UNA-ITCR-UNED

Doctorado en Ciencias Naturales para el Desarrollo
Énfasis en Tecnologías Electrónicas Aplicadas

Estrategia de predicción en procesos biológicos del campo agrícola con datos limitados: casos de aplicación en café y banano.

Documento de tesis sometido a consideración para optar por el grado académico de Doctor en Ciencias Naturales para el Desarrollo con Énfasis en Tecnologías Electrónicas Aplicadas

Luis-Alexánder Calvo-Valverde

Heredia, 4 de mayo del 2020

Declaro que el presente documento de tesis ha sido realizado enteramente por mi persona, utilizando y aplicando literatura referente al tema e introduciendo conocimientos y resultados experimentales propios.

En los casos en que he utilizado bibliografía he procedido a indicar las fuentes mediante las respectivas citas bibliográficas. En consecuencia, asumo la responsabilidad total por el trabajo de tesis realizado y por el contenido del presente documento.

Luis-Alexánder Calvo-Valverde

Heredia, 4 de mayo del 2020

Céd: 1-0749-0252

MIEMBROS DEL TRIBUNAL EVALUADOR

Dr. Álvaro Martín Parada Gómez
Representante del Consejo Central de Posgrado

Dr. Víctor Hugo Granados Fernández
Representante de la Coordinación General del DOCINADE

Dr.-Ing. José Pablo Alvarado Moya
Tutor de tesis

Dr. Carlos González Alvarado
Miembro del Comité Asesor

Dr. José Castro Mora
Miembro del Comité Asesor

Ing. Luis Alexánder Calvo Valverde
Sustentante

Resumen

En la época actual se vive una creciente demanda por contar con alimentos en mayores cantidades y a menor costo para la población. Pero a la vez, las áreas destinadas a la producción de alimentos agrícolas, en lugar de aumentar, han tendido a decrecer; esto fruto de la urbanización, los requerimientos de la industria y la extracción de recursos naturales. En este contexto, los organismos internacionales han invitado a proveer apoyo tecnológico para responder a esta problemática.

Para aportar en la solución y en el marco del Doctorado en Ciencias Naturales para el Desarrollo (DOCINADE), es que se desarrolla la presente tesis doctoral. La idea central es aplicar técnicas del aprendizaje automático al mundo agrícola con el fin de ayudar a los agricultores en la toma de decisiones, aportándoles predicciones basadas en datos históricos de sus procesos biológicos y de variables climatológicas. Concretamente, este trabajo propone una estrategia para la aplicación del aprendizaje automático en la predicción de procesos biológicos en el campo agrícola, mostrando casos de aplicación en los cultivos del banano y del café.

Acorde con los fines del DOCINADE, si bien la estrategia tiene un uso abierto para la comunidad mundial, su construcción estuvo orientada a pequeños y medianos productores, quienes normalmente no cuentan con conjuntos de datos provenientes de sensores de alta calidad y costo, y más bien se trata de apoyar el trabajo colaborativo entre los productores.

De la estrategia se resaltan los siguientes aportes: propone una estrategia esquemática que favorece la repetibilidad del proceso, no requiere predecir variables meteorológicas, propone un método de aumento de datos, no requiere contar con imágenes para iniciar con la experimentación, propone una manera de trabajar con el espacio paramétrico de manera heurística, permite una optimización multiobjetivo, aprovecha el aprendizaje por transferencia y contribuye en la selección de atributos.

Palabras clave: aprendizaje automático, procesos biológicos en el campo agrícola, estrategia de solución, predicción, banano, café.

Abstract

Nowadays humanity is experiencing an increasing demand for food for a growing population at low cost. At the same time the agricultural areas are decreasing. That is the result of urbanization, industry requirements and the extraction of natural resources. In this situation, international organizations have invited to provide technological support to respond to this problem.

To overcome this worldwide problem and within the framework of the Doctoral Program in Natural Sciences for Development (DOCINADE), the present work contributes to the solution of this problem.

The central idea is to apply machine learning techniques to the agricultural world in order to support farmers in their decision making, providing them with predictions based on historical data of their biological processes and climatic variables. Specifically, this work proposes a strategy for the application of machine learning in the prediction of biological processes in the agricultural field, taking banana and coffee crops as use cases.

According to the DOCINADE goals, although the strategy has an open use for the world community, its construction was aimed at small and medium producers, who usually do not have data sets from sensors of high quality and cost. It rather supports collaborative work among producers.

The following contributions are highlighted from the strategy: a schematic strategy that favors the repeatability of the process, it does not require to predict meteorological variables nor requires images to start the experimentation. It proposes a method for data augmentation and a way of working with the parametric space in a heuristic way. It also integrates multi-objective optimization, takes advantage of transfer learning, and contributes to the selection of attributes.

Keywords: machine learning, agricultural biological process, strategic of solution, forecasting, banana, coffee.

a mi esposa Anayansie Fallas Badilla, a mis hijos: Mary Ángel y Anthony, a mi madre Elena Valverde (q.d.D.g), a mi padre Sigifredo Calvo, a mis hermanos y hermanas.

Agradecimientos

Si bien esta tesis tiene solo un autor, es innegable que poder llegar a terminarla, sólo es posible gracias a la colaboración de muchas personas e Instituciones a lo largo del proceso.

Un agradecimiento especial al Dr. Pablo Alvarado Moya, sin cuyo seguimiento y guía a lo largo de todos estos años, no habría sido posible llegar a este punto.

Gracias al Dr. Carlos González Alvarado y al Dr. José Castro Mora, quienes como profesores asesores estuvieron siempre dispuestos a colaborar y a revisar lo que correspondiera.

Gracias a mi familia, por su comprensión y apoyo durante tantos años de estudio, en los que muchas veces fueron los sacrificados al no poder dedicarles todo el tiempo que se merecían.

Gracias al Instituto Tecnológico de Costa Rica por apoyarme financieramente en mis estudios y en particular a la Escuela de Ingeniería en Computación, que creyó en mi y desde el inicio aprobó mi participación en este proceso de formación.

Gracias al DOCINADE, por ofrecer una opción de calidad para realizar mis estudios doctorales.

Finalmente, pero no menos importante, gracias al personal de los Centros de Investigación de la Corporación Bananera Nacional y del Instituto del Café de Costa Rica, quienes me dieron el honor de conocerles, trabajar y aprender en equipo multidisciplinario.

