,c,d} \phi_1(a, b) \phi_2(b, c) \phi_3(c, d) \phi_4(a, d)$ [15].

2.1.5.3 Modelos Ocultos de Márkov

Los Modelos Ocultos de Márkov (MOM) son modelos temporales probabilísticos en los que el estado es descrito utilizando una única variable aleatoria. Los valores posibles de las variables son los estados posibles del sistema. Cuando un modelo requiere

de dos o más variables estado, éstos se fusionan en una única mega variable cuyos valores son todas las tuplas de los valores de las variables estado individuales. Esto restringe la estructura de los MOM y permite una implementación matricial simple y elegante de todos los algoritmos básicos [27]. Esto se consigue a expensas de crear una matriz cuyo tamaño es dependiente de la cantidad de las variables estado del sistema, y en consecuencia, aumentando el costo en recursos y el costo computacional de los algoritmos que se realicen sobre ésta.

2.1.5.4 Redes Bayesianas Dinámicas

Las RBD son una extensión a las RB. Permiten representar no solamente el estado en un momento dado del evento que se está modelando, sino que además permiten modelar su evolución a través del tiempo [15].

Inicialmente se hace una simplificación que consisten en utilizar tiempo discreto. Se asume que las mediciones del sistema van a ser tomadas en intervalos regulares y una granularidad de tiempo Δ . Esto permite representar el estado de un sistema, con variables de la forma X_i , con $i \geq 0$, en el tiempo $t\Delta$ como: X_0, X_1, \dots, X_t . Es decir, el estado de todas las variables desde el tiempo 0 hasta el tiempo t , ó $P(X^{0:T}) = P(X^0) \prod_{t=0}^{T-1} P(X^{t+1} | x^{0:t})$, donde $P(X^{0:T}) = P(X^0, \dots, X^T)$. Lo que implica que, para representar el estado t del sistema, es necesario representar todos los estados anteriores [15].

Evidentemente resulta muy costoso representar tanta información en un modelo, más aún si se está modelando un sistema con una gran cantidad de variables e información. Para resolver este problema es necesario presentar dos conceptos que se van a asumir para este modelo: la Propiedad de Márkov y la estacionalidad. Con la Propiedad de Márkov, expresada como $(X^{t+1} \perp X^{0:t-1} | X^t)$, se asume que el futuro es condicionalmente independiente del pasado, dado el presente. En otras palabras, que el sistema carece de memoria [15].

2.1. Antecedentes teóricos

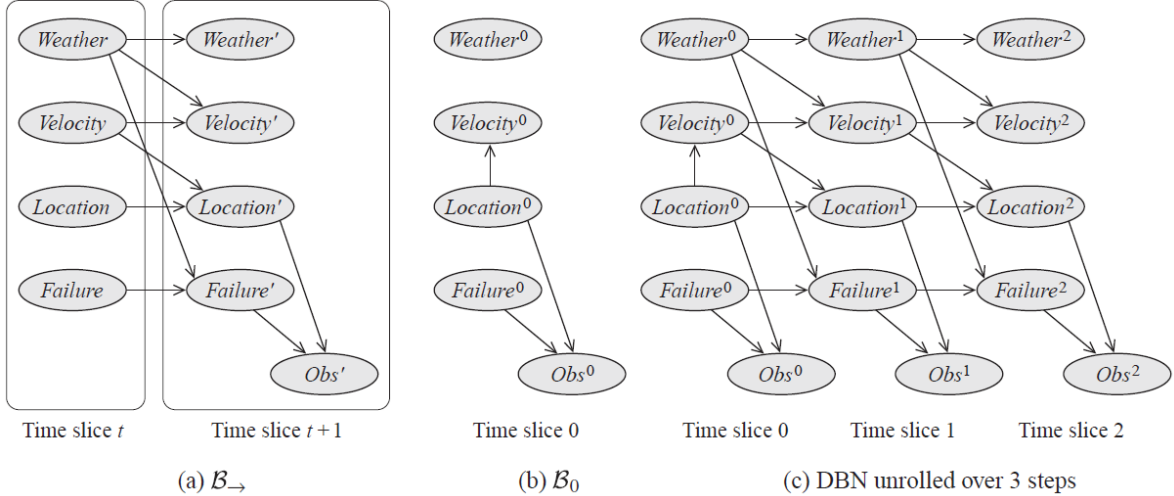


Figura 2.2: RBD para monitoreo de vehículos: a) la 2TBN; b) la red en el tiempo 0; c) la RBD resultante sobre 3 fragmentos de tiempo.

La Propiedad de Márkov nos permite representar la distribución que se está modelando de manera más compacta. Ahora el estado del sistema en un tiempo t , se puede representar como: $P(X^{0:T}) = P(X^0) \prod_{t=0}^{T-1} P(X^{t+1} | X^t)$. A un sistema que cumple con la Propiedad de Márkov para todo $t \geq 0$, se le conoce como un sistema de Márkov [15].

La otra suposición en la que estamos interesados es la homogeneidad o estacionalidad. Suponiendo un modelo donde $P(X' | X)$, donde X es el estado de la red en el tiempo actual y X' el tiempo siguiente. Un sistema de Márkov es homogéneo si: $P(X^{t+1} | X^t) = P(X' | X)$ es igual para todo t . Esto quiere decir que las dinámicas del modelo no cambian a través del tiempo [15].

Con estas dos suposiciones, solamente necesitamos representar el estado inicial de la distribución y el modelo de transición $P(X' | X)$. El modelo de transición que representa las dinámicas del modelo $P(X' | X)$, es conocido como Plantilla de Modelo de Transición. Esta transición es una distribución probabilística condicional, que puede representarse como una Red Bayesiana Condicional. Es decir una red en el tiempo t , que depende del estado anterior de la red [15].

Ahora podemos construir el concepto de Redes Bayesianas de dos fragmentos de

tiempo (2TBN por sus siglas en inglés). Los nodos de una 2TBN sobre las variables X_1, \dots, X_n , incluirá todo las variables X' y un subconjunto de las variables de X . En un 2TBN solamente los nodos X'_1, \dots, X'_n tendrán padres y su propia distribución conjunta, y define la siguiente distribución conjunta: $P(X' | X) = \prod_{i=1}^n P(X'_i | Pa_{X'_i})$ [15].

Como se puede ver en la figura 2.2[15] a, solamente las variables en el tiempo $t + 1$, es decir X' , tiene padres. Y no todas las variables de X son copiadas para X' . La figura 2.2 muestra además en (b), el estado inicial de la red. Aplicando múltiples veces la 2TBN, se crea una red desenrollada. Como se muestra en la parte c de la misma figura, en la cual se muestra la RBD resultante de desenrollar 3 fragmentos de tiempo [15].

Debe notarse que un MOM puede ser representado como una RBD con un único nodo oculto y un único nodo observado. Además, todas las RBD con variables discretas pueden ser representadas como un MOM y viceversa [27]. Lo que hace atractivas las RBD es la manera compacta en la que representan la información. Por ejemplo, dada una RBD con 20 nodos booleanos cada uno con 3 nodos padres, entonces el modelo de transición tendría $20 * 2^3 = 160$ probabilidades, mientras que en un MOM se tendrían 2^{20} estados y 2^{40} , o casi mil billones, de probabilidades en una matriz de transición. Esto es malo por tres razones: 1) el MOM requiere de muchísimo espacio, 2) una matriz tan grande haría la inferencia más cara y 3) aprender esa cantidad tan grande de parámetros no es factible para problemas grandes [27].

La capacidad de las RBD de representar el tiempo, aunado a su compacta representación, fue lo que finalmente hizo que fueran seleccionadas como modelo para la implementación de la herramienta que se desarrollará en esta investigación.

Descripción general de la investigación

3.1 Planteamiento del problema

La presente investigación consiste en determinar si las Redes Bayesianas Dinámicas pueden ser utilizadas para predecir el avance de la Sigatoka Negra y la productividad del cultivo.

Para esto se desarrollará una herramienta de Software basada en RBD, y se utilizarán los datos históricos provistos por CORBANA[6]. Además, se creará una segunda implementación utilizando RBs para hacer un análisis comparativo de los resultados obtenidos con cada implementación.

Existen otros modelos e implementaciones que han sido probadas para este mismo fin, inclusive sobre el mismo conjunto de datos [5]. Pero la presente investigación se restringe a las RBD.

3.2 Trabajos Relacionados

¹ Las Redes Bayesianas (RB) han sido utilizadas para la predicción de enfermedades anteriormente. En [29] se utilizan Redes Bayesianas para representar la relación entre los síntomas y las enfermedades. Para tal fin se utilizan la BN, en conjunto con un algoritmo de aprendizaje incremental. Con ello se obtiene un método de diagnóstico eficiente.

Además en [13] se busca cómo crear un modelo que permita la predicción en tiempo real de la presencia de una enfermedad. En el estudio, se plantea la utilización -no solamente las fuentes de datos tradicionales- sino de los datos provenientes de fuentes como: reportes de salas de emergencias, venta de fármacos y reportes de laboratorio. Los datos provenientes de estas fuentes tienden a ser incompletos y además, suelen estar disponibles con cierto retraso en comparación a los datos tradicionales. Se demuestra

¹Este trabajo está relacionado con la tesis doctoral de Luis-Alexander Calvo-Valverde en el DOCI-NADE [5].

3.3. Hipótesis

que su modelo basado en Redes Bayesianas Dinámicas, puede utilizarse con el fin de lograr predicción en tiempo real y a su vez permiten la utilización de fuentes de datos heterogéneas.

En este mismo ámbito, en [1] las RBDs son utilizadas en conjunto con algoritmos de Evolución Diferencial para el diagnóstico de cáncer de hígado. Estos algoritmos se utilizan para maximizar las características en los datos. Con las RBDs se lograron inferir relaciones temporales entre estas características de manera exitosa. Muestra de ello fue que se identificaron relaciones entre genes que no se conocían previamente. En [3] también se usaron con la finalidad de encontrar relaciones entre genes de la levadura y se compararon con otras alternativas. En este estudio se observó que las RBDs encontraron una mayor cantidad de relaciones entre los genes que las otras alternativas evaluadas.

Este tipo de redes también han sido utilizadas en otras áreas. En [9] son utilizadas para descomponer objetos en 3D en sus partes que los componen y a la vez crear representaciones *esqueléticas* de las mismas. Se planteó un modelo probabilístico para estimar partes faltantes, y se mostró la efectividad de este enfoque.

3.3 Hipótesis

El uso de RDBs supera en capacidad de predicción a las RBs en cuanto al avance de la Sigatoka Negra y la productividad del cultivo.

3.4 Objetivos

3.4.1 Objetivo general

Evaluar el uso de RBDs para la predicción del avance de la enfermedad Sigatoka negra y la productividad del cultivo.

3.4.2 Objetivos específicos

1. Modelar RDB's que permitan predecir el avance de la Sigatoka negra y la productividad del banano.
2. Modelar RBs no dinámicas que permitan predecir el avance de la Sigatoka negra y la productividad del banano.
3. Comparar los resultados obtenidos entre ambos modelos.

3.5 Resumen del Experimento

Para realizar este experimento primero se modeló e implementó una RB (ver sección 3.6.4) y una RBD (ver sección 3.6.3).

El experimento contó con varios factores cuyos valores fueron ajustados en las distintas ejecuciones. Los factores que fueron utilizados son listados en la sección 3.5.1 y explicados en detalle en la sección 3.6.2.

Con este experimento se trató de estudiar el comportamiento de los RBs y RBDs utilizando dos fenómenos: 1) el avance de la enfermedad en los cultivos y 2) la producción del cultivo. Se utilizaron datos climáticos y las mediciones de cada factor para predecir el comportamiento de cada fenómeno. Para ello se utilizaron los conjuntos de datos descritos en la sección 3.6.2.2.

Los conjuntos de datos fueron divididos en dos: el conjunto de entrenamiento y el conjunto de prueba. El conjunto de entrenamiento se utilizó para entrenar cada tipo de red. Una vez que se entrenaron las redes se utilizó el conjunto de pruebas para medir la eficacia de predicción de cada red. La predicción fue medida utilizando las métricas F1 micro y F1 macro (ver sección 3.5.2).

Los resultados obtenidos en los experimentos fueron recopilados y analizados. La sección 4.2 muestra el análisis estadístico que se realizó sobre los resultados.

3.5.1 Factores

Los factores de diseño del experimento son:

- Tipo de red:
 - Red bayesiana(RB)
 - Red bayesiana dinámica (RBD)
- Conjunto de datos:
 - Producción (PN)
 - Estado de evolución (EE)
- Tamaño de las variables:
 - 3: variables que pueden tomar valores entre 1 y 3.
 - 5: variables que pueden tomar valores entre 1 y 5.
- Tamaño del slice: la cantidad de semanas utilizadas para entrenar a la red. Se utilizaron entre 1 y 5 semanas.

Un tamaño de slice de 1 indica que se utilizarán los datos de 1 semana para predecir el valor de la variable de salida para la siguiente semana. Del mismo modo, un tamaño de variable de 5 indica que se utilizarán los datos de 5 semanas para predecir el valor de la variable de salida para la siguiente semana.

3.5.2 Variables de respuesta

Las redes aprenden los patrones de los datos climáticos y de la variable de salida, ya sea la producción o el avance de la enfermedad. Luego, a partir de la nueva evidencia que se proporcione, es posible utilizarlas para predecir el valor de la variable de salida.

El trabajo que realizan las redes es el de asignar una etiqueta a la variable de salida, que corresponde al valor predicho, a partir de las observaciones proporcionadas de las variables climáticas. Es un problema de clasificación con múltiples clases, es decir, que se quiere asignar a cada muestra una y solo una etiqueta [23].

Cuando se trata de un problema de clasificación de una etiqueta, la predicción efectuada puede evaluarse como se describe en la tabla 3.1 [2] que muestra una matriz de confusión. Las filas 2 y 3 de la primera columna indican si la etiqueta fue predicha o no por el modelo mientras que los valores de las columnas 2 y 3 indican los valores esperados. Las intersecciones entre lo predicho por el modelo y lo esperado indican el resultado de la clasificación. Por ejemplo, si el modelo predijo A y el valor esperado era A se tiene un Verdadero Positivo (fila 2: etiqueta A predicha, columna 2: etiqueta A esperada)

Para evaluar la predicción se compara si se predijo o no la clase para la muestra y, al comparar con el valor esperado, si la clase asignada fue correcta. En esta figura la clasificación sería evaluada en:

- VP (verdadero positivo): cuando la muestra pertenece a una clase y fue clasificada para esa clase.
- FN (falso negativo): cuando la muestra pertenece a una clase y no fue clasificada para esa clase.
- FP (falso positivo): cuando la muestra no pertenece a una clase y fue clasificada para esa clase.
- VN (verdadero positivo): cuando la muestra no pertenece a una clase y no fue clasificada para esa clase.

Van Rijsbergen (1975) (citado en [2]) introdujo la métrica Efectividad, también llamada F1. Esta métrica se calcula a su vez utilizando dos métricas: la Precisión y

3.5. Resumen del Experimento

	Etiqueta A esperada	Etiqueta A no esperada
Etiqueta A predicha	Verdadero positivo (VP)	Falso positivo (FP)
Etiqueta A no predicha	Falso negativo (FN)	Verdadero negativo (VN)

Table 3.1: Evaluación de clasificación para la etiqueta A.

Exhaustividad. Utilizando la evaluación anterior es posible calcular estas métricas para determinar la eficacia del clasificador siguiendo estas fórmulas:

$$precision = \frac{VP}{VP + FP} \quad (3.1)$$

$$exhaustividad = \frac{VP}{VP + FN} \quad (3.2)$$

La efectividad se expresa en función de la precisión y la exhaustividad de la siguiente manera:

$$F1 = \frac{precision * exhaustividad}{precision + exhaustividad} \quad (3.3)$$

Como puede verse, tanto la evaluación del clasificador como las métricas, se expresan desde el punto de vista de una clase. Se evalúa de manera binaria, es decir si el clasificador tuvo éxito o no al asignar la etiqueta para una clase. Para poder evaluar integralmente al clasificador, se realiza esta evaluación para cada una de las clases. Esto permite calcular cómo se comporta el modelo a nivel general y no solamente desde el punto de vista de una etiqueta en particular. Los resultados de esta evaluación hecha desde el punto de vista de cada clase pueden ser promediados de dos formas: utilizando el promedio micro o macro.

El promedio macro le da el mismo peso a cada una de las clases, mientras que el promedio micro le da el mismo peso a cada una de las clasificaciones individuales [2]. Como se ve en la ecuación 3.3, las métricas F1 se calculan como la media armónica de la precisión y la exhaustividad [2].

F1 micro da la relación entre las apariciones de las etiquetas y aquellas predichas

por el clasificador basado en la suma de decisiones por entrada. Por otro lado el F1 macro da la relación entre las apariciones de las etiquetas y aquellas predichas por el clasificador basado en un promedio por clase [28].

Las ecuaciones 3.4 y 3.5 muestran cómo calcular la versión micro de la precisión y la exhaustividad, mientras que las ecuaciones 3.6 y 3.7 muestra la versión macro [2].

$$precision_micro = \sum_1^c VP_i / (VP_i + FP_i) \quad (3.4)$$

$$exhaustividad_micro = \sum_1^c VP_i / (VP_i + FN_i) \quad (3.5)$$

$$precision_macro = (\sum_1^c VP_i / (VP_i + FP_i)) / C \quad (3.6)$$

$$exhaustividad_macro = (\sum_1^c VP_i / (VP_i + FN_i)) / C \quad (3.7)$$

3.5.3 Prueba de hipótesis

Para poder sacar conclusiones de los resultados obtenidos en los experimentos, es necesario primero realizar un análisis estadístico para determinar si los resultados se deben al factor en sí o a se deben al azar.

El primer paso de este análisis fue determinar si la población de las variables de respuesta era normal. Para esto se utilizó la prueba de normalidad de Shapiro-Wilk [24] así como gráficos cuantil-cuantil para analizar la distribución. Se planeaba utilizar la prueba ANOVA [24], pero los resultados de las pruebas de normalidad indicaron que las poblaciones de las variables de respuestas no eran normales (ver sección 4.1.1). Es por esta razón que se utilizaron métodos no paramétricos para analizar la significancia estadística que cada uno de los factores tuvo sobre los resultados. Se utilizó

específicamente la prueba Wilcoxon-Mann-Whitney [24] para factores con 2 valores, y Kruskal-Wallis [24] para factores con 2 o más valores.

Estas pruebas indican si un factor tiene significancia estadística en el F1 micro y el F1 macro. Para determinar si un factor es estadísticamente significativo sobre las variables de respuesta se plantea la siguiente hipótesis:

H_0 : La distribución de la variable de respuesta para cada valor del factor es la misma.

H_1 : La distribución de la variable de respuesta para cada valor del factor no es la misma.

En la sección 4.1.1 se detallan los resultados obtenidos al realizar este análisis.

3.6 Detalles del Experimento

3.6.1 Ambiente de desarrollo

Detalles del software:

- Para la RB
 - Ambiente integrado de desarrollo para Python: Spyder 3.1.3 [25]
 - Python versión 2.7.13 [11]
 - LibPGM 1.1 [4]
- Para la RBD
 - Matlab R2016a 9.0.0.341360 [12]
 - LibBNT 1.0.7 [18]
- Para análisis estadístico
 - R x64 3.3.2 [10]

– R Studio 1.0.136 [26]

Detalles de la computadora utilizada para correr los experimentos e implementar las redes:

- Sistema operativo Windows 10 64bits
- Procesador: Intel I7 3770k @3.5GHz
- Memoria RAM: 16GB DDR3 @686MHz
- Tarjeta madre: Asus P8Z68-V Gen3

3.6.2 Descripción de los factores

En esta sección se describen cada uno de los factores que se utilizaron en los experimentos y sus valores.

3.6.2.1 Tipo de red

Para los experimentos se utilizó una implementación de RB y otro de RBD. La RB se implementó en lenguaje de programación Python utilizando la biblioteca LibPGM [4] (ver 3.6.1). La RBD fue implementada en Matlab utilizando el biblioteca LibBNT [18] (ver 3.6.1). En las secciones 3.6.4 y 3.6.3 se detalla el diseño e implementación de las RB y las RBD respectivamente.

3.6.2.2 Conjunto de datos

Para los experimentos se utilizaron los datos de CORBANA[6]. Se utilizaron 2 archivos de datos: uno para el avance de la enfermedad y otro para la producción. Ambos consisten en las mediciones de cada fenómeno y de las variables climáticas correspondientes al periodo en el que se realizó la medición.

3.6.2.2.1 Avance de la enfermedad

Se utilizaron los datos climáticos y del avance de la enfermedad en la finca La Rita. Las variables climáticas utilizadas fueron: humedad, velocidad del viento, precipitación y temperatura. Para medir el avance de la enfermedad, se utiliza la técnica del Preaviso Biológico. Esta técnica observa el avance y velocidad de la enfermedad, y consiste en la detección temprana de los síntomas en las hojas más jóvenes de la planta [16].

Se tienen 675 registros, correspondientes a una muestra semanal hecha entre el año 2001 y el 2014. Estos datos fueron discretizados en dos tamaños 3 y 5. Las variables con tamaño 3 pueden tomar valores entre 1 y 3, mientras que las variables de tamaño 5 pueden tomar valores entre 1 y 5. Estos valores permiten utilizar el modelo como un sistema de alerta temprana, y detectar cambios en el comportamiento de la variable de salida.

Los valores fueron discretizados agrupando los valores de cada variable en la cantidad de rangos homogéneos correspondiente al tamaño de la variable deseada. Por ejemplo, para un tamaño de variable 3, se agrupan los valores en 3 rangos homogéneos lo que permite traducir los valores: 1, 2 y 3.

3.6.2.2.2 Producción

Para medir la producción se utiliza la medida de peso neto. Se utilizaron los datos climáticos y de Producción de la finca *28 Millas*. Las variables climáticas utilizadas fueron: humedad, velocidad del viento precipitación y temperatura. En el caso de la producción se cuenta con 159 muestras, que corresponden a 159 semanas. Estos datos se discretizaron para crear dos archivos: uno con valores entre 1 y 3, y el otro con valores entre 1 y 5.

Los valores fueron discretizados agrupando los valores de cada variable en la cantidad de rangos homogéneos correspondiente al tamaño de la variable deseada. Por ejemplo, para un tamaño de variable 3, se agrupan los valores en 3 rangos homogéneos lo que

permite traducir los valores: 1, 2 y 3.

3.6.2.3 Tamaño de los nodos

Durante la evaluación de las bibliotecas se observó que la biblioteca de LibBNT [18] requiere de mucha memoria RAM para poder representar cada uno de posibles valores de las variables. La memoria requerida por esta biblioteca se incrementa sustancialmente al aumentar el tamaño de los nodos. Es decir, conforme aumenta la cantidad de valores que el nodo puede tomar.

En las pruebas iniciales hechas con la biblioteca LibBNT inclusive se recibió una advertencia de Matlab indicando que se requeriría más de 100 giga bytes de memoria RAM para poder almacenar los valores utilizando nodos de tamaños 25. Además, el tiempo requerido para entrenar la red se incrementa significativamente al utilizar variables más grandes.

Estos datos se discretizaron para crear dos archivos: uno con valores entre 1 y 3, y el otro con valores entre 1 y 5.

3.6.2.4 Tamaño del slice

Tal cómo se observó con el tamaño de los nodos (3.6.2.3), el tamaño del slice tiene un impacto significativo en el uso de recursos. Durante una prueba utilizando un conjunto de datos externos de la UCI [20], el entrenamiento de la red implementada con LibBNT corrió durante una semana antes de que Matlab fallará y, basado en el avance que se obtuvo durante este tiempo, se estimó que requeriría de al menos 24 días para terminar de entrenarse.

Se utilizaron tamaños de slice desde 1 a 5.

3.6.3 Diseño e implementación de la Red Bayesiana Dinámica

Se realizó un análisis de los datos antes de diseñar la estructura de la red. Lo que se requería era representar con la red las relaciones que existen entre las variables climáticas y el comportamiento del avance de la enfermedad y de la producción. Tomando en cuenta que en ambos casos se tenían numerosas variables climáticas relacionadas a una variable de salida, el avance de la enfermedad o la producción, se utilizó la misma estructura para todos los experimentos sin importar cuál conjunto de datos se utilizó.

En las RBDs las variables para las cuáles se conocen sus valores se modelan como nodos observables. Cada conjunto de datos constaba de 4 variables de entrada correspondiente a los datos climáticos. Dado que los valores de cada una de las variables eran conocidos, es decir que se cuenta con las mediciones de cada variable en cada muestra, estas variables fueron representadas en la red como nodos observables.

El valor de la variable de salida, la producción o el avance de la enfermedad dependiendo del conjunto de datos que se estuviese usando en el experimento, también era conocido en cada una de las muestras del conjunto de datos. Por lo que también se modela como un nodo observable.

La forma en la que la RBD codifica las relaciones que aprende de los datos es a través de nodos especiales llamados nodos ocultos. Los nodos ocultos modelan las relaciones que existen entre los nodos observados. La red aprende al ajustar los valores de los nodos ocultos según las observaciones que se provean como entrada. Es por esto que fue necesario agregar un nodo oculto para representar la relación que existe tras el comportamiento de la variable de salida y las variables observadas. Para los conjuntos de datos usados en estos experimentos, esto se traduce en la relación que existe entre las variables climáticas y la variable de salida.

Una vez que la red fue entrenada, es posible utilizarla para contestar a la pregunta: "Dadas estas observaciones de las variables climáticas y de la variable de salida en

3.6. Detalles del Experimento

tiempos anteriores, ¿Cuál es el valor más probable para la variable de salida?”. Es decir que es posible inferir el valor de la variable de salida a partir de sus valores anteriores y de las observaciones de las condiciones climáticas.

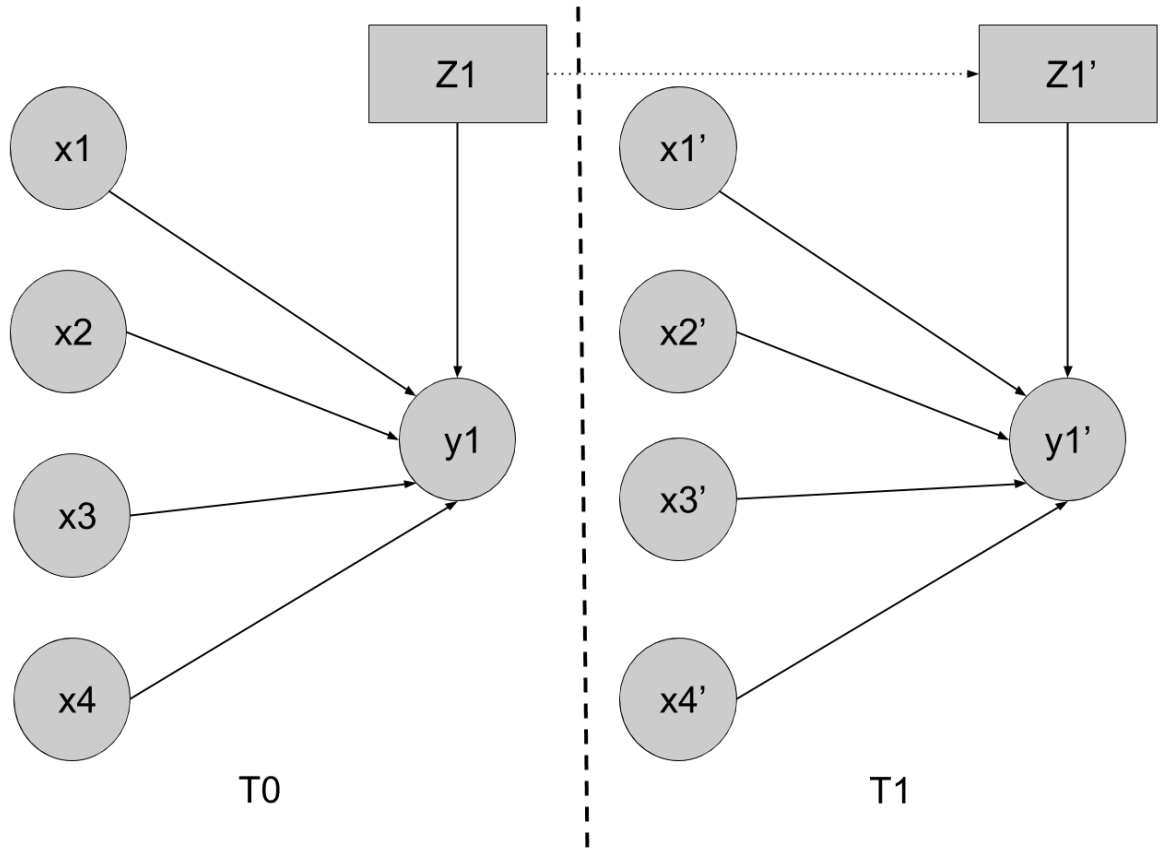


Figura 3.1: Diseño de la Red Bayesiana Dinámica.

Los arcos entre los nodos de la red representan que un nodo puede influenciar a otro nodo de la red. La influencia fluye a través de la red en los casos en los que los caminos entre los nodos a través de los arcos que los unen sean caminos activos [14].

Cuando existe una estructura en V , es decir dos nodos con arcos a un mismo nodo descendiente, la influencia fluirá por él solamente en el caso de que se cuente con evidencia del valor del nodo descendiente [14].

Tomando en cuenta esta característica de las redes se diseñó la red de forma tal que tanto los nodos de las variables climáticas como el nodo oculto tienen arcos hacia el nodo

de salida. Ningún arco se origina desde el nodo de salida. Estos arcos corresponden a las relaciones dentro del mismo slice.

Los arcos entre slices representan cuando un nodo tiene incidencia en el valor del mismo nodo en el siguiente periodo. En la red aquí diseñada existe un único arco entre el nodo oculto en un slice y sí mismo en el siguiente slice. Esto sucede porque se quiere que la información que el nodo oculto va aprendiendo, se propague en el tiempo.

La figura 3.1 muestra la estructura final de la red. La línea vertical punteada separa un slice del siguiente. En la parte izquierda se muestra la estructura de la red en el T_0 , es decir, el estado inicial de la red. En la parte derecha está el estado de la red en el tiempo T_1 . Los arcos con líneas punteadas representan las relaciones entre slices, las negras las relaciones intra slices. Los nodos observados se representan con círculos, mientras que los ocultos con rectángulos. Los nodos x_i corresponden a los nodos observados de las variables climáticas, el nodo y_1 es el nodo de salida, el nodo del cual la red va a predecir el valor, y el nodo z_1 es el nodo oculto. El conjunto de arcos entre los slices es lo que se conoce como el 2TBN, y es lo que permite que la red se expanda a través del tiempo.

El nodo oculto aprende las relaciones entre los valores de los nodos observados al ajustar su valor según cada tiempo T . Ya que se quiere que esta información sea propagada a lo largo del tiempo sobre la red, como puede observarse en la figura 3.1, la única relación que hay entre slices es la del nodo oculto. No hay otro arco entre slices dado que el resto de los nodos son observados y sus valores son independientes entre sí.

3.6.4 Diseño e implementación de la Red Bayesiana

El diseño de la RB utilizada se basó en el diseño de la RBD descrito en la sección 3.6.4. De igual manera se utilizó la misma estructura de la RB en los experimentos sin importar el conjunto de datos. La estructura de la RB es esencialmente la misma que la de la RBD en el tiempo T_0 .

A diferencia de las RBD, las RB no representan datos temporales de forma directa.

Para hacerlo, es necesario expandir o 'desenrollar' la red acorde con la cantidad de slices que se esté utilizando.

La figura 3.2 muestra el diseño de la RB equivalente a una RBD para un tamaño de slice 2. Esto es una red desenrollada en dos tiempos, específicamente en dos semanas. Los nodos x_{ij} corresponden al i -ésimo nodo observado en el tiempo j , en este caso los nodos correspondientes a las variables climáticas. El nodo y_{1j} corresponde al nodo de salida en el tiempo j . Los arcos representan las relaciones entre los nodos.

Cada nodo tiene asociada una tabla de distribución condicional que representa la probabilidad de su valor dado el valor de cada uno de sus padres. La información en este tipo de red fluye ya que cada nodo de salida es padre del nodo de salida en el siguiente slice, influyendo así su valor. En este ejemplo (figura 3.2) el nodo y_{1_0} es padre de y_{1_1} , permitiendo que la información del primer periodo pase al siguiente.

3.6.4.1 Entrenamiento y evaluación

Ambas redes fueron entrenadas y evaluadas siguiendo el mismo procedimiento. El 70% de los datos se utilizaron como conjunto de entrenamiento. Luego el 30% de los datos restantes se usaron como conjunto de pruebas y fueron utilizados para medir la eficacia de la red para predecir el valor correspondiente.

Con cada una de las muestras del conjunto de entrenamiento se comparó el valor predicho por la red con el valor esperado y se evaluó como describe la tabla 3.1. La información de esta evaluación se guardó y sobre ésta se calcularon las métricas descritas en 3.5.2.

Este procedimiento se realizó combinando distintos valores de los factores. Finalmente se resumió el resultado de todos estos experimentos en una tabla sobre la cuál se realizó el análisis descrito en el capítulo 4.

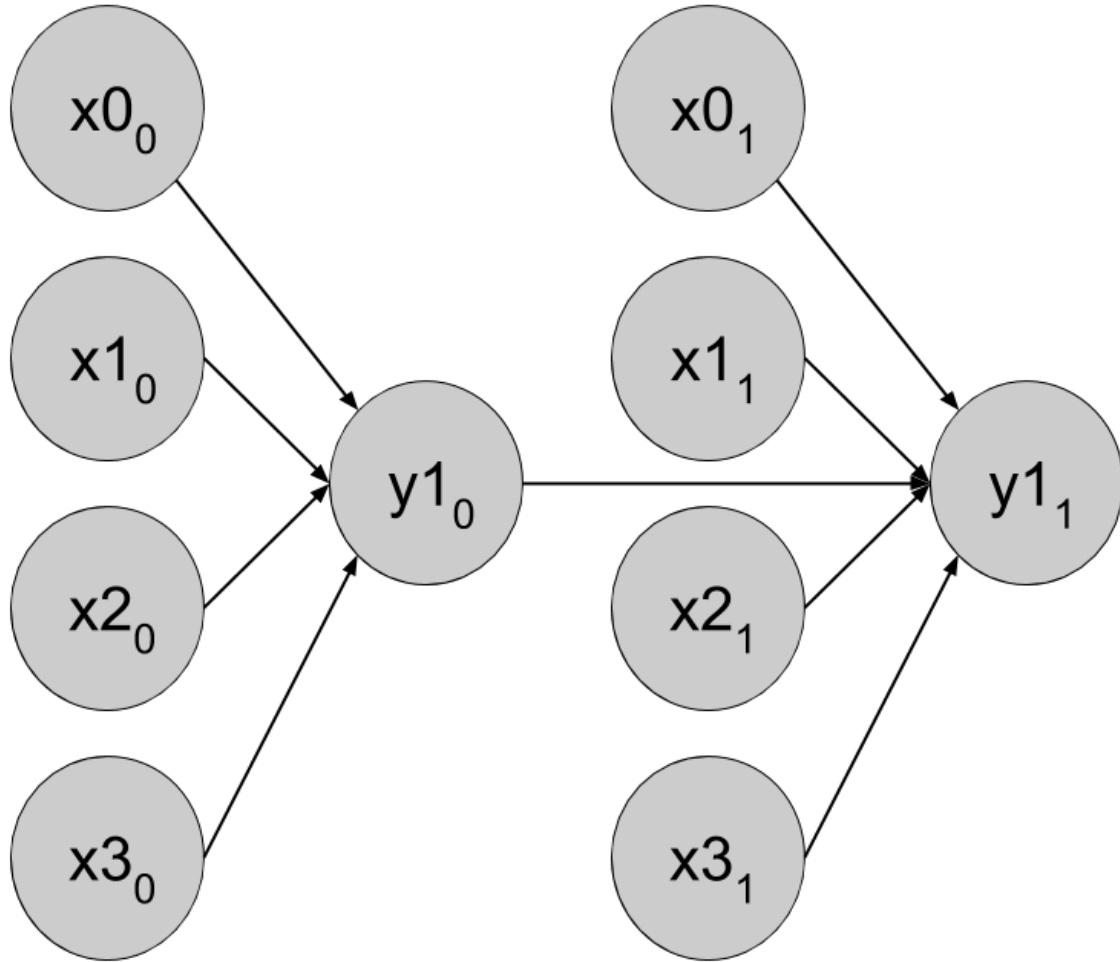


Figura 3.2: Ejemplo del Diseño de la Red Bayesiana desenrollada para un periodo de dos semanas.