Luis-Alexánder Calvo-Valverde

Heredia, 4 de mayo del 2020

Simbología, Glosario y Estructuras

Simbología

CA	Conocimiento para el aprendizaje, conjunto
ca	Conocimiento para el aprendizaje, objeto
cvm	Coeficiente de variación multidimensional
D	Concatenación vertical entre X e y
EPB	Experimento de proceso biológico, conjunto
epb	Experimento de proceso biológico, objeto
ha	Hectáreas
I	Matriz identidad
q	Resultado de multiplicar el cvm por 100 y tomar la parte entera
R^2	Coeficiente de determinación
RCA	Repositorio de conocimiento aprendido
$RMSE$	Raíz cuadrada del error cuadrado medio
VCA	Variable de conocimiento aprendido, conjunto
vca	Variable de conocimiento aprendido, objeto
vca_dat	Datos de un vca
divergencia-KL	Divergencia de Kullback-Leibler
t-SNE	t-Distributed stochastic neighbor embedding

Glosario

- **Divergencia de Kullback-Leibler:** Es un indicador de la similitud entre dos funciones de distribución.
- **Fanega de café:** En Costa Rica, una fanega equivale a 20 cajuelas y una cajuela es aproximadamente 20 litros.
- **Grados de libertad:** Se refiere a la cantidad de información suministrada por los datos que se pueden utilizar para estimar los valores de parámetros de población desconocidos y calcular la variabilidad de esas estimaciones.
- **Multiplicadores de Lagrange:** Es un método para encontrar extremos de una función de varias variables restringidas a un subconjunto dado.

- **Perplejidad:** Es una medida de lo bien que una distribución de probabilidad o modelo de probabilidad predice una muestra.
- **Probabilidad condicional:** Probabilidad existente de que suceda un evento A, conociendo que además ocurre otro evento B.
- **Probabilidad conjunta:** Probabilidad de que los eventos A y B sucedan al mismo tiempo.
- **Pruebas ANOVA:** En estadística, una prueba de análisis de varianza es una forma de averiguar si los resultados de un experimento son significativos. En otras palabras, le ayudan a determinar si necesita rechazar la hipótesis nula o aceptar la hipótesis alternativa. One-way / Two-way, se refiere al número de variables independientes en la prueba de análisis de varianza.

Estructuras EPB: Experimento de Proceso Biológico, corresponderá al conjunto de objetos que representan un experimento. Cada objeto en el EPB, llamado *epb*, tendrá la siguiente estructura:

- *id_epb*: Identificador único, será una hilera de caracteres.
- *descripcion*: Texto explicativo del objetivo del experimento.
- *variables*: Vector que contendrá los *id_vca* de los *vca* que conforman el experimento y donde el último *id_vca* en el vector corresponderá a la variable a predecir.
- *ca_estudio*: *id_ca* del *ca* que se estudia en el experimento.
- C: Vector con todos los *id_ca* de los *ca* utilizados como entrenamiento en el experimento.
- Pat: Vector que contendrá todos los patrones a experimentar.
- T: Vector que contendrá las técnicas a aplicar.
- T': Vector que contendrá las técnicas en T que tienen uno o más parámetros y que por tanto requieren que se determine su espacio paramétrico.
- E: Matriz $\in \mathbb{R}^{n \times m}$, para $i \in \{1,2,3,...,n\}$ y $j \in \{1,2,3,...,m\}$, donde $e_{i,j}$ corresponderá al espacio paramétrico de la i-ésima técnica en T', para su j-ésimo parámetro.
- O: Matriz $\in \mathbb{R}^{n \times m}$, para $i \in \{1,2,3,...,n\}$ y $j \in \{1,2,3,...,m\}$, donde $o_{i,j}$ contendrá los valores seleccionados por el proceso de optimización heurística para la i-ésima técnica en T', en su j-ésimo parámetro.
- M: Vector que contendrá las métricas a calcular.
- MEV: Método de entrenamiento y validación. Será una hilera de caracteres.
- S: Objeto que contendrá todas las matrices con patrones generadas.
- R: Matriz $\in \mathbb{R}^{n \times m \times p}$, para $i \in \{1,2,3,...,n\}$, $j \in \{1,2,3,...,m\}$ y $k \in \{1,2,3,...,p\}$, que contendrá los resultados obtenidos en el proceso de entrenamiento y validación, y donde $r_{i,j,k}$ corresponderá al resultado de la i-ésima matriz de patrones en S, para la j-ésima técnica en T y para la k-ésima métrica en M.

- AE: Objeto que contendrá los resultados de realizar el análisis de varianza a los resultados en R. Podrán ser documentos de texto, gráficos, tablas de datos, entre otros.
- U: Objeto que contendrá los resultados en R que pertenecen al frente de Pareto determinado en el experimento.
- *observaciones*: Objeto que contendrá las observaciones que el equipo investigador considera deben ser guardadas respecto al experimento. Podrán ser documentos de texto, gráficos, videos, audios, tablas de datos, entre otros.

CA: Conocimiento para el Aprendizaje, corresponderá al conjunto de objetos que representan datos para el aprendizaje. Cada objeto en el CA, llamado *ca*, tendrá la siguiente estructura:

- *id_ca*: Identificador único, será una hilera de caracteres.
- *id_epb*: Identificador del *epb* que le dio origen.
- *tipo_aumento_datos*: Texto que indicará el tipo de aumento de datos utilizado. Si el *ca* no proviene de aumento de datos, este atributo contendrá una hilera nula.
- A: Atributos que se incluirán en el *ca*. Se representa como un vector que contendrá los *id_vca* del *vca* que lo conforman. a_i será el i-ésimo atributo en A, para $i \in \{1,2,3,...m\}$. Los primeros $m-1$ elementos corresponden a los atributos independientes y a_m corresponde al atributo a predecir
- N: Serán los atributos para los que se desean probar sus combinaciones. N se representará como un vector conformado por un subconjunto propio de los primeros $m-1$ elementos en A (Subconjunto propio de los atributos independientes).
- X: Matriz $\in \mathbb{R}^{n \times m-1}$, para $i \in \{1,2,3,...,n\}$ y $j \in \{1,2,3,...,m-1\}$, donde $x_{i,j}$ corresponderá al i-ésimo vector en su j-ésimo atributo.
- y: Vector columna de datos numéricos con n elementos, en donde y_i contendrá el valor del atributo a_m para la i-ésima fila en X. Atributo independiente.
- *detalles*: Para cada uno de los *id_vca* en el vector A, se tendrá un *ca_dat*, que será un objeto con la siguiente estructura.
  ca_dat:
  - *ca_id_vca*: Identificador único.
  - *periodicidad*: Intervalo temporal entre dato y dato.
  - *marca_temporal_inicio*: Indicación cronológica del dato más antiguo utilizado.
  - *marca_temporal_fin*: Indicación cronológica del dato más reciente utilizado.
  - *maximo*: Valor máximo permitido. Determinado por el experto en el dominio, el cual sirve para detectar valores atípicos.
  - *minimo*: Valor mínimo permitido. Determinado por el experto en el dominio, el cual sirve para detectar valores atípicos.