3.6.4.2 Alternativas exploradas

Antes de seleccionar LibPGM [4] y LibBNT [18] para implementar las redes, se evaluaron varias opciones. Esta sección describe brevemente las opciones evaluadas y la razón por la cuál no fueron utilizadas.

La biblioteca Mocapy [22] escrita en C++ para el sistema operativo Linux, fue evaluada como alternativa para implementar las RBD. Esta opción presentó numerosos problemas de compatibilidad e interoperabilidad con versiones más modernas de las herramientas de las cuales depende.

Una vez que se logró configurar, se implementó una red de prueba. Sin embargo, al realizar pruebas se identificó que los resultados no eran correctos. Se intentó resolver estos problemas pero se concluyó que la implementación de la biblioteca para la parte de inferencia en RBD tenía errores y finalmente se descartó como una posible opción.

También se evaluó la biblioteca LibPMG [4], escrita en Python, como alternativa para implementar las RBD. Ésta fue descartada para implementar las RBD ya que para el momento de su evaluación no contaba con soporte para hacer inferencia sobre este tipo de redes. Sin embargo fue seleccionada para implementar las RB.

Finalmente se evaluó el programa Netica [7]. Se utilizó su ambiente gráfico para realizar pruebas. En su versión gratis se limitan las redes a menos de 10 nodos, incluyendo los nodos al desenrollar la red, por lo que los experimentos fueron bastante limitados. De todas las opciones evaluadas Netica representaba la mayor inversión monetaria, su costo oscila entre \$265 y \$685 dependiendo del tipo de licencia. A pesar de que parecía una opción viable, se descartó su uso debido a la ausencia de acceso a sus APIs sin contar con una licencia, ya que esto imposibilitó evaluarla de manera completa.

Finalmente se seleccionó la biblioteca LibBNT [18] para la implementación las RBD y la biblioteca LibPGM [4] para las RB.

3.7 Alcances y limitaciones

Luego de evaluar y finalmente seleccionar las opciones existentes para poder implementar las redes, fue necesario ajustar algunos de los factores utilizados en los experimentos.

Inicialmente se planeaba trabajar utilizando datos continuos, pero debido a que no existe un soporte completo para este tipo de datos fue necesario discretizar los datos y utilizar nodos discretos. Un ejemplo de estos es la biblioteca LibPMG [4]. La biblioteca permite la creación y carga de datos continuos en una RBD, pero solamente soporta

algoritmos de inferencia sobre datos discretos.

El tamaño del slice y el tamaño de los nodos también tuvieron que modificarse. Esto debido al intensivo uso de recursos computacionales requeridos por las bibliotecas. En el caso particular de LibBMT [18], eran necesario más de 100 giga bytes de memoria para poder utilizar nodos con tamaño 25.

Debido a que no se contaba con un hardware con más recursos en el que se pudieran preparar y correr los experimentos utilizando estos valores se limitaron los valores de estos factores en los experimentos. La sección 3.6.2 describe los valores que fueron utilizados en los experimentos.

Resultados

Antes de analizar los resultados obtenidos se realizó un análisis estadístico de los mismos. Esto con el fin de determinar si los resultados obtenidos en los experimentos se debieron al experimento en sí o a se debieron al azar.

En primera instancia, se realizó un análisis de normalidad sobre cada una de las variables de respuesta. En todos los casos las pruebas indicaron que los datos no eran normales. Debido a que la prueba ANOVA [24] tiene como supuesto que los datos son normales, no fue posible utilizarla. Por lo que se utilizaron métodos no paramétricos para realizar este análisis (sección 4.2).

Para determinar la normalidad de los datos se utilizó la prueba de normalidad de Shapiro-Wilk [24]. Y para determinar que las poblaciones de cada factor para una variable de respuesta era independientes se utilizaron las pruebas no paramétricas Wilcoxon-Mann-Whitney, [24] para factores con 2 valores, y Kruskal-Wallis [24] para factores con 2 o más valores.

4.1 Resumen de resultados

Los resultados obtenidos en los experimentos realizados se resumen en la tabla 4.1 utilizando la siguiente codificación:

- Tipo de red
 - RB: Red bayesiana
 - RBD: Red bayesiana dinámica
- Conjunto de datos
 - PN: productividad del cultivo medida en peso neto
 - EE: Estado de evolución de la enfermedad
- Tamaño del nodo

4.1. Resumen de resultados

- 3: nodo que puede tomar valores entre 1 y 3
- 5: nodo que puede tomar valores entre 1 y 5

- Tamaño del slice
 - T1: slice tamaño de una semana
 - T2: slice tamaño de dos semanas
 - T3: slice tamaño de tres semanas
 - T4: slice tamaño de cuatro semanas
 - T5: slice tamaño de cinco semanas

- F1 micro

- F1 macro

4.1. Resumen de resultados

Tipo de red	Conjunto de datos	Tamaño del nodo	Tamaño del slice	F1 micro	F1 macro
RB	PN	3	T1	0.49153	0.16384
RB	PN	3	T2	0.49153	0.16384
RB	PN	3	T3	0.49587	0.16529
RB	PN	3	T4	0.49573	0.16524
RB	PN	3	T5	0.49573	0.16524
RB	PN	5	T1	0.19672	0.039344
RB	PN	5	T2	0.19672	0.039344
RB	PN	5	T3	0.21212	0.042424
RB	PN	5	T4	0.21538	0.043077
RB	PN	5	T5	0.22857	0.045714
RB	EE	3	T1	0.41989	0.13996
RB	EE	3	T2	0.41783	0.13928
RB	EE	3	T3	0.41783	0.13928
RB	EE	3	T4	0.41573	0.13858
RB	EE	3	T5	0.4136	0.13787
RB	EE	5	T1	0.014388	0.0028777
RB	EE	5	T2	0.014388	0.0028777
RB	EE	5	T3	0.014388	0.0028777
RB	EE	5	T4	0.014388	0.0028777
RB	EE	5	T5	0.014388	0.0028777
RBD	PN	3	T1	0.275	0.091666667
RBD	PN	3	T2	0.381818182	0.127272727
RBD	PN	3	T3	0.292682927	0.097560976
RBD	PN	3	T4	0.42519685	0.141732283
RBD	PN	3	T5	0.208955224	0.069651741
RBD	PN	5	T1	0.138888889	0.027777778
RBD	PN	5	T2	0.170731707	0.034146341
RBD	PN	5	T3	0.14084507	0.028169014
RBD	PN	5	T4	0.24137931	0.048275862
RBD	PN	5	T5	0.247933884	0.049586777
RBD	EE	3	T1	0.311345646	0.103781882
RBD	EE	3	T2	0.40917782	0.136392607
RBD	EE	3	T3	0.271386431	0.090462144
RBD	EE	3	T4	0.306451613	0.102150538
RBD	EE	3	T5	0.297520661	0.099173554
RBD	EE	5	T1	0.156626506	0.031325301
RBD	EE	5	T2	0.173669468	0.034733894
RBD	EE	5	T3	0.225596529	0.045119306
RBD	EE	5	T4	0.194373402	0.03887468
RBD	EE	5	T5	0.177285319	0.035457064

Table 4.1: Resumen de los resultados de los experimentos

4.1.1 Análisis estadístico de los resultados

En esta sección se realiza un análisis estadístico de los resultados obtenidos.

4.1.2 F1 Micro

4.1.2.1 Análisis de normalidad

Resultados de la prueba de normalidad Shapiro-Wilk aplicada sobre la variable de respuesta F1 Micro.

Prueba: Shapiro-Wilk

Dato: f1 micro

Valor p: 0.02454

El valor p obtenido es menor que 0.05 en la prueba Shapiro-Wilk pruebas de normalidad. Por esto se puede concluir con un nivel de certeza del 95% que el F1 micro no tiene una distribución normal. La figura 4.1 es el gráfico cuantil-cuantil del F1 Micro donde se visualiza el resultado de manera gráfica.

Dado que el F1 Micro no sigue una distribución normal, se hizo su análisis utilizando pruebas no paramétricas. Para factores con dos niveles se utilizó la prueba Wilcoxon, para los factores con más niveles se utilizaron las pruebas Kruskal-Wallis y la prueba post-hoc de Nemenyi.

4.1.2.2 Tipo de red

Prueba: Wilcoxon-Mann-Whitney

Dato: F1 micro por tipo de red

Valor p: 0.3884

Como el valor p es mayor que 0.05 ($0.3884 > 0.05$), el resultado de la prueba indica que según las muestras con las que se cuenta, no hay evidencia suficiente para determinar si las poblaciones son independientes. El gráfico 4.2 muestra cómo los valores del F1 micro con ambos tipos de redes se intersecan, evidenciando lo obtenido en la prueba Wilcoxon-Mann-Whitney.

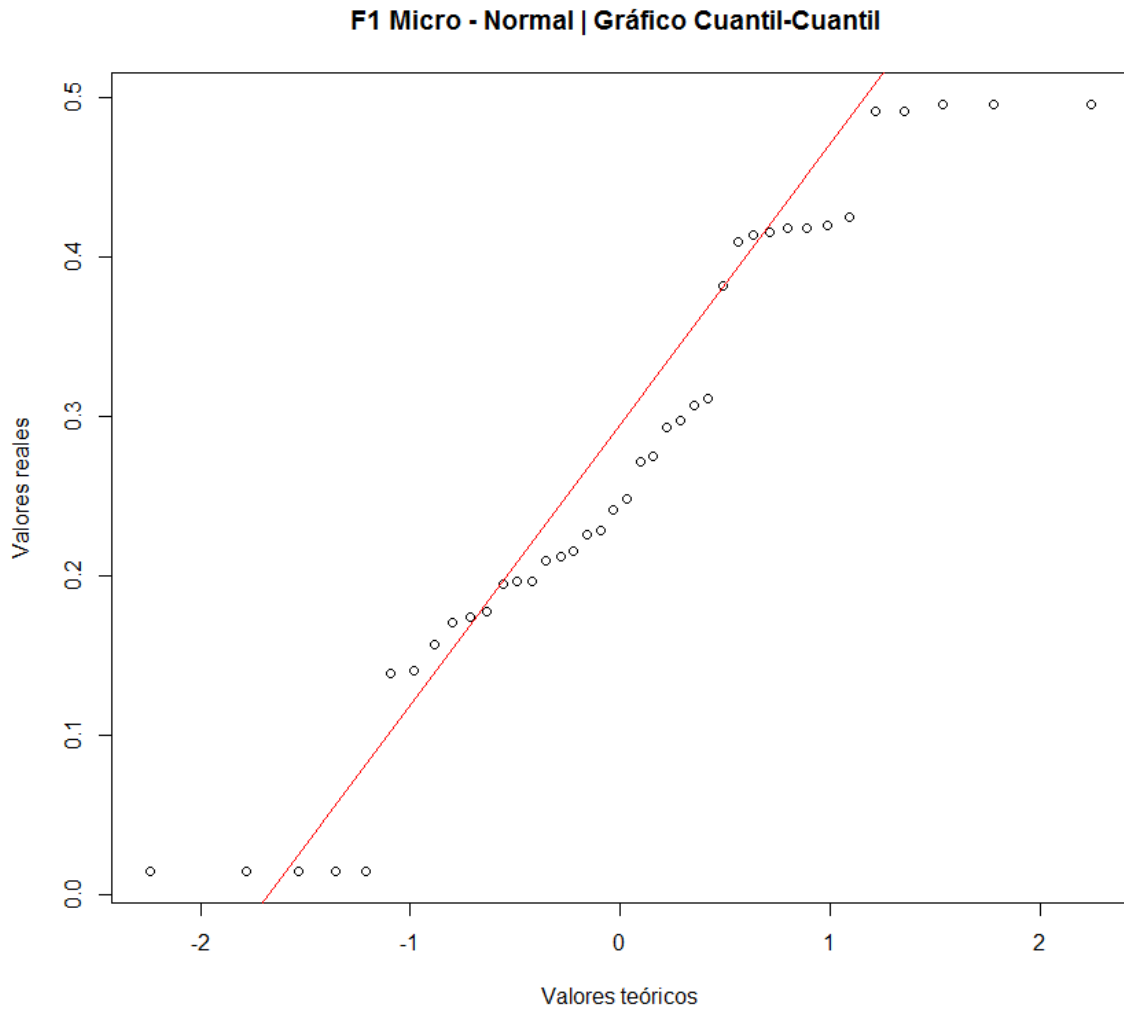


Figura 4.1: Gráfico cuantil-cuantil del F1 micro.

4.1.2.3 Conjunto de datos

Prueba: Wilcoxon-Mann-Whitney

Dato: F1 micro por conjunto de datos

Valor p : 0.01447

Como el valor p obtenido con la prueba Wilcoxon-Mann-Whitney es menor que el valor alfa ($0.01447 < 0.05$), se puede concluir con un nivel de certeza del 95% que las poblaciones son independientes. Este resultado puede observarse en el gráfico 4.3.

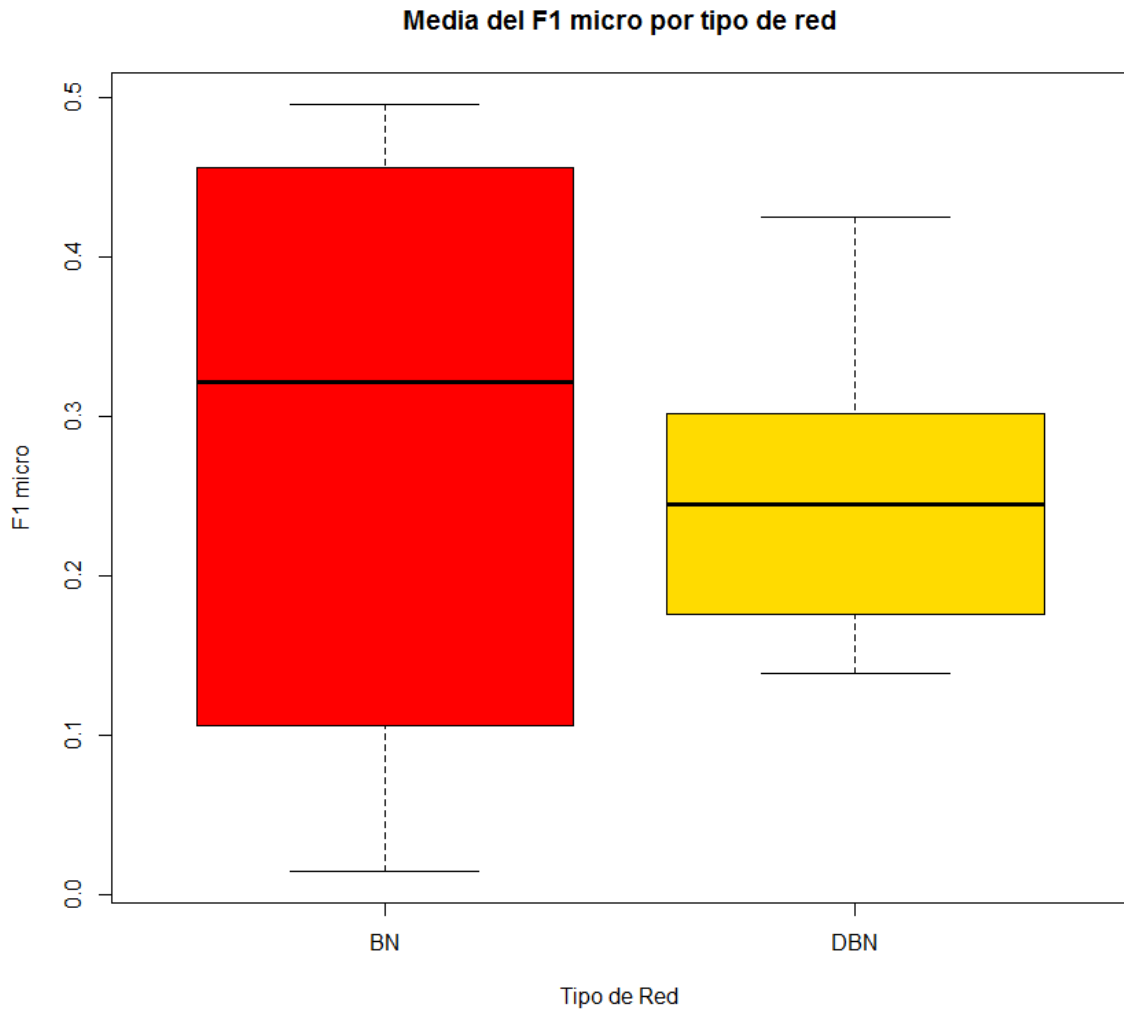


Figura 4.2: Gráfico de las medias del F1 micro por tipo de red.

4.1.2.4 Tamaño del nodo

Prueba: Wilcoxon-Mann-Whitney

data: F1 micro por tamaño del nodo

Valor p: 0.0001113

Como el valor p obtenido con la prueba Wilcoxon-Mann-Whitney es menor que el valor alfa ($0.0001113 < 0.05$), se puede concluir con un nivel de certeza del 95% que las poblaciones son independientes. Este resultado puede observarse en el gráfico 4.4.

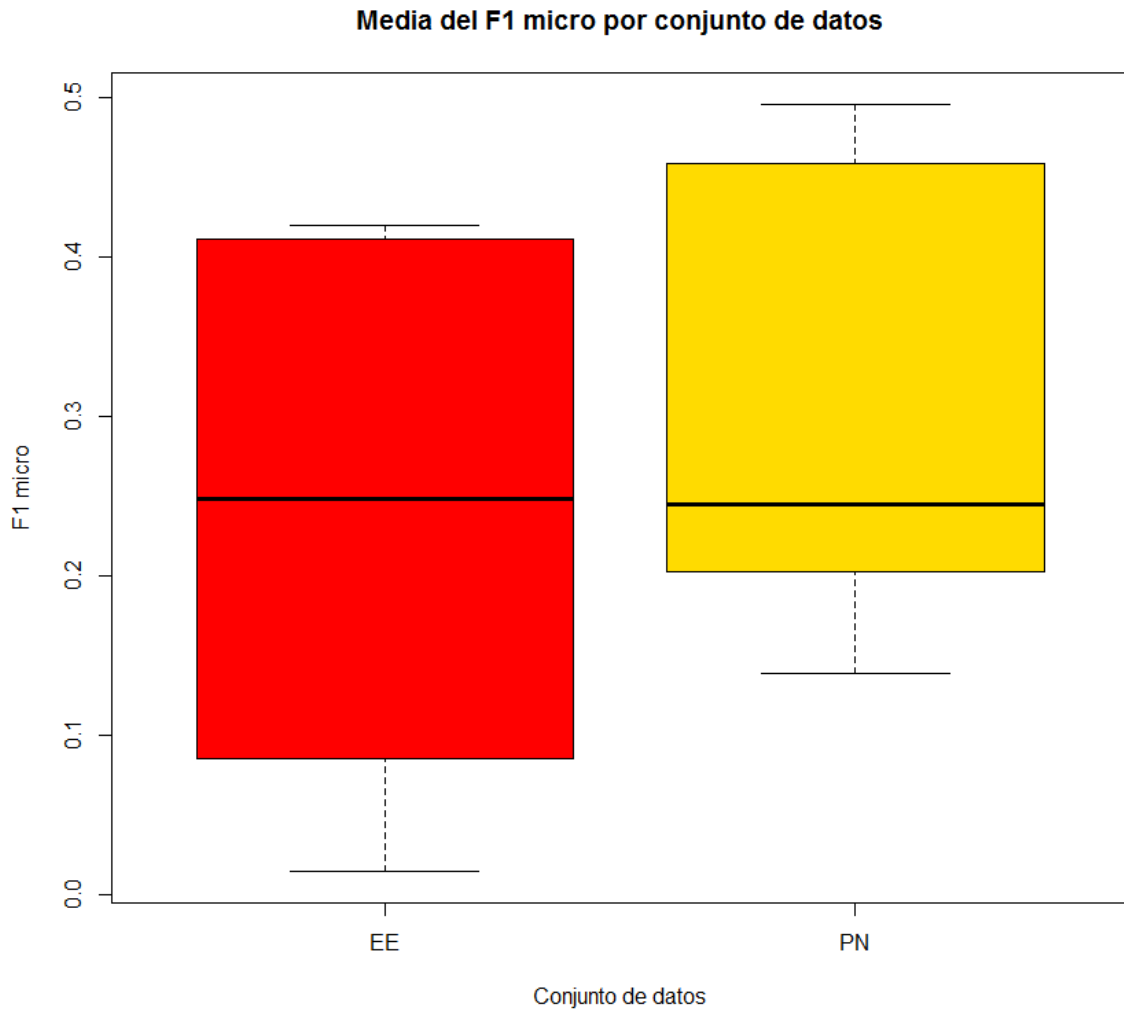


Figura 4.3: Gráfico de las medias del F1 micro por conjunto de datos.

4.1.2.5 Tamaño del slice

Prueba: Kruskal-Wallis

Dato: F1 micro por tamaño del slice

Valor p: 0.9756

Como el valor p es mayor que 0.05 ($0.9756 > 0.05$), el resultado de la prueba Kruskal-Wallis indica que, según las muestras con las que se cuenta, no hay evidencia suficiente para determinar si las poblaciones son independientes. El gráfico 4.5 muestra cómo

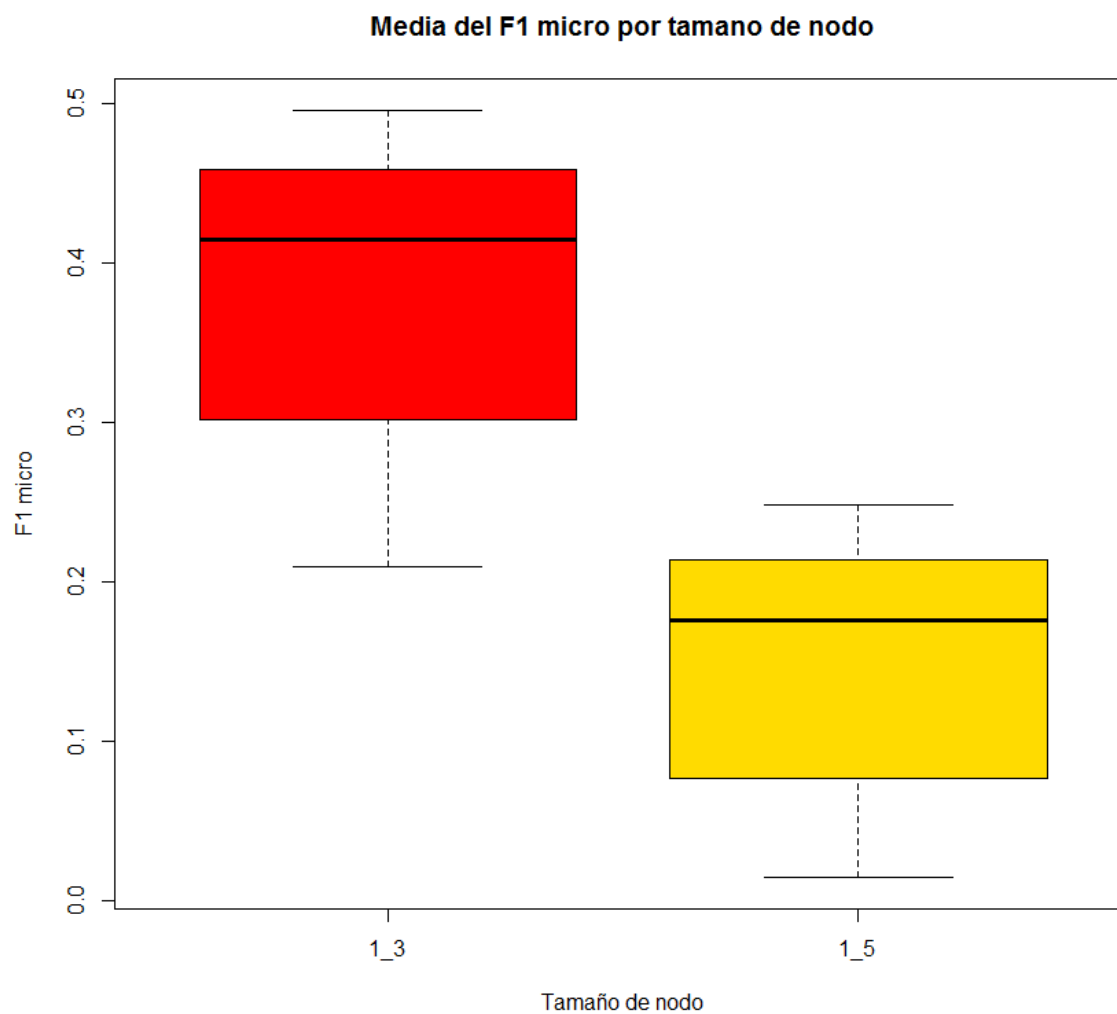


Figura 4.4: Gráfico de las medias del F1 micro por tamaño de nodo.

los valores del F1 micro para todos los tamaños de slice se intersecan, evidenciando el mismo resultado.

4.1.3 F1 Macro

4.1.3.1 Análisis de normalidad

Prueba: Shapiro-Wilk

Dato: F1 macro

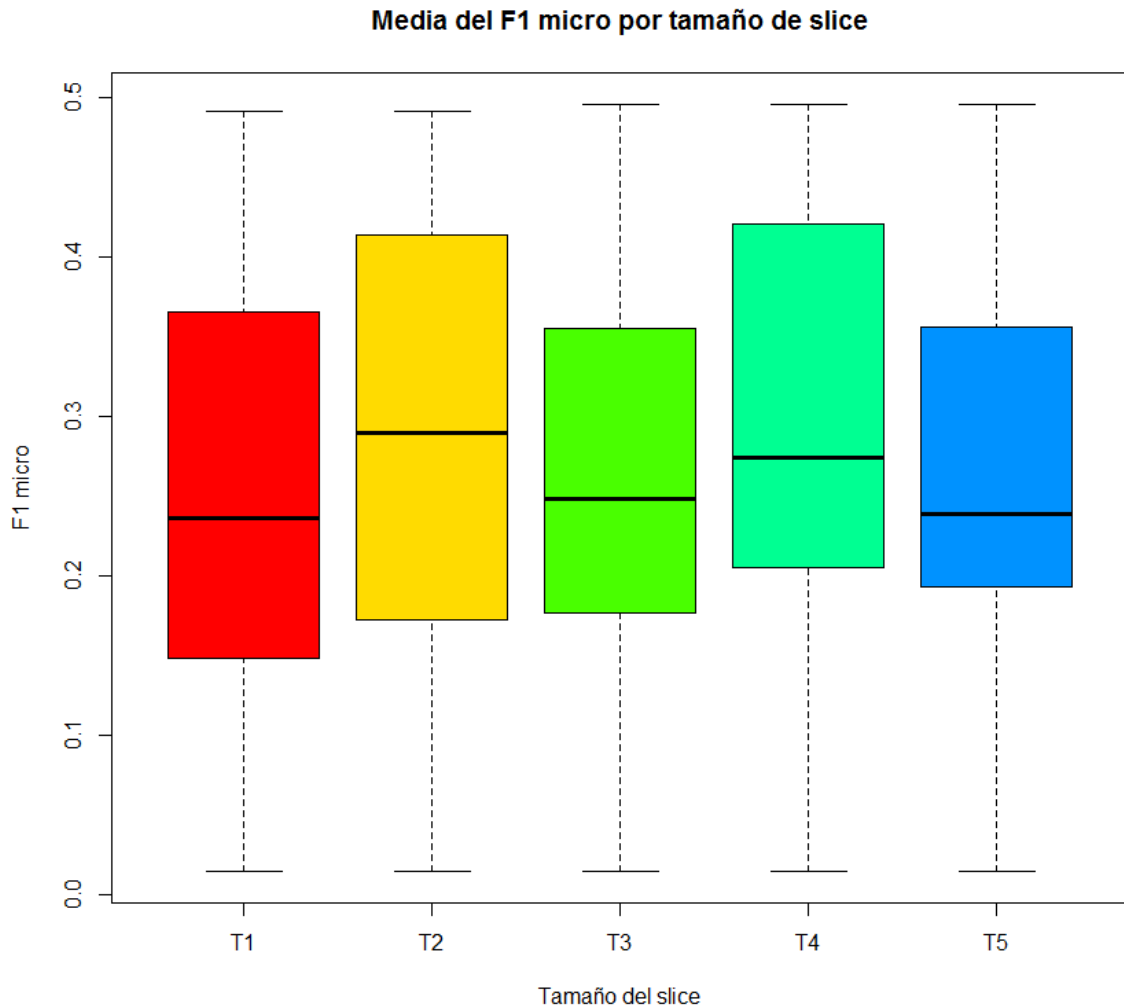


Figura 4.5: Gráfico de las medias del F1 micro por tamaño de slice.

Valor p : 0.00128

El valor p obtenido es menor que 0.05 en la prueba Shapiro-Wilk pruebas de normalidad. Por esto se puede concluir con un nivel de certeza del 95% que el F1 micro no tiene una distribución normal. La figura 4.1 es el gráfico cuantil cuantil del F1 Micro donde se visualiza el resultado de manera gráfica.

Dado que el F1 Macro no sigue una distribución normal, se hizo su análisis utilizando pruebas no paramétricas. Para factores con dos niveles se utilizó la prueba Wilcoxon, para los factores con más niveles se utilizaron las pruebas Kruskal-Wallis y la prueba

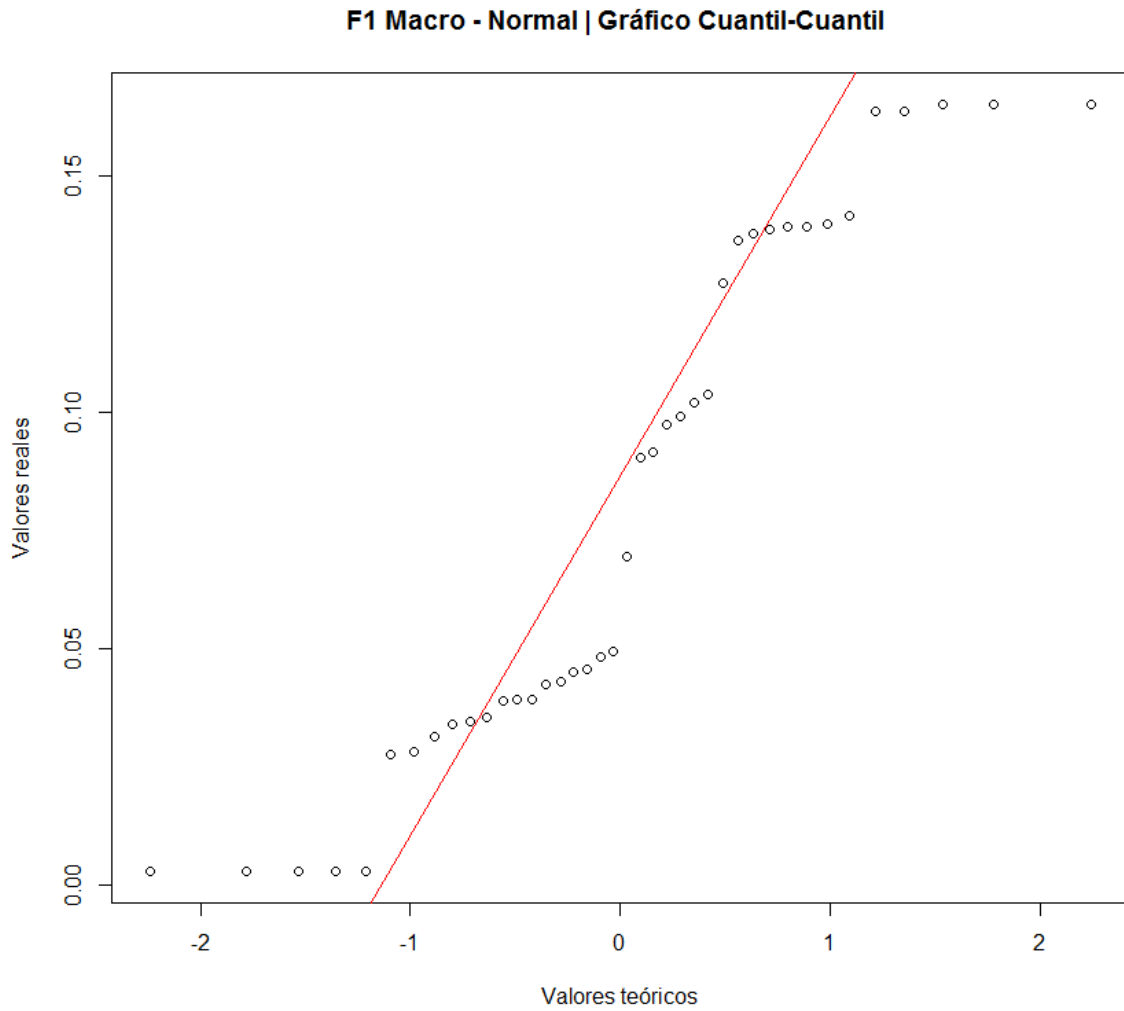


Figura 4.6: Gráfico cuantil-cuantil del F1 macro.

post-hoc de Nemenyi.

4.1.3.2 Tipo de red

Prueba: Wilcoxon-Mann-Whitney

Dato: F1 macro por tipo de red

Valor p : 0.09731

Como el valor p es mayor que 0.05 ($0.09731 > 0.05$), el resultado de la prueba indica que, según las muestras con las que se cuenta, no hay evidencia suficiente para

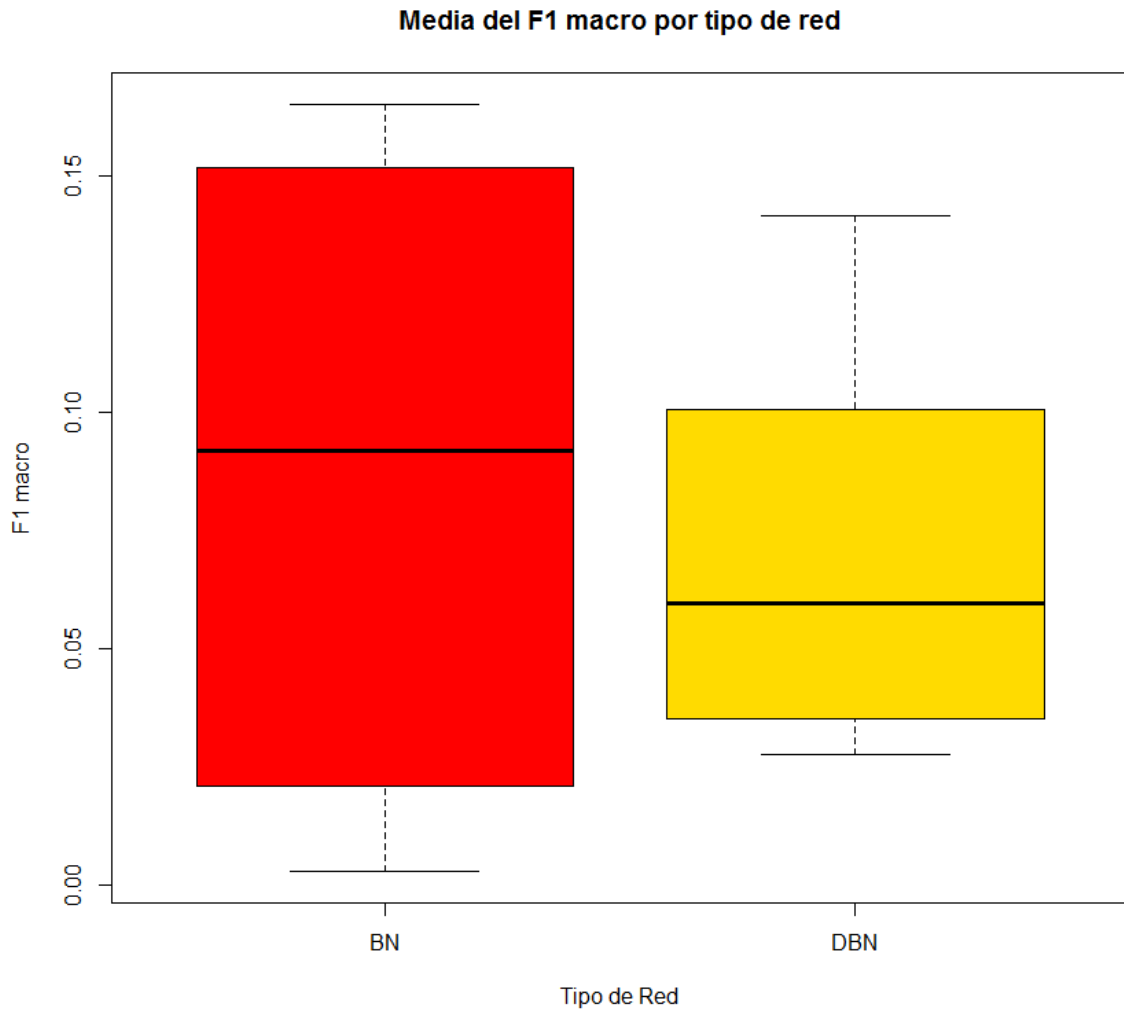


Figura 4.7: Gráfico de las medias del F1 micro por tipo de red.

determinar si las poblaciones son independientes. El gráfico 4.7 muestra como los valores del F1 micro con ambos tipos de redes se intersecan, evidenciando lo obtenido en la prueba Wilcoxon-Mann-Whitney.