VCA: Variable de Conocimiento Aprendido, corresponderá al conjunto de objetos que representan variables que miden algún fenómeno, sea físico o biológico. Cada objeto en el VCA, llamado *vca*, tendrá la siguiente estructura:

- *id_vca*: Identificador único, será una hilera de caracteres.

- *descripcion*: Texto explicativo del tipo de fenómeno que refleja la variable.
- *mostrar_como*: Texto utilizado al mostrar el *id_vca*.
- *unidad*: La unidad en que se mide la variable.
- *origenes*: Cada *vca* podrá contener cero o más *vca_dat*. Cada *vca_dat* será un objeto con la siguiente estructura.
  vca_dat:
  - *id_vca_dat*: Identificador único, será una hilera de caracteres.
  - *id_vca*: Identificador del *vca* al que pertenece.
  - *origen*: Descripción de donde fueron tomados los datos.
  - *datos*: Matriz $\in \mathbb{R}^{n\times 2}$, donde para cada fila i, $i \in \{1,2,3,...,n\}$, la columna uno será el valor de la variable y la columna dos será la marca temporal cuando se tomó el valor de la variable.
- $filtro$: Filtro recomendado a aplicar en caso de requerirse para unificar frecuencias. Será una hilera y es determinado por el experto en el dominio. Se consideran las siguientes opciones, pero con criterio experto podrán definirse adicionales:
  - *suma*: Suma de todos los valores en el rango.
  - *promedio*: Media aritmética de los valores en el rango.
  - *mediana*: Es el valor que se ubica en la posición central al ordenar de menor a mayor los datos en el rango, de existir varios valores que cumplen esta característica, se toma el de mayor valor numérico.
  - *moda*: Es el valor que más se repite en el rango a filtrar, de existir varios valores que cumplen esta característica, se toma el de mayor valor numérico de entre ellos.
  - *maximo*: Es el mayor valor de todos los valores en el rango.
  - *minimo*: Es el menor valor de todos los valores en el rango.
- *metodo_imputacion*: En diálogo con los expertos del área se recomienda el método de imputación de faltantes.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Un tema que ha ocupado la atención de la humanidad desde la antigüedad es el de la producción de alimentos y ésta vista desde perspectivas tales como la calidad de la semilla, el proceso de producción, las enfermedades que afectan la productividad, el efecto del clima y el lugar.

El mundo de hoy vive el reto de producir más y a menor costo. Además, hay realidades en el siglo XXI, como el cambio climático, que imponen restricciones que invitan a ser creativos en las soluciones y a utilizar la capacidad intelectual para lograr este objetivo.

Ahora bien, no basta con pensar en aumentar la producción de alimentos, sino que ésta debe hacerse de un modo sostenible. Al respecto, el *Informe Brundtland* de las Naciones Unidas [109] dice que el Desarrollo Sostenible es el desarrollo que satisface las necesidades de la generación presente, sin comprometer la capacidad de las generaciones futuras para satisfacer sus propias necesidades.

Una respuesta viable ante tal situación, es ofrecer apoyo de alta tecnología al sector agrícola de la región que permita mejorar los rendimientos en la producción a niveles competitivos internacionalmente. Entes mundiales como la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), en su informe *E-Agriculture in Action* [81], reconocen que el papel de las Tecnologías de la Información y la Comunicación (TIC) en la agricultura ha crecido sustancialmente en los últimos años, tanto en escala como en alcance, y que urge desarrollar, adaptar y aplicar estas tecnologías como soluciones de e-agricultura para compensar algunos de los desafíos que enfrenta esta actividad. Se debe producir un 60 % más de alimentos para el año 2050, dado que se espera que la población mundial supere los 9 mil millones para ese año, y la agricultura tiene que aumentar la producción de alimentos nutritivos para satisfacer la creciente demanda y garantizar la seguridad alimentaria para todos [82].

En este sentido, una aplicación concreta de la tecnología en el campo agrícola es la llamada agromática, la cual según Grenón [43] se refiere a la aplicación de los principios y técnicas de la informática y la computación a las teorías y leyes del funcionamiento y manejo de los sistemas agropecuarios (sean estos desde un potrero, una empresa o hasta una región). Las aplicaciones agromáticas van desde sistemas de información que pueden ayudar en el proceso que gestiona el ciclo de vida del producto, hasta proyectos de investigación

científica que se centren en mejorar los procesos productivos. Particularmente la aplicación de la inteligencia artificial, en especial la disciplina denominada aprendizaje automático, adquiere una función primordial como herramienta base de predicción. El aprendizaje automático es el estudio de algoritmos de computadora que mejoran automáticamente a través de la experiencia. En 1959 Samuels afirmó que se trataba de la programación de computadoras para aprender de la experiencia y esperaba que esto eliminaría mucho del esfuerzo de programación [95]. Este tipo de aprendizaje ha sido utilizado en aplicaciones desde la minería de datos que descubren las reglas en grandes conjuntos de datos, hasta sistemas de filtración de información que automáticamente aprenden los intereses de los usuarios.

En la siguiente sección se presenta el por qué, entre las aplicaciones del aprendizaje automático, esta investigación del Doctorado en Ciencias Naturales para el Desarrollo (DOCINADE), opta por impactar el campo agrícola.

1.1. Seguridad alimentaria

En el año 2012, la Organización de las Naciones Unidas para la alimentación y la agricultura, indicó que la principal conclusión de la evaluación realizada a nivel mundial es que la agricultura parece afrontar una expansión impulsada por la demanda que está siendo cubierta principalmente por exportadores nuevos y emergentes, más que por los proveedores tradicionales. A pesar de ello, el aumento en el precio de los insumos y el costo de acceso desde zonas más aisladas, ha provocado subidas en los precios de los alimentos en términos reales. La Organización de las Naciones Unidas para la Alimentación y la Agricultura [80] se cuestiona si la producción será capaz de crecer al mismo ritmo que la demanda en los próximos años, de modo que se logren estabilizar los precios reales a sus pautas históricas, o si esos precios seguirán subiendo por la presión creciente de la demanda.