4.1.3.3 Conjunto de datos

Prueba: Wilcoxon-Mann-Whitney

Dato: F1 macro por conjunto de datos

Valor p: 0.009463

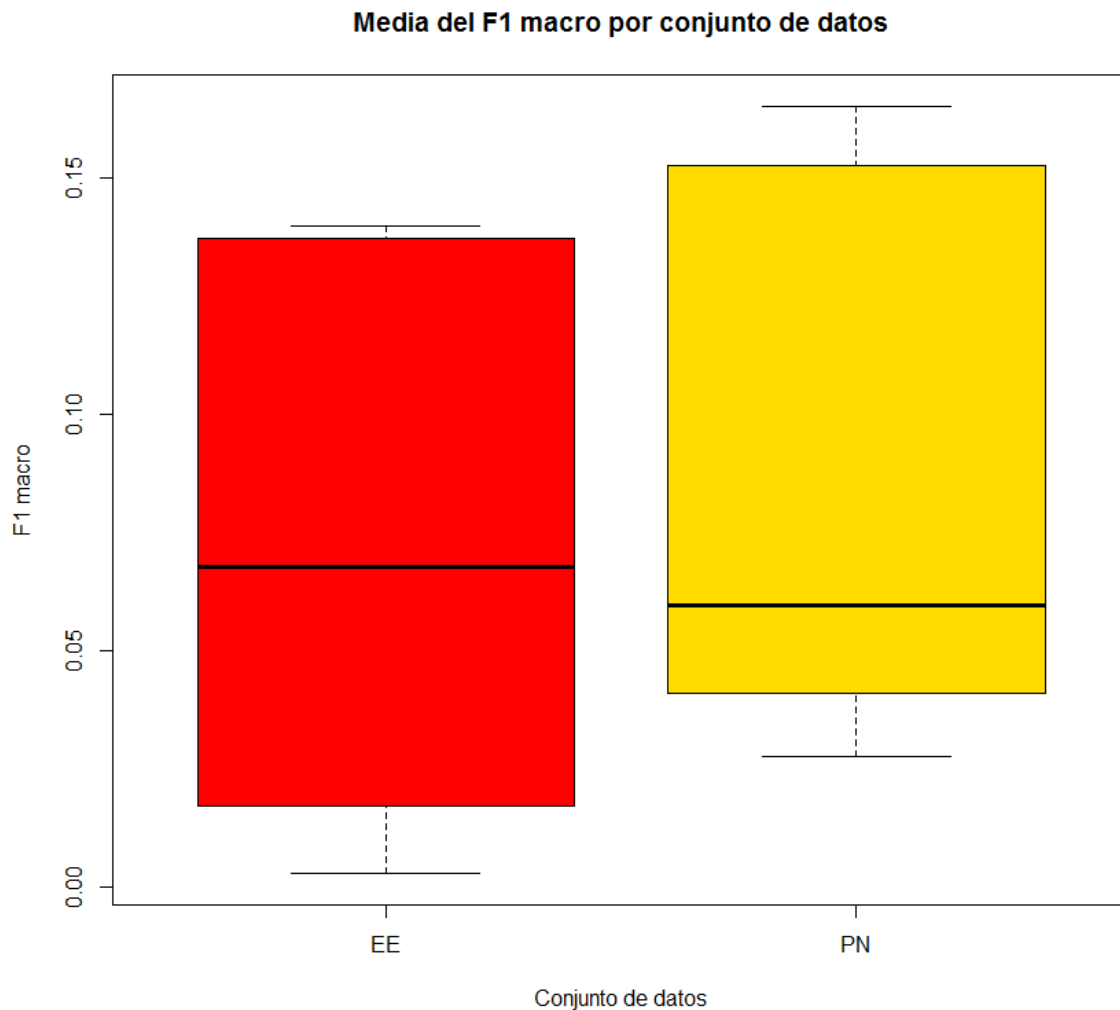


Figura 4.8: Gráfico de las medias del F1 macro por conjunto de datos.

Como el valor p obtenido con la prueba Wilcoxon-Mann-Whitney es menor que el valor α ($0.009463 < 0.05$), se puede concluir con un nivel de certeza del 95% que las poblaciones son independientes. Este resultado puede observarse en el gráfico 4.8.

4.1.3.4 Tamaño del nodo

Prueba: Wilcoxon-Mann-Whitney

Dato: F1 macro por tamaño de nodo

Valor p : $9.542e-05$

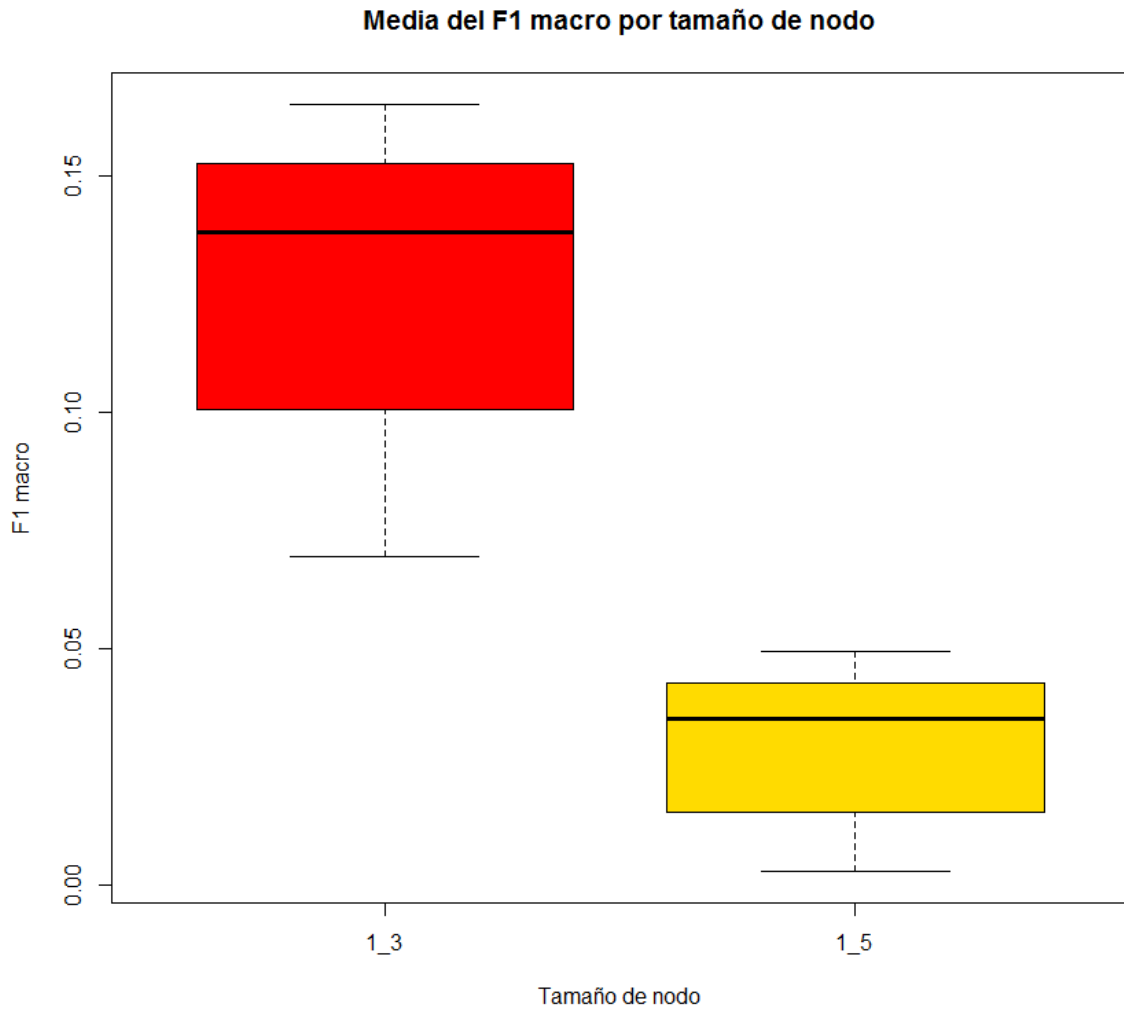


Figura 4.9: Gráfico de las medias del F1 macro por tamaño de nodo.

Como el valor p obtenido con la prueba Wilcoxon-Mann-Whitney es menor que el valor α ($9.542e-05 < 0.05$), se puede concluir con un nivel de certeza del 95% que las poblaciones son independientes. Este resultado puede observarse en el gráfico 4.9.

4.1.3.5 Tamaño del slice

Prueba: Kruskal-Wallis

Dato: F1 macro por tamaño de slice

Valor p : 0.9802

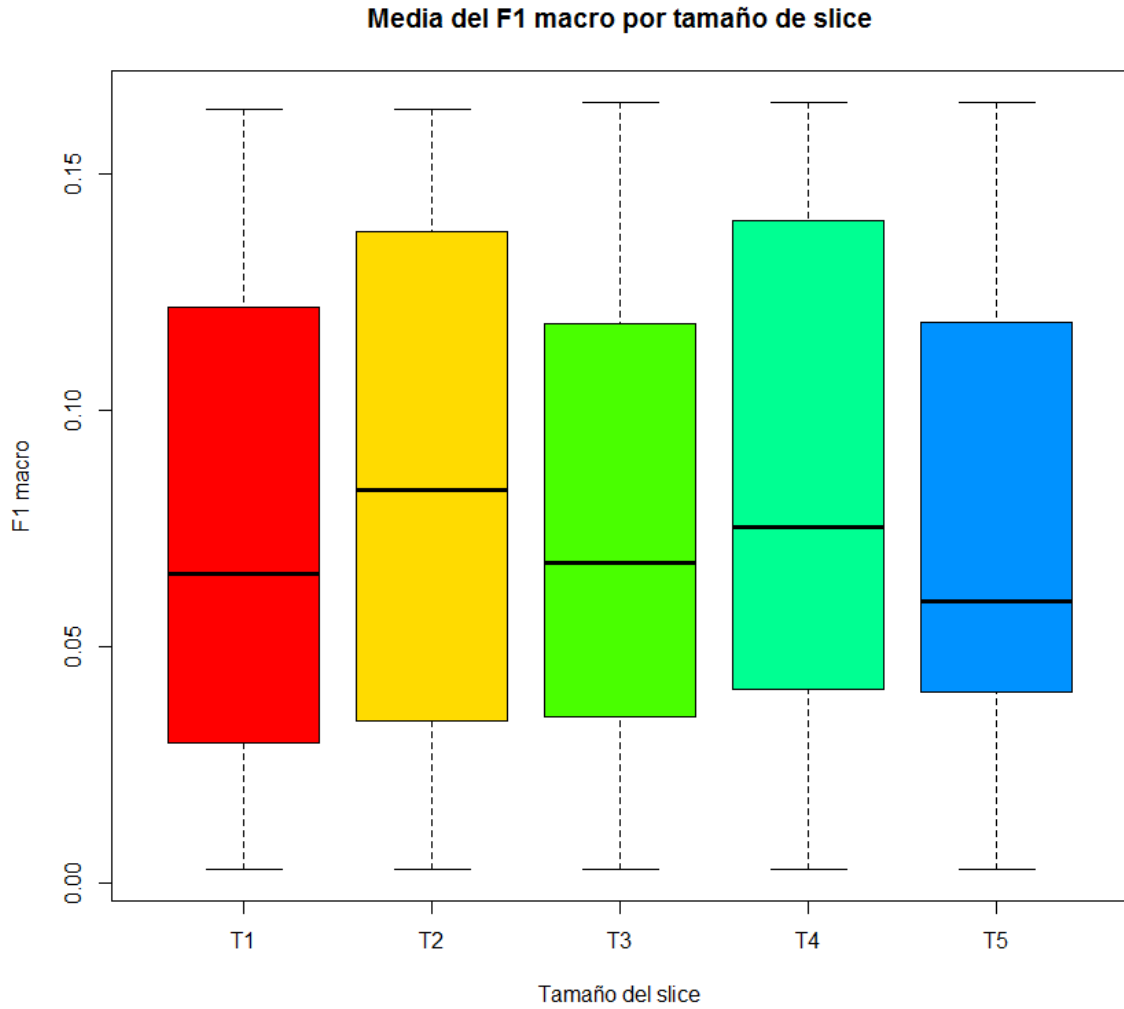


Figura 4.10: Gráfico de las medias del F1 macro por tamaño de slice.

Como el valor p es mayor que 0.05 ($0.9802 > 0.05$), el resultado de la prueba Kruskal-Wallis indica que, según las muestras con las que se cuenta, no hay evidencia suficiente para determinar si las poblaciones son independientes. El gráfico 4.10 muestra como los valores del F1 micro con ambos tipos de redes se intersecan, evidenciando el mismo resultado.

4.2 Análisis de resultados

En esta sección se analizan los resultados obtenidos en la sección 4.1.1.

4.2.1 Tipo de red

Los resultados indican que el tipo de red no tiene una incidencia significativa en las variables de respuesta. Para representar los datos semanales (los slice en la RB) se construyó la red de manera que la misma estructura de una semana se repite y conecta con la de la siguiente. Esto conceptualmente es lo que el modelo de las RBD permite representar ya que, utilizando la 2TBN, representan los slices de manera intrínseca. Es por esta razón que el resultado obtenido no resulta sorprendente.

Este resultado además es un indicador de que ambas implementaciones están prediciendo los datos de manera similar.

4.2.2 Conjunto de datos

Los resultados obtenidos indican que los distintos conjuntos de datos inciden significativamente en los resultados. Esto resulta lógico ya que los conjuntos de datos tienen la información de fenómenos distintos: la productividad del cultivo y el estado de la enfermedad.

Conjunto de datos	Promedio F1 Micro	Promedio F1 Macro
EE	0.2340	0.0713
PN	0.3022	0.0875

Table 4.2: Promedio por conjunto de datos.

Como puede verse en la tabla 4.2, las variables de respuesta siempre fueron mayores, es decir el clasificador fue más efectivo, para los experimentos en los cuáles se utilizaron

los datos de la productividad del cultivo en contraste con los resultados obtenidos al utilizar los datos del estado de la enfermedad.

4.2.3 Tamaño del nodo

Los resultados indican que las redes tienen más efectividad al predecir cuando se utiliza tamaño de nodo 3 en comparación con cuando se utiliza tamaño de nodo 5. El tamaño del nodo, es decir la cantidad de valores que se utilizan para representar cada una de las variables en la red, determina el tamaño de las tablas de distribución conjunta que cada uno de los nodos tiene asociado.

Las redes aprenden a representar el valor de un nodo al ajustar las probabilidades asociadas a cada uno de sus posibles valores según las observaciones de los valores de sus padres y su valor. Con cada observación, en el caso de los experimentos realizados en este trabajo las observaciones en el conjunto de entrenamiento, la red ajusta las probabilidades de cada uno de los nodos. Al presentarse nueva evidencia la red puede calcular basado en las probabilidades ya calculadas el valor esperado o predicho.

Al aumentar el tamaño del nodo lo que se está haciendo es aumentando la granularidad con la que se representa cada uno de las variables en la red. Considerando lo anterior y el hecho de que se utilizó la misma cantidad de datos para entrenar las redes para todos los tamaños de nodo, resulta esperado que cuando se aumentó el tamaño del nodo de 3 a 5 la eficacia de la red disminuyera.

4.2.4 Tamaño del slice

El tamaño de slice, es decir la cantidad de semanas sobre las cuales la red se expande para predecir el valor para la siguiente semana, no tiene una incidencia significativa en las variables de respuesta según los resultados obtenidos.

Este resultado no era el esperado, pero puede deberse a que el tamaño del slice, dadas las características de los datos que se están utilizando, no fue lo suficientemente

grande para que fuera significativo.

4.3 Verificación del modelo

Cuando se comenzó a evaluar el desempeño de la RBD se observó un pobre desempeño por parte de la misma. Dado que se cuenta con archivos de datos con un número reducido de elementos, alrededor de cientos, se sospechó que el mal desempeño de la red se debía a esto.

Se realizó un paso adicional para verificar que la implementación de la RBD es correcta y que ésta puede aprender la información representada en los datos y luego realizar predicciones. Este paso adicional consistió en seguir los mismos pasos descritos en el experimento final para entrenar y evaluar el desempeño de la RBD pero utilizando otros conjuntos de datos.

Primero se crearon archivos de datos basado en los datos del estado de evolución de la enfermedad repitiendo el contenido del archivo 1, 2, 3, 4 y 5 veces. Además se generó un archivo con 500 datos aleatorios que fueron repetidos de la misma manera en archivos con 1, 5, 10, 25 50 los mismos datos. Esto con el fin de generar patrones que podrían ser aprendidos por la red.

Como puede observarse en la figuras 4.11 y 4.12, conforme aumenta la cantidad de datos, la eficacia de la red también aumenta. Este comportamiento se mantuvo para ambos tipos de datos. Los resultados pueden verse en el apéndice C.

Este resultado sugiere que la reducida cantidad de muestras utilizadas hace que baje la eficacia, y que si llegara a contarse con más datos la eficacia de la red aumentaría. Esto sucedería debido a que se estaría aumentando la cantidad de muestras que se utilizan para entrenar a la red y por ende ella podría ajustarse mejor.

Los archivos utilizados en esta prueba pueden encontrarse en el apéndice C.

4.3. Verificación del modelo

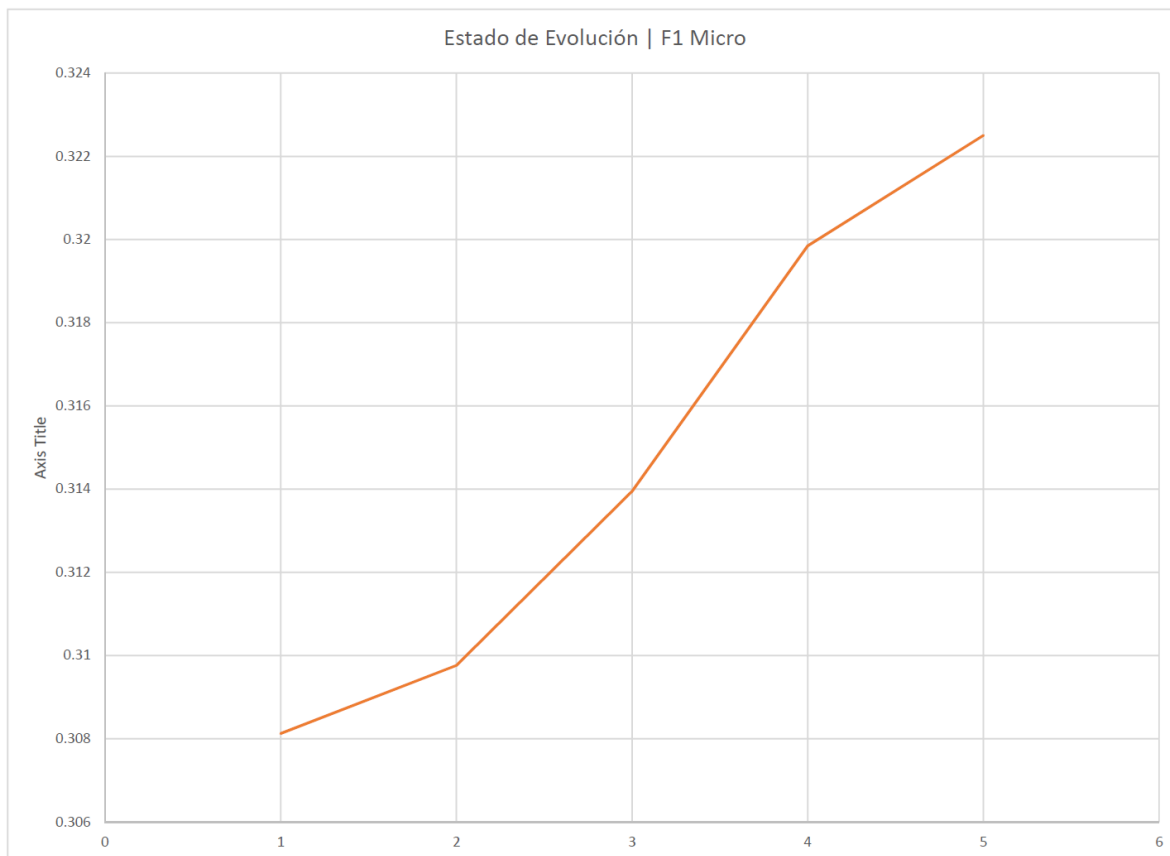


Figura 4.11: F1 micro promedio utilizando datos del Estado de Evolución. El eje X tiene la cantidad de veces que se repitieron los datos en el archivo de entrada.

4.3. Verificación del modelo

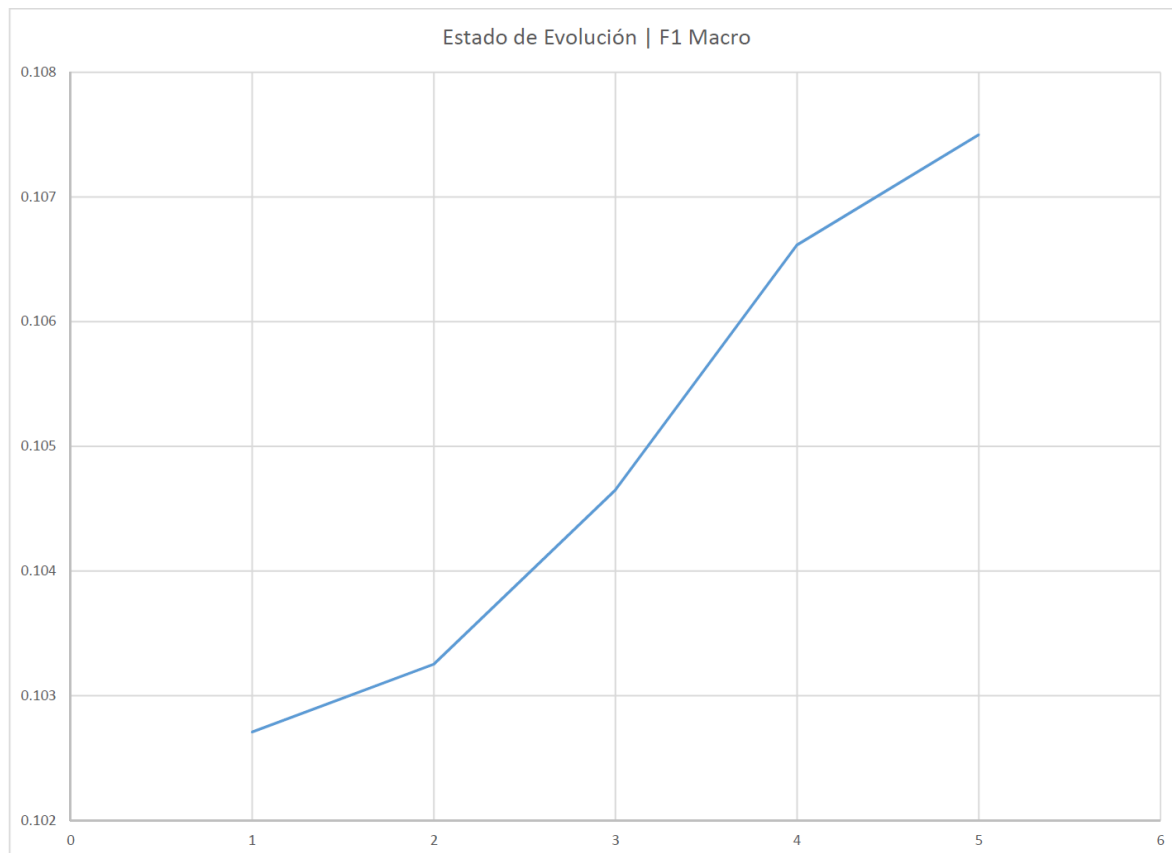


Figura 4.12: F1 macro promedio utilizando datos del Estado de Evolución. El eje X tiene la cantidad de veces que se repitieron los datos en el archivo de entrada.

Conclusiones y recomendaciones

5.1 Conclusiones

Los resultados obtenidos en los experimentos mostraron que la capacidad de predicción de las RBDs no supera la de las RBs utilizando los datos de CORBANA [6]. De hecho no se observó una diferencia significativa entre ambos tipos de red. Por lo anterior se rechaza la hipótesis planteada.

Los resultados obtenidos en los experimentos mostraron una efectividad baja por parte de las RBD. Aún para el mejor resultado, el cual se obtuvo al utilizar los datos de producción y nodos de tamaño 3, la efectividad de predicción de la red fue baja ya que se obtuvo un F1 micro de 0.49 y un F1 macro de 0.16. Por esta razón se concluye que las RBDs no son la mejor alternativa para predecir este tipo de fenómenos cuando se cuenta con datos similares a los utilizados en esta investigación.

La eficacia en la predicción de la red se ve influenciada por la estructura del grafo y el tamaño de los nodos. En el caso del grafo, esto se debe a que los arcos entre los nodos determinan cómo la influencia entre los nodos pasa a través de la red. El tamaño del nodo influye ya que incrementa o disminuye la cantidad de etiquetas que la red deberá aprender a predecir.

Las RBDs en esencia permiten representar datos temporales al replicar y conectar múltiples RBs. Su representación compacta las hace atractivas al trabajar con problemas que tienen un gran número de variables al compararlas con otros MGPs como las RBs y los MOMs.

A pesar de las ventajas que el modelo de las RBDs presenta en la teoría frente a otros MGPs como los MOMs y las RB, en la práctica las limitaciones de las implementaciones disponibles hacen que no sean tan atractivas. Inclusive LibBNT, una de las implementaciones evaluadas más completas que fue creada por los considerados padres del modelo, resultó requerir una cantidad muy alta de recursos computacionales para poder operar correctamente.

5.1.1 Recomendaciones

A partir del análisis de los resultados obtenidos y las observaciones que se realizaron durante las distintas fases de esta investigación se resumen en esta sección algunas recomendaciones con respecto al uso de las RBDs.

- Tamaño de los nodos

Como se discutió en la sección *Conclusiones* (ver sección 5.1). Las implementaciones que soportan RBDs -al menos las evaluadas- no son tan eficientes como el modelo es en teoría. Es por ello que al modelar redes, principalmente redes con gran cantidad de nodos, se debe reducir el tamaño de los nodos en los casos donde se enfrenten problemas con la cantidad de recursos o por tiempos muy elevados al utilizar la red. Por ejemplo, al entrenar la red.

- Estructura de la Red

La estructura de la red debe representar acertadamente las relaciones que existen entre las variables que se quieren modelar. Por lo tanto, se deben estudiar las relaciones entre las variables involucradas y poner especial atención a la hora de traducirlas a la red para hacerlo adecuadamente. Debe considerarse siempre el concepto de la separación d (d-separation en inglés) y asegurarse que la influencia entre los nodos fluya de manera adecuada en la red.

5.1.2 Trabajo futuro

Se presentan seguidamente algunas reflexiones sobre el trabajo a realizar en el futuro:

- Usar distintos tamaños de slice mayores que 5

En esta investigación se utilizaron tamaños de slice entre 1 y 5; lo que representa de 1 a 5 semanas. Se limitó el tamaño del slice a estos valores por el intensivo uso

de recursos computacionales requeridos por las bibliotecas tal como se describe en la sección 3.7. Los resultados obtenidos indican que este factor no tuvo un impacto significativo sobre las variables de respuesta. Es por esto que surge la pregunta de si un tamaño de slice mayor podría haber tenido un mayor impacto y, de ser así, cuál efecto tendría sobre la eficacia de predicción de la red.

- Valores iniciales

Las probabilidades de los nodos de las redes utilizadas para esta investigación fueron inicializadas utilizando valores aleatorios. En la presente investigación esto no fue considerado debido a que se contaba con muy pocos elementos en los conjuntos de datos. De utilizarse parte de ellos para determinar los valores iniciales no quedarían elementos disponibles para verificar el modelo. Queda por determinar el efecto que podría tener inicializarlas siguiendo otro criterio.

- Utilizar otros conjuntos de datos

Sería interesante experimentar con otros conjuntos de datos que cuenten con un mayor número de elementos, y determinar si con ellos se consigue un resultado distinto. En los experimentos realizados

Resultados complementarios

En la tabla A.01 se muestran todos los resultados obtenidos en los experimentos.

Conjunto de datos	Tamaño del nodo	T	fl-micro	fl-macro	r-micro	p-micro	r-macro	p-macro	TP	FP	TN	FN
PN	3	1	0.49153	0.16384	0.93548	0.33333	0.31183	0.11111	29	58	4	2
PN	3	2	0.49153	0.16384	0.93548	0.33333	0.31183	0.11111	29	58	4	2
PN	3	3	0.49587	0.16529	0.96774	0.33333	0.32258	0.11111	30	60	2	1
PN	3	4	0.49573	0.16524	0.96667	0.33333	0.32222	0.11111	29	58	2	1
PN	3	5	0.49573	0.16524	0.96667	0.33333	0.32222	0.11111	29	58	2	1
PN	5	1	0.19672	0.039344	0.19355	0.2	0.03871	0.04	6	24	100	25
PN	5	2	0.19672	0.039344	0.19355	0.2	0.03871	0.04	6	24	100	25
PN	5	3	0.21212	0.042424	0.22581	0.2	0.045161	0.04	7	28	96	24
PN	5	4	0.21538	0.043077	0.23333	0.2	0.046667	0.04	7	28	92	23
PN	5	5	0.22857	0.045714	0.26667	0.2	0.053333	0.04	8	32	88	22
EE	3	1	0.41989	0.13996	0.56716	0.33333	0.18905	0.11111	76	152	116	58
EE	3	2	0.41783	0.13928	0.5597	0.33333	0.18657	0.11111	75	150	118	59
EE	3	3	0.41783	0.13928	0.5597	0.33333	0.18657	0.11111	75	150	118	59
EE	3	4	0.41573	0.13858	0.55224	0.33333	0.18408	0.11111	74	148	120	60
EE	3	5	0.4136	0.13787	0.54478	0.33333	0.18159	0.11111	73	146	122	61
EE	5	1	0.014388	0.0028777	0.0074627	0.2	0.0014925	0.04	1	4	532	133
EE	5	2	0.014388	0.0028777	0.0074627	0.2	0.0014925	0.04	1	4	532	133
EE	5	3	0.014388	0.0028777	0.0074627	0.2	0.0014925	0.04	1	4	532	133
EE	5	4	0.014388	0.0028777	0.0074627	0.2	0.0014925	0.04	1	4	532	133
EE	5	5	0.014388	0.0028777	0.0074627	0.2	0.0014925	0.04	1	4	532	133

Table A.01: Resultados de los experimentos



Figura A.1: Gráfico del tiempo requerido para entrenar y evaluar la RBD por tamaño de slice. Datos sintéticos, nodos tamaño 3.

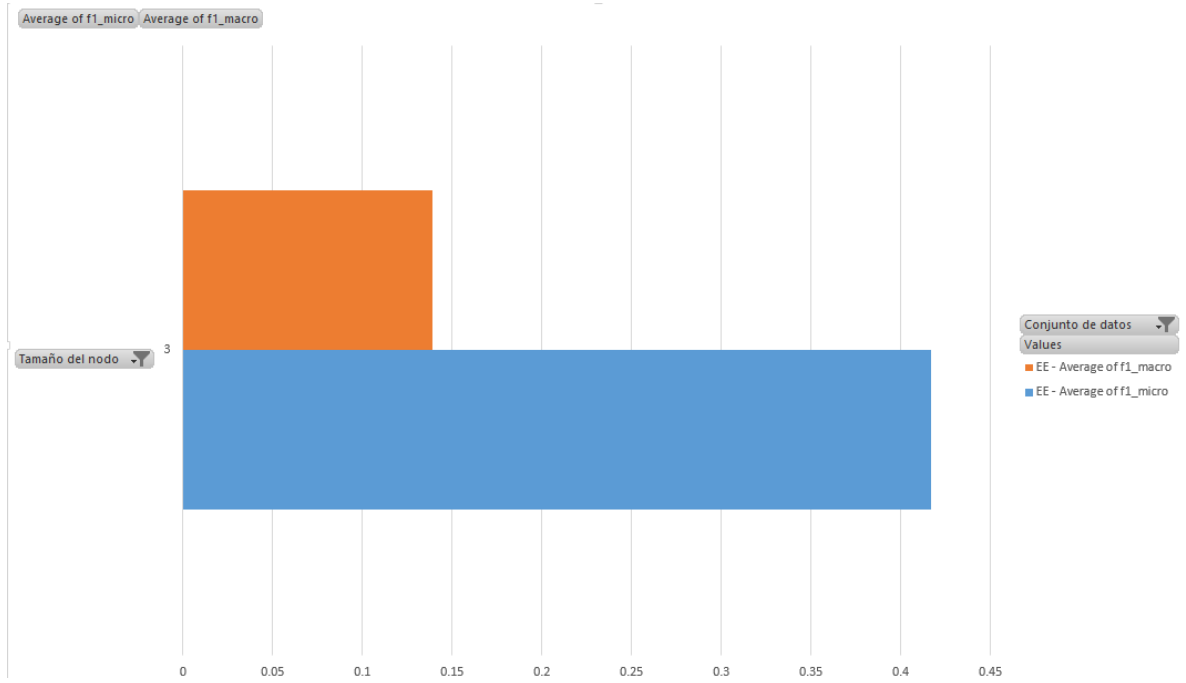


Figura A.2: Gráfico de las medias del F1 micro y macro utilizando el conjunto de datos del Estado de la enfermedad y nodos de tamaño 3.

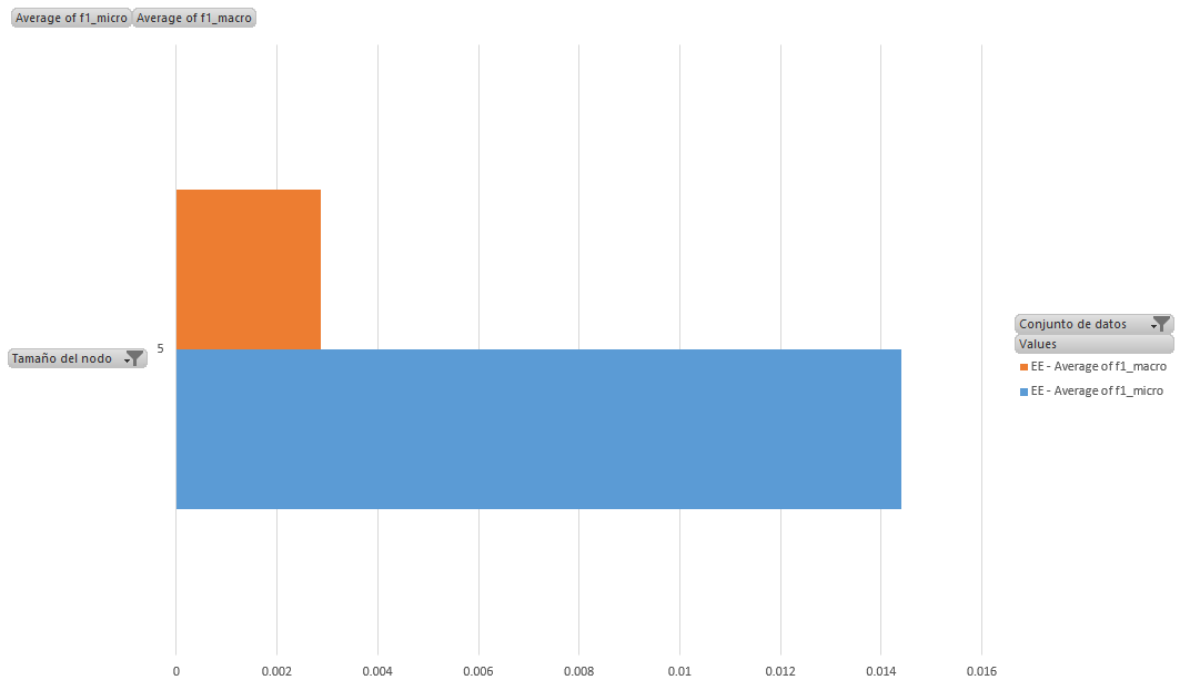


Figura A.3: Gráfico de las medias del F1 micro y macro utilizando el conjunto de datos del Estado de la enfermedad y nodos de tamaño 5.