En la tabla 1.1 se muestran estadísticas que proporciona la FAO sobre la alimentación y la agricultura en el Mundo, en América Latina y el Caribe, particularizando los datos para Costa Rica. Del año 1990 al 2017, la población ha ido en aumento, un 42 % a nivel mundial, un 45 % para América Latina y un 58 % para Costa Rica. En cuanto al uso de la tierra, a nivel mundial, de 1990 al 2002, el área para la agricultura aumentó en un 1 %, pero del 2002 al 2015 disminuyó en un 1 %. Para Costa Rica el área dedicada a la agricultura disminuyó en un 21 % de 1990 al 2015, y para América Latina y el Caribe, las hectáreas con fines de agricultura aumentaron en un 10 % de 1990 al 2015. Este aumento en hectáreas en América Latina, al estudiar cada uno de los países que conforman el área, responde básicamente a dos países: Brasil y Argentina. Por ejemplo, de las 41.137,4 hectáreas de incremento del 2002 al 2015 en América Latina y el Caribe, 16.721 hectáreas corresponden al crecimiento en Brasil y 19.990 a Argentina; juntos representan un 89,24 % de la variación [85]. Se puede concluir que la población va en aumento y la superficie de la tierra con fines de agricultura no crece al mismo ritmo, sino que por el contrario, en muchas zonas más bien va disminuyendo, lo cual implica que se requiere mejorar la

productividad para producir más alimentos en menos hectáreas, y es precisamente a este proceso de eficiencia que la presente tesis pretende colaborar.

Tabla 1.1: Estadísticas de la FAO sobre población y uso de la tierra para fines de agricultura

Aspecto	Año	Mundial	América Latina y el Caribe	Costa Rica
Población total (millones)	1990	5.330,9	445,9	3,1
	2002	6.302,1	540,3	4,1
	2017	7.550,3	645,6	4,9
Uso de la tierra (1000 ha)	1990	4.831.300,5	687.119,6	2.305,0
	2002	4.940.642,4	715.248,5	1.826,0
	2015	4.868.989,5	756.385,9	1.811,0

(Tomado de Organización de las Naciones Unidas para la Alimentación y la Agricultura [85])

Aunado al panorama presentado anteriormente, el mundo se encuentra ante la realidad del cambio climático, el cual impone retos adicionales a la agricultura. La FAO en su publicación titulada "Climate smart agriculture - Building resilience to climate change" [84], acepta que el concepto de agricultura climáticamente inteligente (CSA) está ganando considerable aceptación a nivel internacional y nacional para enfrentar los desafíos de abordar la planificación agrícola bajo el cambio climático. CSA es un concepto que exige la integración de la necesidad de adaptación y la posibilidad de mitigación en las estrategias de crecimiento agrícola para respaldar la seguridad alimentaria.

En respuesta a la problemática esbozada en los párrafos anteriores, la presente tesis propone la utilización del aprendizaje automático para colaborar con los productores, particularmente con los pequeños productores. Como la misma FAO reconoce [84], las pequeñas explotaciones agrícolas y las comunidades rurales de los países en desarrollo son especialmente vulnerables a los efectos del cambio climático, y por tanto, es de esperar que el cambio climático exacerbará los desafíos existentes de escasez de recursos, restricciones crediticias, limitaciones de infraestructura e información y mercados incompletos.

Como casos de estudio de la presente investigación se tomaron dos cultivos: el café y el banano. Del café se analizaron datos relativos a la enfermedad denominada roya, y del banano se analizaron datos de producción y de la enfermedad denominada Sigatoka negra. Se trabajó con el Instituto del Café de Costa Rica (ICAFE) y la Corporación Bananera Nacional S.A. (CORBANA), ubicadas en Costa Rica, América Central. El contar con el apoyo de ambas Instituciones permitió utilizar los datos de estos procesos biológicos, lo cual era necesario para poder evaluar formalmente los métodos propuestos en esta investigación.

En la siguiente sección se presenta en qué conjuntos de datos, en particular, se trabajó para efectos de experimentación y su relevancia para el país.

1.2. Banano y café en Costa Rica

La tabla 1.2 muestra las estadísticas respecto a la producción y área cultivada de ambos productos en Costa Rica [85]. Consistente con la realidad mundial ya expresada en la sección anterior, para ambos productos el área cultivada va decreciendo con el pasar de los años, mientras que la producción va en aumento para el banano, no así para el café. Esto se debe a los efectos de las enfermedades, del cambio climático y a que Costa Rica ha ido cambiando su política de producción al dejar de competir por volumen, y tratar de posicionarse por calidad, aunque con menor volumen.

Tabla 1.2: Estadísticas de la FAO sobre producción y área cultivada de banano y café en Costa Rica para los años 2000, 2010 y 2016

Aspecto	Año	Banano	Café
Área cultivada (hectáreas)	2000	47.982	106.000
	2010	43.031	98.681
	2016	42.410	84.133
Producción (hectogramos/hectárea)	2000	454.545	15.226
	2010	469.389	9.284
	2016	568.154	10.399

(Tomado de Organización de las Naciones Unidas para la Alimentación y la Agricultura [85])

1.2.1. Banano

Desde un punto de vista económico y de acuerdo al último anuario estadístico disponible de la Promotora de Comercio Exterior de Costa Rica (PROCOMER) [26], el ingreso de divisas provenientes de las exportaciones de banano totalizó US$1.043,2 millones en 2017. La tabla 1.3 muestra la participación del banano en la economía nacional [26].

Tabla 1.3: El banano en la economía de Costa Rica (2017)

Concepto	Contribución
Valor	$1.043,2 USD Mill.
Participación en las exportaciones totales	9,4 %
Participación en las exportaciones agrícolas	36,9 %

(Tomado de Comercio Exterior de Costa Rica (PROCOMER) [26])

En Costa Rica, la Sigatoka negra se trata con frecuencia con fungicidas químicos. Dependiendo de la zona de producción y las condiciones climáticas, se requieren de 45 a 55 ciclos por año de aplicaciones de fungicidas para mantener esta enfermedad bajo control y para producir la calidad de fruta esperada para los mercados internacionales. Esto representa un costo por hectárea anual en el rango de US$1600 a US$2000; lo cual representa aproximadamente de entre US$0.64 a US$0.80 de los costos de producción para una caja de 18.14 kg, que corresponde de un 10 % a un 12 % del total de los costos de producción. En

ausencia de medidas de combate a la enfermedad, la Sigatoka negra puede reducir hasta en un 50 % el peso del racimo de banano y causar pérdidas del 100 % de la producción debido al deterioro en la calidad (longitud y grosor del fruto) [45].

Desde el punto de vista de la producción (floración), en Costa Rica se produjeron 2,195,736,00 toneladas en el año 2014, 2,008,155,00 en el 2015, 2,417,876,00 en el 2016 y 2,553,420,00 en el 2017. Y para mostrar la importancia relativa de este cultivo en la producción nacional, cabe indicar que para el año 2017 la producción de bananos representó un 20 % de la producción total de Costa Rica [89].

1.2.2. Café

Respecto al café, las exportaciones del café oro representaron un 3 % [26] del total exportado. La tabla 1.4 muestra la participación del café en la economía nacional.

Tabla 1.4: El café en la economía de Costa Rica (2017)

Concepto	Contribución
Valor	$299,2 USD Mill.
Participación en las exportaciones totales	3,0 %
Participación en las exportaciones agrícolas	11,0 %

(Tomado de Comercio Exterior de Costa Rica (PROCOMER) [26])

Desde un punto de vista económico y para poner en contexto los costos asociados a la producción del café, la renovación de cafetales en Costa Rica —cosecha 2018-2019— alcanzó los US$8.784,37 por hectárea [56]. Por su parte, para fincas productoras de café, de entre 50 a 70 fanegas por hectárea, los costos totales de producción por hectárea alcanzaron los US$5.238,44 —para los mismos años 2018 a 2019— [55].

Finalmente, aclarado el por qué de la selección del dominio sobre el que se efectuó la presente investigación, en la siguiente sección se delimita el enfoque con que se trabajó la aplicación del aprendizaje automático al dominio propuesto.

1.3. Delimitaciones al alcance de la investigación

En un contexto más amplio, la presente investigación apunta a la aplicación del aprendizaje automático a la predicción en cultivos agrícolas. Esto a través de la utilización de sensores e información recabada con cualquier otro medio oportuno, para tomar esta información y descubrir patrones que pueden ayudar a los especialistas en la materia a tomar decisiones basadas en los pronósticos que generan los resultados de esta investigación.

Ahora bien, de manera más específica, este trabajo tiene como objeto responder a llamados como el que hace la FAO en su informe titulado: “The state of food and agriculture - Leveraging food systems for inclusive rural transformation” [83], en el que expresa que los pequeños agricultores deben tener la escala necesaria para acceder a los mercados y

adoptar nuevas tecnologías (destacando la importancia de los servicios rurales públicos y la acción colectiva de los agricultores) o el acceso a tecnologías específicamente adaptadas a las operaciones en pequeña escala.

En consecuencia con lo indicado anteriormente, para la presente investigación se establecieron tres requerimientos a considerar en la solución. El primero es que la propuesta pueda ser utilizada por pequeños y medianos productores, por lo que no se desea incluir el uso de imágenes, ni hiperespectrales ni multiespectrales, dado que normalmente este tipo de imágenes no están al alcance de dicho tipo de productores. El segundo es que la propuesta equilibre la maximización de la capacidad de generalización con la minimización del error empírico, lo cual implica una optimización multiobjetivo (por ejemplo, el frente de Pareto [72] entre el coeficiente de determinación, R^2, y la raíz del error cuadrado medio, $RMSE$); esto para buscar transferir el aprendizaje adquirido con unos productores a otros y así promover la cooperación entre los mismos. Y en tercer lugar, que la propuesta no requiera estimar variables climatológicas (como precipitación, temperatura, humedad, radiación solar, entre otras) sino que solo se utilicen sus valores observados y la predicción sea sobre la variable que representa el proceso biológico en estudio, por ejemplo, el nivel de producción o el grado de desarrollo de una enfermedad. Esto debido a que de las variables climatológicas interesa el efecto ya producido en el proceso biológico en estudio y la predicción de las mismas requeriría contar con gran cantidad de registros históricos, los cuales normalmente este tipo de productores no posee. Además, como las variables climatológicas representan fenómenos caóticos, su predicción no es confiable, particularmente en micro climas.

Por tanto, el principal aporte de la presente investigación es una estrategia para iniciar con pequeños y medianos productores que no cuentan con los volúmenes de datos, ni tiene como requisito contar con grandes cantidades de imágenes etiquetadas, sino que permite trabajar de manera colaborativa con los expertos en el proceso biológico en estudio, para aplicar técnicas de aprendizaje automático con fines de predicción en los conjuntos de datos disponibles e ir produciendo una base de datos de conocimiento aprendido que permita hacer transferencia del aprendizaje para beneficiar otros pequeños y medianos productores.

Las condiciones expuestas anteriormente limitan la aplicabilidad de técnicas de moda en el aprendizaje automático (como por ejemplo las redes neuronales profundas o las redes bayesianas), pues éstas se caracterizan por una elevada cantidad de parámetros a ajustar, lo que presupone la existencia de volúmenes de datos inalcanzables en los contextos descritos. Esto no es óbice para afirmar que la estrategia es flexible para que en el momento en que se cuente con grandes volúmenes de datos, las técnicas anteriormente mencionadas y otras más, pueden ser incorporadas sin ningún cambio adicional a la estrategia propuesta.

Delimitado el alcance, en la siguiente sección se presenta la estrategia de solución.

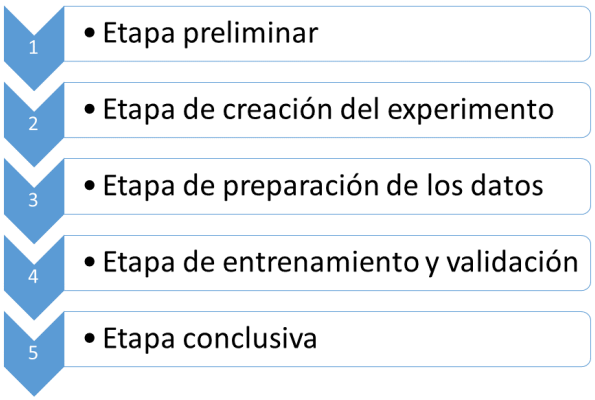


Figura 1.1: Etapas de la estrategia propuesta

1.4. Estrategia de solución

La estrategia propuesta se resume en la figura 1.1 y se detalla en el Capítulo 3.

La idea central detrás de la presente estrategia es conformar un conjunto de modelos de predicción, llamado Repositorio de Conocimiento Aprendido (RCA), que permita ir aprendiendo de los mismos modelos que se vayan conformando, de manera que se vuelva un medio colaborativo entre los productores. Para lograr esto, la estrategia se compone de cinco etapas. En la primera etapa, preliminar, se explicitan las condiciones de uso y se propone un lenguaje común. En la segunda etapa, creación del experimento, se configura el experimento a realizar; esto según los requerimientos del equipo investigador. En este punto se puede hacer uso de información ya contenida en el RCA. En la tercera etapa, preparación de los datos, se procesan los datos según la configuración definida en la etapa anterior. En la cuarta etapa, entrenamiento y validación, se ejecuta el diseño experimental definido y se registran los resultados obtenidos. Finalmente, en la quinta etapa, conclusiva, se analizan los resultados obtenidos, se trasladan a los interesados y se actualiza el RCA, de ser necesario.