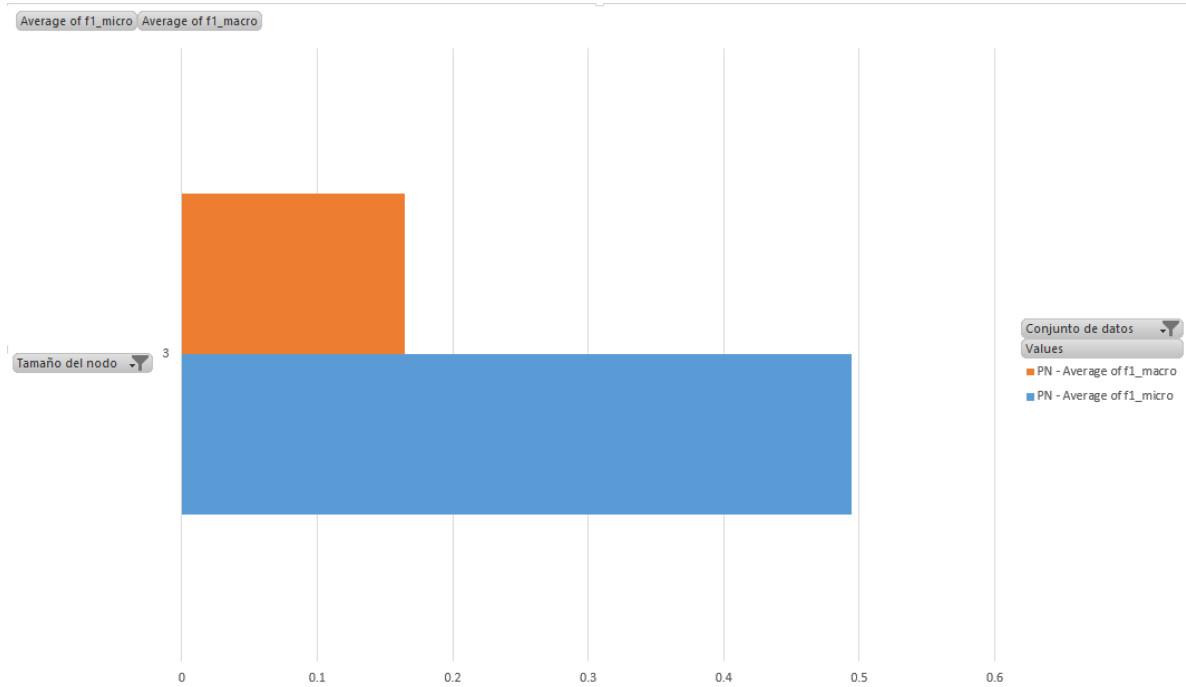


Figura A.4: Gráfico de las medias del F1 micro y macro utilizando el conjunto de datos de la Producción y nodos de tamaño 3.

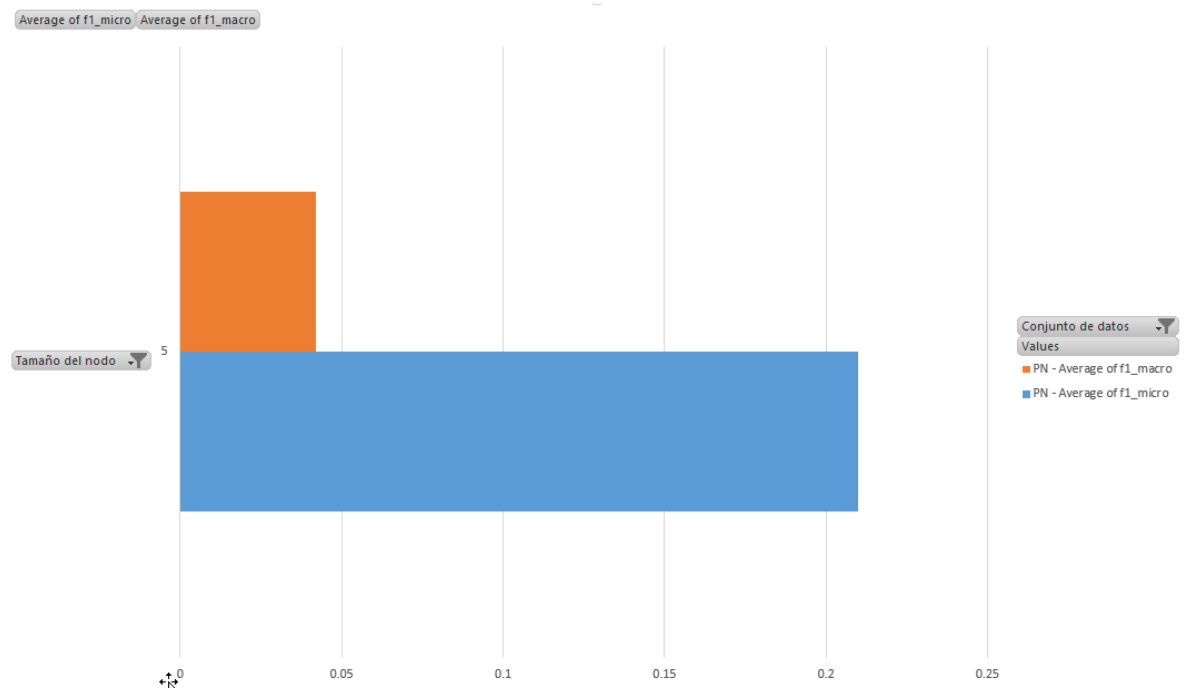


Figura A.5: Gráfico de las medias del F1 micro y macro utilizando el conjunto de datos de la Producción y nodos de tamaño 5.

Recursos en línea

En el sitio web: https://bitbucket.org/sebastian_arguello/black-sigatoka-thesis, puede encontrarse todo el código fuente que fue utilizado para esta investigación incluyendo:

- Código de prueba de las alternativas de biblioteca evaluadas
- El código que genera las RBDs y RB que fueron utilizadas en los experimentos
- El código que ejecuta los experimentos y recompila los resultados
- Scripts en R para el análisis estadístico

Verificación del modelo

Este apéndice contiene los datos utilizados para la verificación de la implementación de la RBD. La tabla C.02 corresponde a los datos que fueron generados de manera aleatoria. La tabla C.01 tiene los datos del Estado de Evolución. El contenido de ambas tablas fue utilizado repitiendo su contenido entre 1 y 5 veces para generar archivos con mayor cantidad de muestras. Variando entre los 500 hasta las 2500 entradas.

Table C.01: Datos del avance de la enfermedad utilizados para la verificación del modelo.

Inicio de la tabla				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	3	1	2	1
2	3	1	1	1
2	3	1	1	1
3	3	1	1	2
2	3	1	1	2
3	3	1	1	1
2	3	1	1	1
2	3	1	1	2
2	3	1	1	1
2	3	1	1	1
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
3	3	1	1	2
3	3	1	1	2
2	3	1	1	2
3	2	1	1	1

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	2	1	1	1
2	3	2	1	2
3	3	1	1	1
3	3	1	1	1
3	3	1	1	1
3	3	1	1	2
3	3	1	1	2
3	3	2	1	2
3	3	1	1	2
2	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
2	3	1	1	2
3	3	1	1	3
3	3	1	1	2
3	3	2	1	2
3	3	2	1	2
3	3	1	1	2
3	3	2	1	2
3	3	1	1	3
3	3	1	1	3
3	3	1	1	3

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	3	1	1	3
3	3	1	1	2
3	3	1	1	2
3	3	1	1	3
3	3	1	1	2
3	3	1	1	3
3	3	1	1	3
3	3	1	1	2
3	3	1	1	3
2	3	1	1	3
2	3	1	1	3
3	3	1	1	2
2	3	1	1	3
2	3	2	1	3
2	3	1	1	3
1	3	2	1	3
2	3	1	2	3
2	3	1	2	3
1	3	1	2	2
2	3	1	2	2
2	3	1	2	3
2	2	1	3	3
2	3	1	3	3
2	3	1	3	2

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
2	3	1	3	2
2	3	1	3	3
2	3	2	3	3
2	3	1	3	3
3	2	1	3	3
2	3	1	3	3
3	3	1	3	3
2	3	1	3	2
2	3	1	2	2
3	3	1	3	2
2	3	3	2	2
3	3	3	2	2
2	3	2	1	2
2	3	1	2	2
3	3	1	1	3
3	3	1	2	3
3	3	1	2	2
3	3	1	2	2
3	3	1	2	2
3	3	1	2	2
2	3	1	2	2
3	3	1	2	2
3	3	1	2	3
3	3	1	1	3

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	3	1	2	3
3	3	1	2	3
2	3	1	1	3
3	3	1	2	3
3	3	1	2	3
3	3	1	2	3
3	3	1	2	3
3	3	1	2	3
3	3	1	2	3
3	3	1	1	3
3	3	1	2	2
3	3	1	1	2
3	3	1	2	2
1	3	1	3	2
2	3	1	3	2
3	3	1	1	2
2	2	1	3	2
2	3	1	2	2
2	3	2	1	2
1	3	1	2	2
1	3	3	1	1
2	3	3	1	2
1	3	2	1	2
1	3	1	1	3

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
1	3	1	1	1
2	3	1	1	2
1	3	1	1	2
2	3	1	1	3
2	3	1	1	3
2	3	1	1	3
1	3	1	1	3
1	3	1	1	3
2	3	1	1	2
2	3	1	1	3
2	3	1	1	3
2	3	1	1	3
2	3	1	1	2
2	3	1	1	2
2	3	1	1	3
3	3	1	1	3
3	3	1	1	3
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	3
3	3	1	1	3
3	3	1	1	3

Continuación de la tabla C.01

Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	3	1	1	3
3	3	1	1	3
3	3	1	1	3
3	3	1	1	3
2	3	1	1	3
2	3	1	1	3
3	3	1	1	2
3	3	1	1	2
2	3	1	1	2
2	3	1	1	2
3	3	1	1	2
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	3
3	3	1	1	3
2	3	1	1	2
2	3	1	1	3
2	3	2	1	2
1	2	1	1	2
2	3	1	1	2

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
2	3	1	1	2
3	3	1	1	2
2	3	1	1	2
1	3	1	1	2
1	3	1	1	2
2	3	1	1	2
1	3	1	1	2
1	3	1	1	2
2	3	1	1	2
2	3	1	1	2
3	2	1	1	2
1	3	1	1	1
3	3	1	1	1
2	3	1	1	2
2	3	1	1	1
2	3	1	1	1
3	3	1	1	2
3	3	1	1	3
3	3	1	1	2
3	2	1	1	2
3	2	1	1	2
2	3	1	1	2
2	3	1	1	3
3	3	1	1	3

Continuación de la tabla C.01

Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
2	3	1	1	3
3	3	1	1	3
3	3	1	1	3
3	3	1	1	3
3	3	1	1	2
3	3	2	1	2
3	3	2	1	3
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
2	3	1	1	2
3	3	1	1	2
2	3	1	1	2
2	3	1	1	2
3	3	1	1	2
2	3	1	1	2
3	3	1	1	2
2	3	1	1	2
3	3	1	1	2
2	3	1	1	2
2	3	1	1	2

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
1	3	2	1	2
1	3	2	1	2
2	3	1	1	1
2	3	1	1	2
1	3	1	1	2
1	3	1	1	2
1	3	1	1	2
1	3	1	1	2
1	3	1	1	2
1	3	1	1	2
1	2	1	1	2
1	3	1	1	2
1	2	1	1	3
1	2	1	1	2
1	3	1	1	2
1	3	1	1	2
1	2	1	1	2
2	2	1	1	1
1	2	1	1	2
1	3	1	1	2
1	3	1	1	1
2	2	1	1	1
2	3	1	1	1
1	3	1	1	1

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
2	3	1	1	1
2	3	1	1	1
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
3	3	1	1	2
3	3	2	1	2
3	3	1	1	2
2	3	1	1	2
3	3	1	1	2
2	3	1	1	1
2	3	1	1	1
2	3	2	1	1
2	3	1	1	1
2	3	1	1	1
3	3	1	1	1
2	3	1	1	1
2	3	1	1	1
2	3	1	1	1
3	3	1	1	1
2	3	1	1	2
3	3	1	1	2
2	3	1	1	2

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
1	3	1	1	2
1	2	2	2	2
1	3	3	1	2
1	3	2	1	1
1	3	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
2	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	2	1	1	1
1	3	3	1	1
1	3	1	1	1
1	3	1	1	2
1	3	1	1	1
1	3	2	1	1
1	3	1	1	1
2	3	1	1	1

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
2	3	1	1	1
3	3	1	1	1
2	2	1	1	1
3	3	1	1	1
2	3	1	1	1
3	3	1	1	1
3	3	1	1	1
2	3	2	1	1
3	3	1	1	1
3	3	1	1	2
3	3	1	1	3
3	3	1	1	2
3	3	1	1	3
3	3	1	1	3
3	3	1	1	3
2	3	1	1	2
2	3	1	1	1
3	3	1	1	2
3	3	1	1	3
3	3	2	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	3
3	3	1	1	2

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
2	3	2	1	2
3	3	1	1	2
3	3	2	1	2
2	3	1	1	2
2	3	2	1	2
2	3	1	1	2
3	3	1	1	1
2	3	1	1	2
2	3	1	1	1
2	3	1	1	2
2	3	1	1	2
1	3	1	1	2
1	3	3	1	2
3	3	1	1	2
2	3	1	1	2
3	3	1	1	2
2	3	1	1	2
2	3	2	1	2

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	3	1	1	2
1	3	2	1	1
2	3	1	1	1
3	3	1	1	1
3	3	2	1	1
3	3	1	1	2
3	3	1	1	1
2	3	1	1	1
3	3	1	1	2
3	3	1	1	2
3	3	1	1	1
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
2	3	1	1	3
3	3	2	1	2
3	3	1	1	3
3	3	1	1	2
2	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	3	1	1	2
3	3	3	1	2
3	3	2	1	3
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	2	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
2	3	2	1	2
1	3	1	1	1
2	3	1	1	2
2	3	1	1	1
1	3	3	1	1
1	3	1	1	1
1	3	1	1	1
1	3	1	1	1
1	3	1	1	1
1	3	3	1	1
2	3	1	1	1

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
2	3	1	1	1
1	3	1	1	1
2	3	1	1	1
2	3	1	1	1
2	3	1	1	1
2	3	1	1	1
2	3	1	1	1
2	3	1	1	1
2	3	1	1	1
2	3	1	1	1
3	3	1	1	2
2	3	1	1	1
2	3	1	1	1
3	3	1	1	1
3	3	1	1	2
2	3	1	1	1
3	3	1	1	1
3	3	1	1	1
3	3	1	1	1
3	3	1	1	2
3	3	3	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
2	3	1	1	2
3	3	1	1	3
3	3	1	1	2
3	3	1	1	2
3	3	1	1	3
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	2	1	2
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
2	3	1	1	2
1	3	3	1	2
1	3	1	1	2
1	3	1	1	1
1	3	3	1	1
2	3	1	1	1

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
2	3	1	1	1
1	3	1	1	1
2	3	1	1	2
1	3	1	1	1
2	3	1	1	1
1	3	1	1	1
2	3	1	1	1
2	3	1	1	2
2	3	1	1	1
2	3	1	1	1
2	3	1	1	1
2	3	1	1	1
2	3	1	1	1
2	3	1	1	1
2	3	1	1	2
3	3	1	1	2
2	3	1	1	2
3	3	1	1	2
2	3	1	1	2
3	3	1	1	2
3	3	1	1	1
3	3	1	1	2
3	3	2	1	1
3	3	1	1	2
3	3	1	1	2

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
2	3	1	1	2
3	3	1	1	2
3	3	1	1	2
2	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	2	1	2
2	3	1	1	2
3	3	1	1	1
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
2	3	1	1	2
3	3	1	1	2
2	3	1	1	2
3	3	1	1	2
2	3	1	1	2
2	3	2	1	2
2	3	1	1	2
1	3	3	1	2

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
1	3	1	1	2
2	3	1	1	1
3	3	1	2	2
2	3	1	2	2
2	3	2	2	2
2	3	1	2	1
3	3	1	3	2
2	3	1	2	1
3	3	1	2	1
2	3	1	2	1
2	3	1	2	1
2	3	1	2	1
2	3	1	2	1
2	3	1	1	1
1	3	2	2	1
2	3	1	2	1
3	3	1	2	2
3	3	1	2	2
3	3	1	1	2
3	3	1	2	2
3	3	1	2	2
3	3	1	2	2
3	3	1	2	2
3	3	1	2	2
3	3	1	2	2

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	3	1	2	2
3	3	1	1	2
3	3	1	1	2
3	3	1	2	2
3	3	1	2	2
3	3	1	1	2
3	3	1	1	2
3	3	1	2	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	2	1	3
3	3	1	2	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	2	2
3	3	1	1	2
3	3	1	2	2
3	3	1	1	2
3	3	1	1	2

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	2	1	2
3	3	1	1	2
3	3	1	1	2
2	3	1	2	2
3	3	1	2	2
3	3	1	2	2
3	3	1	2	2
2	3	1	2	2
2	3	1	2	2
2	3	1	2	1
2	3	1	2	1
3	3	1	2	2
2	3	1	3	2
3	3	1	2	1
3	3	1	2	1
3	3	1	2	1
3	3	1	2	1
3	3	1	3	1
3	3	1	2	1

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	3	1	2	2
3	3	1	2	1
3	3	1	1	2
3	3	1	2	2
3	3	1	2	2
3	3	1	2	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	3	2	2
3	3	2	1	2
3	3	1	1	2
3	3	3	1	2
3	3	3	1	2
3	3	3	2	2
3	3	2	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	1
3	3	1	1	1
3	3	1	2	2
3	3	1	1	1
3	3	1	1	1

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	3	1	1	1
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
2	3	2	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	2	2
3	3	2	1	2
1	3	3	2	1
2	3	1	2	1
3	3	1	2	1
3	3	1	2	1
2	3	2	3	2
3	3	1	2	1
2	3	1	2	1
1	3	2	1	1
2	3	1	2	2
2	3	1	2	2
2	3	2	1	1
3	3	1	2	2
2	3	1	2	2
3	3	1	3	1