Esbozada la estrategia de solución, en la siguiente sección se presentan los objetivos de la investigación.

1.5. Objetivos del estudio

El objetivo general es diseñar una nueva estrategia de predicción en procesos biológicos del campo agrícola con datos limitados.

Los objetivos específicos son los siguientes:

- Caracterizar los casos de aplicación de la estrategia propuesta.
- Elaborar una arquitectura de aprendizaje automático capaz de modelar los datos del proceso biológico en estudio.
- Proponer un proceso para validar el modelo propuesto.

- Precisar una manera para aprender de las iteraciones de la aplicación de la estrategia.

Finalmente, dados estos objetivos, el resto del documento es organizado de la siguiente manera: en el Capítulo 2 se presenta la teoría necesaria para comprender la propuesta y el estado del arte relacionado a la presente investigación. En el Capítulo 3 se presenta y explica la estrategia propuesta. En el Capítulo 4 se muestran y analizan los resultados producto de aplicar la estrategia a varios procesos biológicos. Finalmente, en el Capítulo 5 se presentan las conclusiones de la investigación y se plantean opciones de trabajo futuro.

Capítulo 2

Marco conceptual de la propuesta

Este capítulo incluye los conceptos necesarios para la adecuada intelección de la presente propuesta y el estado del arte del campo de estudio.

Con respecto a la presentación de conceptos y con el fin de tener un hilo conductor, se utiliza el título del presente trabajo, a saber, *Estrategia de predicción en procesos biológicos del campo agrícola con datos limitados: casos de aplicación en café y banano.* El orden de presentación de los conceptos es: estrategia, aprendizaje automático y predicción, procesos biológicos en el campo agrícola, y posteriormente se introducen conceptos que se requieren para comprender los procesos relacionados con las etapas de la presente propuesta, los cuales se agrupan de la siguiente manera: criterios de evaluación y herramientas matemáticas.

2.1. Definición de estrategia y términos similares

En general, se considera una estrategia como un conjunto de acciones planificadas y coordinadas que se llevan a cabo para lograr un determinado fin. White [114] la define como una serie coordinada de acciones que involucran el despliegue de recursos a los que se tiene acceso para el logro de un propósito determinado, de manera que la estrategia es como una idea unificadora que vincula el propósito y la acción. Es por la definición anterior, que la presente propuesta se presenta como una estrategia, pues propone un conjunto de acciones a realizar y organiza el aporte de los involucrados, de manera que se logre realizar el pronóstico deseado.

Ahora bien, en las publicaciones científicas relacionadas con la aplicación del aprendizaje automático en el campo agrícola, se encuentran propuestas con términos como los siguientes: **metodología** (*methodology*), Green methodology for soil organic matter analysis using a national near infrared spectral library in tandem with learning machine [97], Improved machine learning methodology for high precision agriculture [107]; **acercamiento** (*approach*), Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review [24], An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning [33]; **marco de trabajo**

(*framework*), A framework for detection and classification of plant leaf and stem diseases [2], A framework for the management of agricultural resources with automated aerial imagery detection [94]; **modelo** (*model*), Crop Production - Ensemble Machine Learning Model for Prediction [12], Predictive models in horticulture: A case study with Royal Gala apples [65]. Dichos artículos no entran a definir concretamente qué entienden por cada uno de estos términos, pero al analizar su contenido se constata que no se alejan del concepto de estrategia (*strategy*) utilizado en el presente trabajo, en cuanto presentan un conjunto de pasos a seguir para lograr un objetivo, algunos con mayor o menor detalle del proceso que los otros.

Clarificado qué es estrategia y cómo se relaciona con ciertos conceptos similares, en la siguiente sección se presenta qué es aprendizaje automático y predicción, reflexión necesaria dado que la presenta propuesta utiliza aprendizaje automático con fines de predicción.

2.2. Aprendizaje automático y predicción

Para Murphy [79], el aprendizaje automático (del inglés *machine learning*) es un conjunto de métodos que pueden automáticamente detectar patrones en los datos, y entonces usar los patrones descubiertos para predecir datos futuros, o para ejecutar otra clase de toma de decisión bajo incertidumbre (como por ejemplo planificar cómo recolectar más datos). El autor propone dividir el aprendizaje automático en tres tipos [79]:

- Aprendizaje supervisado: también llamado predictivo. El objetivo es aprender a mapear entradas x a salidas y, dado un conjunto etiquetado de pares de entrada-salida $D = \{(x_i, y_i)\}_{i=1}^{N}$, donde D es llamado el conjunto de entrenamiento y N es el número de ejemplos de entrenamiento. Cuando y es categórico, el problema es conocido como clasificación o reconocimiento de patrones, y cuando y es un valor real, el problema es conocido como regresión.

- Aprendizaje no supervisado: también llamado descriptivo. En este caso se tienen solo las entradas $D = \{x_i\}_{i=1}^{N}$ y el objetivo es encontrar lo que llama el autor *patrones interesantes* en los datos. A este tipo se le suele llamar descubrimiento de conocimiento.

- Aprendizaje por refuerzo: es utilizado para aprender cómo actuar o cómo comportarse cuando se dan señales ocasionales de premio o castigo.

Por su parte, Ayodele [10] considera que el aprendizaje automático consiste en diseñar algoritmos que permiten que una computadora aprenda. El aprendizaje no necesariamente implica consciencia, sino que aprender es una cuestión de encontrar regularidades estadísticas u otros patrones en los datos. Este autor propone una clasificación más amplia [10]:

- Aprendizaje supervisado: donde el algoritmo genera una función que asigna entradas a las salidas deseadas. Una formulación estándar de la tarea de aprendizaje

supervisado es el problema de clasificación, donde se trata de aprender de varios ejemplos de entrada y salida, para luego asignar una de varias clases.

- Aprendizaje no supervisado: que modela un conjunto de entradas y donde no se cuenta con un conjunto de ejemplos etiquetados.
- Aprendizaje semi-supervisado: que combina ejemplos etiquetados y no etiquetados para generar una función o clasificador apropiado.
- Aprendizaje por refuerzo: donde el algoritmo aprende una política de cómo actuar dada una observación del mundo. Cada acción tiene algún impacto en el entorno, y el entorno proporciona información que guía el algoritmo de aprendizaje.
- Transducción: similar al aprendizaje supervisado, pero no construye explícitamente una función. En su lugar, intenta predecir nuevos resultados en función de las entradas del entrenamiento, las salidas del entrenamiento y las nuevas entradas.
- Aprender a aprender: donde el algoritmo aprende su propio sesgo inductivo basado en experiencia previa.