Continuación de la tabla C.01				
Temperatura	Humedad	Precipitación	Velocidad Viento	Estado Evolución
3	3	1	2	2
3	3	1	2	2
3	3	1	2	2
3	3	1	2	2
3	3	1	1	2
3	3	1	2	1
3	3	1	1	2
3	3	1	1	2
3	3	3	1	2
3	3	2	1	2
3	3	3	1	2
3	3	1	1	2
3	3	1	1	3
3	3	2	1	2
3	3	3	2	2
3	3	1	2	2
3	3	3	1	2
3	3	1	1	2
3	3	1	1	2
3	3	1	1	2
3	3	3	2	2
3	3	1	1	2
3	3	1	1	2
3	3	1	2	2

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
1	2	2	1	2
2	3	2	1	2
2	3	3	2	3
1	2	1	2	2
3	2	3	2	3
2	3	3	2	3
1	3	2	1	2
2	2	2	2	2
3	3	2	1	3
3	3	2	2	3
2	2	2	3	3
2	2	2	1	2
2	2	2	3	3
3	2	2	2	3
3	2	2	3	3
1	3	3	1	2
2	2	3	2	3
1	3	3	1	2
2	2	3	2	3
1	3	2	3	3
1	2	2	3	2
2	1	1	3	2
1	2	2	2	2
2	3	2	3	3

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
3	3	3	3	3
3	3	1	2	3
3	1	2	3	3
1	3	1	3	2
2	1	3	2	2
1	1	3	1	2
2	1	3	3	3
2	3	3	1	3
1	2	2	3	2
3	2	2	1	2
2	1	3	2	2
2	1	3	3	3
2	3	3	3	3
1	3	1	3	2
1	3	2	2	2
2	3	1	2	2
1	1	2	3	2
1	1	2	3	2
1	1	2	3	2
2	2	2	3	3
1	1	3	2	2
2	3	3	2	3
3	1	1	3	2
1	3	2	1	2

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
2	1	1	2	2
2	2	3	1	2
2	3	3	1	3
1	3	3	1	2
1	3	3	3	3
1	2	2	3	2
3	1	2	1	2
3	1	3	2	3
1	3	1	1	2
1	1	3	1	2
1	1	2	2	2
2	2	2	2	2
3	2	3	1	3
2	3	2	3	3
1	2	2	1	2
1	1	2	2	2
1	3	1	2	2
1	1	1	3	2
3	1	3	1	2
1	1	1	2	2
2	1	2	2	2
2	2	1	2	2
3	2	3	1	3
3	2	3	2	3

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
1	2	1	2	2
2	3	2	2	3
2	3	3	3	3
3	2	1	1	2
1	1	3	3	2
1	2	2	1	2
1	2	2	2	2
2	2	2	1	2
2	2	1	3	2
2	2	1	3	2
3	3	2	1	3
2	2	1	2	2
1	1	1	2	2
3	1	3	1	2
3	2	3	1	3
2	2	1	1	2
1	2	2	2	2
2	1	3	2	2
1	3	2	1	2
1	1	2	2	2
3	1	3	3	3
3	2	1	1	2
2	2	3	3	3
2	3	2	1	2

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
1	2	1	1	2
1	2	1	3	2
2	2	1	3	2
2	1	1	2	2
3	1	3	2	3
2	2	3	1	2
3	1	1	2	2
1	2	2	3	2
3	3	2	3	3
1	1	2	3	2
2	3	2	3	3
2	1	2	2	2
1	2	2	3	2
1	1	2	2	2
2	1	1	3	2
1	2	2	3	2
3	2	3	1	3
1	2	3	2	2
3	3	1	1	2
1	1	1	2	2
3	2	1	3	3
1	1	1	1	1
1	1	3	1	2
2	3	3	1	3

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
3	3	3	3	3
3	3	1	2	3
1	1	3	1	2
2	2	3	3	3
3	2	2	2	3
2	2	1	2	2
1	2	3	2	2
2	2	2	1	2
1	3	3	3	3
2	1	3	3	3
1	1	1	2	2
2	2	1	1	2
1	3	3	1	2
1	3	2	1	2
1	1	3	1	2
2	2	1	1	2
1	2	2	1	2
3	3	2	1	3
3	3	1	3	3
3	2	2	3	3
1	2	1	3	2
2	3	1	2	2
1	2	3	2	2
2	1	1	1	2

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
1	2	3	2	2
1	2	3	3	3
2	2	3	3	3
1	2	1	3	2
2	1	3	1	2
3	2	3	1	3
2	1	1	2	2
1	2	1	2	2
2	3	3	2	3
1	1	3	2	2
2	1	2	3	2
3	3	1	2	3
2	2	1	3	2
2	3	2	1	2
2	2	2	1	2
3	3	3	2	3
1	3	2	3	3
1	2	1	3	2
1	2	3	1	2
2	1	1	1	2
3	1	2	1	2
2	3	2	1	2
1	2	1	1	2
2	2	2	3	3

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
3	2	2	3	3
3	3	2	2	3
3	3	2	3	3
3	1	1	2	2
3	3	2	3	3
1	1	2	1	2
3	2	3	3	3
1	2	2	1	2
2	1	1	1	2
2	2	2	2	2
1	1	1	1	1
1	1	1	1	1
3	1	3	2	3
2	1	1	3	2
3	1	2	1	2
2	2	2	2	2
2	3	2	1	2
3	3	2	3	3
3	1	1	1	2
1	2	2	2	2
2	1	2	3	2
1	1	3	2	2
2	3	2	1	2
3	2	1	3	3

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
1	2	2	2	2
3	3	1	2	3
3	2	3	3	3
2	2	1	3	2
1	1	1	3	2
3	1	3	3	3
2	3	2	3	3
3	3	1	3	3
3	1	1	2	2
1	2	1	3	2
2	2	3	2	3
3	2	3	3	3
1	2	1	3	2
1	3	1	3	2
2	1	3	3	3
1	1	3	1	2
1	1	1	3	2
2	2	1	3	2
2	1	1	3	2
2	3	2	3	3
1	1	3	2	2
3	1	3	3	3
3	1	1	2	2
3	1	1	1	2

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
1	2	3	1	2
2	1	1	3	2
2	1	1	2	2
3	2	2	3	3
3	2	3	1	3
2	2	3	1	2
2	3	1	1	2
3	2	1	2	2
3	1	1	2	2
3	2	1	3	3
1	1	1	2	2
2	2	2	3	3
3	1	2	1	2
1	3	3	2	3
1	2	2	2	2
1	3	3	1	2
2	3	1	2	2
2	1	2	1	2
3	3	2	1	3
2	3	1	3	3
3	3	3	1	3
2	3	3	3	3
1	1	1	1	1
2	3	1	1	2

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
2	3	1	2	2
3	3	1	2	3
1	2	3	1	2
1	2	2	2	2
1	2	3	1	2
3	2	2	3	3
3	1	1	1	2
2	2	2	3	3
1	2	1	3	2
3	2	1	2	2
1	1	1	2	2
1	3	3	2	3
3	3	2	3	3
2	3	1	3	3
3	2	1	2	2
1	1	3	1	2
2	1	3	2	2
2	3	3	2	3
3	1	2	2	2
1	3	3	2	3
3	1	1	2	2
1	3	2	1	2
1	2	2	3	2
3	2	2	2	3

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
3	1	2	2	2
3	1	3	3	3
3	1	1	1	2
3	2	3	1	3
3	2	1	1	2
3	3	3	2	3
2	1	3	2	2
1	3	1	3	2
3	3	3	3	3
2	2	3	1	2
1	1	2	1	2
1	1	3	2	2
2	1	1	3	2
1	3	3	1	2
2	3	3	3	3
2	2	1	2	2
1	3	3	1	2
1	2	3	2	2
3	2	1	3	3
3	2	3	3	3
1	3	2	3	3
1	1	1	2	2
3	3	1	3	3
3	2	2	2	3

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
2	3	2	1	2
2	3	3	3	3
2	3	1	3	3
2	2	2	2	2
2	3	3	3	3
3	2	1	3	3
2	3	2	3	3
1	3	3	1	2
1	2	1	2	2
3	2	1	2	2
2	3	1	2	2
1	3	3	3	3
2	3	1	3	3
2	1	2	2	2
1	3	2	3	3
3	2	3	3	3
2	2	2	1	2
1	3	3	2	3
2	3	2	3	3
3	3	1	1	2
2	3	1	3	3
3	1	2	1	2
3	1	3	1	2
2	1	2	3	2

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
1	1	1	2	2
3	1	2	3	3
1	3	1	2	2
1	2	2	1	2
1	3	3	3	3
1	2	1	1	2
3	3	2	1	3
1	2	2	1	2
3	2	3	1	3
3	1	3	3	3
1	2	1	1	2
3	1	3	1	2
1	2	3	3	3
1	1	1	2	2
3	1	1	1	2
3	1	2	2	2
1	3	2	2	2
2	1	1	1	2
1	2	2	3	2
3	2	1	1	2
3	3	1	1	2
2	3	3	2	3
3	3	1	3	3
1	2	2	2	2

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
2	3	1	1	2
2	2	2	2	2
2	2	3	2	3
3	1	2	2	2
3	1	3	2	3
3	1	1	3	2
3	3	1	2	3
2	3	2	1	2
1	3	3	2	3
2	3	1	1	2
3	2	3	2	3
3	1	1	3	2
2	3	1	2	2
2	3	3	3	3
1	1	2	1	2
2	2	2	2	2
2	1	3	1	2
2	3	1	1	2
3	3	3	3	3
2	1	2	3	2
2	3	1	2	2
3	2	2	1	2
2	2	1	2	2
2	2	3	3	3

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
3	3	2	3	3
3	1	2	2	2
3	2	2	2	3
1	1	1	2	2
2	3	1	3	3
3	3	1	1	2
2	2	1	3	2
3	2	3	2	3
2	1	2	3	2
1	1	1	1	1
2	2	2	1	2
3	1	3	2	3
3	2	1	1	2
1	3	1	2	2
1	3	2	1	2
3	3	1	1	2
1	3	2	2	2
1	2	3	2	2
3	3	2	3	3
1	2	2	2	2
3	3	1	3	3
3	1	1	3	2
2	1	1	2	2
3	1	2	1	2

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
1	3	3	2	3
2	2	3	2	3
1	1	3	2	2
2	3	1	2	2
2	2	1	2	2
2	2	1	3	2
3	3	3	1	3
2	3	1	3	3
3	3	3	3	3
1	2	2	3	2
1	1	2	3	2
2	3	1	1	2
2	3	1	3	3
1	1	3	3	2
3	3	2	1	3
2	3	3	3	3
3	2	2	3	3
1	1	2	2	2
1	1	2	2	2
1	2	2	2	2
3	1	3	2	3
2	3	2	2	3
2	1	2	3	2
1	1	1	2	2

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
1	1	2	1	2
2	2	3	3	3
1	2	3	2	2
2	2	3	3	3
3	2	2	3	3
3	2	1	2	2
1	3	2	3	3
1	2	1	3	2
3	1	3	3	3
2	2	1	3	2
2	3	2	1	2
3	2	3	3	3
1	1	1	3	2
2	3	3	3	3
1	2	2	3	2
3	3	3	3	3
1	2	3	3	3
3	1	2	3	3
2	3	2	1	2
1	2	3	3	3
2	2	2	2	2
1	3	1	1	2
1	2	1	1	2
3	3	2	2	3

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
1	2	1	1	2
3	3	2	3	3
1	2	3	3	3
2	2	3	3	3
1	3	2	2	2
3	1	1	3	2
1	1	3	3	2
2	3	3	1	3
3	1	3	3	3
2	2	2	1	2
1	3	1	2	2
2	2	3	1	2
3	3	2	2	3
1	3	2	1	2
3	3	1	1	2
3	3	2	2	3
2	2	3	1	2
2	2	3	3	3
2	1	3	2	2
2	3	2	2	3
1	1	2	3	2
1	2	3	1	2
3	3	3	1	3
1	2	1	3	2

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
2	2	1	1	2
2	2	2	1	2
2	1	2	1	2
3	3	2	1	3
2	1	2	1	2
1	3	3	3	3
2	1	1	2	2
1	3	3	2	3
3	2	3	2	3
1	1	1	1	1
2	3	2	1	2
3	1	2	3	3
1	2	1	3	2
3	2	3	1	3
1	1	2	3	2
1	1	2	3	2
1	2	1	2	2
2	2	1	3	2
3	3	3	3	3
3	2	2	3	3
1	3	2	3	3
1	3	3	1	2
1	1	2	3	2
1	3	2	2	2

Continuación de la tabla C.02				
x_0	x_1	x_2	x_3	y
2	1	3	1	2
1	1	1	3	2
2	1	3	3	3
2	1	1	2	2
1	1	2	2	2
3	2	2	3	3
1	2	3	2	2
1	3	2	1	2
2	2	3	1	2
1	3	3	1	2
2	3	3	1	3
1	2	1	1	2
Fin de la tabla				

Bibliografía

- [1] Arinze Akutekwe, Huseyin Seker, and Sunday Iliya. An optimized hybrid dynamic bayesian network approach using differential evolution algorithm for the diagnosis of hepatocellular carcinoma. In *Adaptive Science & Technology (ICAST), 2014 IEEE 6th International Conference on*, pages 1–6. IEEE, 2014.
- [2] Vincent Van Asch. Macro- and micro-averaged evaluation measures [[basic draft]], 2013. URL <http://www.cnts.ua.ac.be/~vincent/pdf/microaverage.pdf>.
- [3] Norhaini Baba, Mohd Saberi Mohamad, Abdul Hakim Mohamed Salleh, Mohd Hanafi Ahmad Hijazi, Lian En Chai, Muhammad Mahfuz Zainuddin, and Safaai Deris. Continuous dynamic bayesian network for gene regulatory network modelling. In *Computational Science and Technology (ICCST), 2014 International Conference on*, pages 1–5. IEEE, 2014.
- [4] Charles Cabot, James Ulrich, and Mark Raugas. LibPGM: Probabilistic graphical models on Python, 2012–. URL <http://pythonhosted.org/libpgm/#documentation/>. [Online; accedido Febrero 2017].
- [5] Luis Calvo. *Aprendizaje de Máquina aplicado al pronóstico en cultivos agrícolas*. PhD thesis, Doctorado en Ciencias Naturales para el Desarrollo, 2017.
- [6] CORBANA. Corporación nacional bananera., 2017. URL "<https://www.corbana.co.cr/categories/quienes-somos>". [Online; accedido Mayo 2017].
- [7] Norsys Corporation. Netica application, 1995–. URL "<https://www.norsys.com/netica.html>". [Online; accedido Abril 2017].
- [8] Ministerio de Educación Pública de Costa Rica Unesco. Conceptos y herramientas sobre sistemas de alerta temprana y gestión del riesgo para la comunidad educativa., 2012. URL "<http://www.cridlac.org/digitalizacion/pdf/spa/doc19078/doc19078-contenido.pdf>". [Online; accedido Abril 2017].

- [9] Tarek El-Gaaly, Vicky Froyen, Ahmed Elgammal, Jacob Feldman, and Manish Singh. A bayesian approach to perceptual 3d object-part decomposition using skeleton-based representations. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15*, pages 3762–3768. AAAI Press, 2015. ISBN 0-262-51129-0. URL <http://dl.acm.org/citation.cfm?id=2888116.2888238>.
- [10] The R Foundation for Statistical Computing. R version 3.3.2, 2016. URL "<https://www.r-project.org/>". [Online; accedido Mayo 2017].
- [11] Python Software Foundation. Python language reference, version 2.7, 2009–. URL "<http://www.python.org>". [Online; accedido Mayo 2017].
- [12] The MathWorks Inc. Matlab r2016a, 1984–. URL "<https://www.mathworks.com/products/matlab.html>". [Online; accedido Mayo 2017].
- [13] Masoumeh Izadi, Charland Katia, and Buckeridge David. Using dynamic bayesian networks for incorporating non-traditional data sources in public health surveillance. *AAAI-14 Workshop*, 2014.
- [14] Kiran Karkera. *Building Probabilistic Graphical Models with Python*. Packt Publishing, 2014. ISBN 1783289007 9781783289004.
- [15] Daphne Koller and Friedman Nir. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.
- [16] Douglas Marín and Romero Ronald. El combate de la sigatoka negra. *Boletín número 4 departamento de investigaciones CORBANA*, 1992.
- [17] Mehryar Mohri and Rostamizadeh Afshin. *Foundations of Machine Learning*. The MIT Press, 2012.
- [18] Kevin Murphy. The bayes net toolbox for matlab. *Computing Science and Statistics*, 2001. URL <http://people.cs.ubc.ca/~murphyk/Papers/bnt.pdf>. [Online; accedido Agosto 2016].
- [19] Kevin Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.
- [20] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. Uci repository of machine learning databases, 1998. URL <http://www.ics.uci.edu/~mlearn/MLRepository.html>. [Online; accedido Enero 2017].
- [21] Dusheyko S Hua S Poliakov A Shabalov I Smirnova T Grigoriev IV Dubchak I Nordberg H, Cantor M. The genome portal of the department of energy joint genome institute: 2014 updates, 2014. URL <http://genomeportal.jgi-psf.org/Mycfi2/Mycfi2.home.html>. [Online; accedido Agosto 2014].

- [22] Martin Paluszewski and Thomas Hamelryck. Mocapy++ - a toolkit for inference and learning in dynamic bayesian networks. *BMC Bioinformatics*, 2010. URL <http://dx.doi.org/10.1186/1471-2105-11-126>. [Online; accedido Noviembre 2016].
- [23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL <http://scikit-learn.org/stable/modules/multiclass.html>. [Online; accedido Marzo 2017].
- [24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL <https://www.R-project.org/>. [Online; accedido Mayo 2017].
- [25] Pierre Raybaut. Spyder: Scientific python development environment, 2009–. URL "<https://github.com/spyder-ide/spyder>". [Online; accedido Mayo 2017].
- [26] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., 2015. URL <http://www.rstudio.com/>. [Online; accedido Mayo 2017].
- [27] S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Series in Artificial Intelligence. Prentice Hall, 2010. ISBN 9780136042594.
- [28] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, jul 2009.
- [29] Yungang Zhua, Liu Dayou, Chen Guifen, Jia Haiyang, and Yu Helong. Mathematical modeling for active and dynamic diagnosis of crop diseases based on bayesian networks and incremental learning. *Mathematical and Computer Modeling*, pages 514–523, 2011.