Aunque la estrategia propuesta no se limita a un conjunto específico de técnicas de aprendizaje automático, se presentan seguidamente las utilizadas en los casos de estudio presentados en el Capítulo 4.

2.2.1. Regresión ordinaria de mínimos cuadrados (OLSR)

Del inglés *ordinary least squares regression* (OLSR). Sea D un conjunto de datos.

$$D = \{(\mathbf{x}_i, y_i) \mid i = 1 \dots n\} \tag{2.1}$$

compuesto de n vectores de atributos d-dimensionales $\mathbf{x}_i \in \mathbb{R}^d$ y las respuestas correspondientes [1] y_i. El OLSR ajusta un modelo lineal

$$\tilde{y}_i = f(\mathbf{x}_i) = \langle \mathbf{w}, \mathbf{x}_i \rangle \tag{2.2}$$

donde $\langle \cdot, \cdot \rangle$ denota el producto interno de vectores, tal que la suma de los cuadrados de los residuales $(\tilde{y}_i - y_i)$ es minimizado.

Sea X la matriz de diseño $n \times d$ que contiene la i-ésima muestra de datos $\mathbf{x}_i^T$ en su i-ésima fila y sea $\mathbf{y}$ el vector de todas las respuestas y_i correspondientes a cada fila; entonces la regresión del cuadrado mínimo se halla en

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} E(\mathbf{w}) \tag{2.3}$$

con el error de la función

$$E(\mathbf{w}) = \|X\mathbf{w} - \mathbf{y}\|_2^2 \tag{2.4}$$

[1] Se utiliza la convención de que el primer componente de cada vector $\mathbf{x}_i$ es 1.

La solución de forma cerrada se halla por medio de la matriz pseudo-inversa

$$\hat{\mathbf{w}} = X^{+}\mathbf{y} = (X^{T}X)^{-1}X^{T}\mathbf{y} \tag{2.5}$$

que se puede calcular de forma numéricamente robusta con la descomposición en valores singulares de X [88]. Métodos de descenso de gradiente iterativos son también aplicables para minimizar este error [16].

En este trabajo se utilizó el método de descomposición en valores singulares de X [86].

2.2.2. Regresión de red elástica (ENR)

En la regresión de red elástica (del inglés *elastic net regression*) (ENR), se agregan términos adicionales de regularización a la función de error (2.4) con el fin de imponer más restricciones a la solución.

Por ejemplo, en la regresión de cresta (del inglés *ridge regression*) (RR) un término de regularización a priori $\alpha \|\mathbf{w}\|_2^2$ se incluye para preferir soluciones con normas pequeñas.

En regresión lazo (del inglés *lasso regression*) (LR) se usa en su lugar un término $\lambda \|\mathbf{w}\|_1$ [106], el cual permite seleccionar un subconjunto de atributos disponibles poniendo en cero los pesos de los atributos ignorados.

Si la dimensión d de los datos es mucho más grande que el número de las n muestras de datos, lazo seleccionará un máximo de n variables.

La ENR, [120], combina ambos términos de los estimadores de cresta y lazo por medio de la función de error.

$$E(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \alpha \|\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Por lo tanto, OLSR, RR, y LE son casos particulares de ENR.

La combinación de los términos de regularización permite aprender un modelo esparcido con solo unos pocos pesos que sean no-cero como en el caso de lazo, pero manteniendo las propiedades de regularización de la regresión de cresta [86].

La regresión de red elástica es útil cuando se correlacionan múltiples atributos: la regresión lazo probablemente elegirá uno de estos al azar, mientras que la regresión de red elástica probablemente elegirá ambos.

2.2.3. Regresión con vectores de soporte (SVR)

En la regresión con vectores de soporte (del inglés *support vector regression*) (SVR), la función de regresión está usualmente formulada como

$$\tilde{y} = f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \tag{2.6}$$

Los pesos se seleccionan en un procedimiento de optimización convexo [99]:

$$\begin{aligned} \text{minimizar} \quad & E(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{sujeto a} \quad & \begin{cases} y_i - \langle \mathbf{w},\mathbf{x}_i\rangle - b & \leq \epsilon + \xi_i \\ \langle \mathbf{w},\mathbf{x}_i\rangle + b - y_i & \leq \epsilon + \xi_i^* \\ \xi_i,\xi_i^* & \geq 0 \end{cases} \end{aligned} \tag{2.7}$$

donde ϵ es la desviación máxima permitida de los $\tilde{y}_i$ objetivo de las respuestas y_i, las variables de holgura ξ_i y ξ_i^* permiten hacer frente a las restricciones no factibles para el problema de optimización, y la constante $C > 0$ controla el balance entre la capacidad f y la tolerancia a las desviaciones mucho mayores que ϵ.

Dado que OLSR y ENR usan una función de error al cuadrado, los datos atípicos tendrán una fuerte influencia en los pesos resultantes $\mathbf{w}$. En la formulación de SVR, sin embargo, el uso de la norma L_2 combinada con las variables de holgura, considerablemente restringe, o completamente bloquea, la influencia de esos valores atípicos.

El problema SVR es entonces reformulado como un problema de optimización dual en [99]:

$$\mathbf{w} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathbf{x}_i \tag{2.8}$$

donde $\alpha_i,\alpha_i^* \in [0,C]$ son los multiplicadores de Lagrange sujetos a $\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0$. En este, llamado expansión con vectores de soporte, los pesos son expresados como una combinación lineal de los patrones de los conjuntos de datos $\mathbf{x}_i$. Insertando (2.8) en (2.6) dirige a

$$f(\mathbf{x}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\,\langle \mathbf{x}_i,\mathbf{x}\rangle + b \tag{2.9}$$

Ambos multiplicadores de Lagrange α_i,α_i^* son distintos de cero solo para aquellos puntos donde $|f(\mathbf{x}_i) - y_i| \geq \epsilon$. Por tanto, la expansión de $\mathbf{w}$ en términos de $\mathbf{x}_i$ es dispersa. Aquellos puntos de datos con coeficientes que no desaparecen son llamados *vectores de soporte* [113].

Adicionalmente, en (2.9) es posible emplear el llamado *truco del núcleo* y reemplazar los términos $\langle \mathbf{x}_i,\mathbf{x}\rangle$ con la evaluación de cualquier núcleo Mercer.

$$K(\mathbf{x}_i,\mathbf{x}) = \langle \phi(\mathbf{x}_i),\phi(\mathbf{x})\rangle \tag{2.10}$$

donde $\phi(\cdot)$ es un mapeo no lineal del espacio de entrada en un espacio característico de una dimensión mucho mayor (incluso infinita).

La evaluación del núcleo hace innecesaria la evaluación explícita del mapeo no lineal, y permite resolver regresiones no lineales en el espacio de entrada al mapear implícitamente las muestras a través del núcleo en el espacio dimensional superior, donde ocurre la regresión lineal [5].

Los núcleos usados en los casos de estudio son:

- núcleo lineal:
$$K(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i{}^T\mathbf{x}_j \tag{2.11}$$
- núcleo gaussiano:
$$K(\mathbf{x}_i,\mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \tag{2.12}$$
donde γ es inversamente proporcional a la varianza de la curva gaussiana.
- núcleo sigmoide:
$$K(\mathbf{x}_i,\mathbf{x}_j) = \tanh\left[c + \gamma\mathbf{x}_i^T\mathbf{x}_j\right] \tag{2.13}$$
donde γ es el vector de escalamiento, y $c \geq 0$.
- núcleo polinomial:
$$K(\mathbf{x}_i,\mathbf{x}_j) = \left(c + \gamma\mathbf{x}_i^T\mathbf{x}_j\right)^d \tag{2.14}$$
donde γ es el vector de escalamiento, d es el grado del polinomio y $c \geq 0$.

Finalmente, considerando lo presentado en esta sección, se puede decir que la presente propuesta versa sobre la aplicación del aprendizaje automático de tipo supervisado y con fines predictivos en procesos biológicos en el campo agrícola, por lo que la siguiente sección aclara qué se debe entender por este tipo de procesos.

2.3. Procesos biológicos en el campo agrícola

Según Gu, Kwok, Lam y col. [44], los procesos biológicos son una serie de reacciones bioquímicas, eventos y funciones moleculares que ocurren en los organismos vivos y son esenciales para que un organismo pueda vivir. Estos procesos son específicamente pertinentes a la función de las células vivas, tejidos y organismos. La presente tesis se enfoca en aplicar la estrategia propuesta a este tipo de procesos, particularmente a los relacionados con el campo agrícola. En las siguientes secciones se presentan tres de estos procesos biológicos, de los cuales tratan los casos de estudio que se utilizarán para mostrar la aplicación de la propuesta: la Sigatoka negra, la roya y la floración del banano.

2.3.1. Enfermedad del banano: Sigatoka negra

El hongo *Mycosphaerella fijiensis Morelet* causa la enfermedad denominada Sigatoka negra, la cual es el mayor problema patológico de las plantaciones de banano en América Central, Panamá, Colombia y Ecuador, como también en zonas de África y Asia [70]. Esta enfermedad ataca las hojas de la planta produciendo un rápido deterioro del área de la hoja. Esto afecta el crecimiento y la productividad de las plantas debido a la afectación en el proceso de fotosíntesis. Adicionalmente, causa una reducción de la calidad de la fruta y promueve una maduración prematura de los racimos, la cual es la mayor causa de pérdidas de producto asociada con la Sigatoka negra.

(a) (b) (c)

Figura 2.1: Ejemplos de tres estados de la Sigatoka negra: (a) Estado inicial. (b) Estado intermedio, y (c) Estado avanzado.
(Tomado de Marín Vargas y Romero Calderón [70])

Por lo anterior, se han desarrollado sistemas de preaviso para detectar la enfermedad y monitorear su proceso. Como ejemplo está el sistema de preaviso biológico, desarrollado por [39] y modificado por [38] para el control de la Sigatoka amarilla en Camerún, el cual luego fue adaptado por [104] y [36] a la Sigatoka negra. Este sistema de preaviso biológico está basado en observaciones semanales sobre un conjunto de plantas seleccionadas de modo representativo en el lugar de estudio. Se determina de manera manual el estado de desarrollo en las tres hojas más jóvenes de cada planta [35] y luego estos registros son utilizados para calcular indicadores empíricos, los cuales cuantitativamente describen la progresión de la enfermedad. Se usan coeficientes empíricos basados en la incidencia y severidad del desarrollo de la enfermedad para calcular dos variables: la suma bruta y el estado de evolución. La suma bruta está basada en el estado presente y un coeficiente empírico, el cual incrementa con el avance de los síntomas en la juventud de la hoja. El estado de evolución es calculado usando la suma bruta y el periodo de emisión foliar [71]. La figura 2.1 muestra un ejemplo de tres estados de desarrollo de la Sigatoka negra.

Según Chuang y Jeger [25], las tasas pasadas y presentes del desarrollo de la enfermedad pueden, en principio, ser usados para pronosticar su comportamiento futuro y determinan si un programa particular de fungicidas será capaz de tratar efectivamente la enfermedad de una manera que resulte también económicamente factible.

Finalmente, hay evidencia que los datos meteorológicos junto con las observaciones directas del desarrollo de los síntomas en el campo es un método común utilizado para pronosticar la evolución de esta enfermedad [21].

2.3.2. Enfermedad del cafeto: Roya

Según Barquero Miranda [13], la roya del cafeto (*Hemileia vastatrix*) es una de las enfermedades de mayor importancia económica que afectan el café. En esta enfermedad, el principal factor que condiciona su desarrollo es la relación entre el hospedante (plantas de cafeto), el patógeno (la roya) y el ambiente (variación del clima).

Se trata de un hongo que se desarrolla únicamente en el tejido vivo de la planta que lo hospeda, en este caso las hojas del cafeto. La figura 2.2 muestra varias de las etapas por las que pasa la enfermedad. Al inicio, los síntomas consisten en pequeñas lesiones

Figura 2.2: Muestra de los síntomas provocados por la roya del cafeto. (A) Manchas traslúcidas, (B) Progreso de la infección, (C) Lesiones viejas de roya.
(Tomado de Barquero Miranda [13])

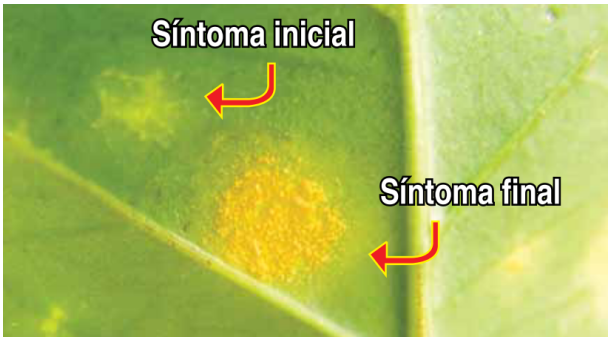


Figura 2.3: Lesiones viejas de roya presentes durante la época seca
(Tomado de Barquero Miranda [13])

o manchas redondas, color amarillo pálido, de 1 a 3 milímetros de diámetro (A). Esta mancha es traslúcida, pero aumenta gradualmente de tamaño al iniciarse la formación de esporas por el envés de la hoja y puede alcanzar los 2 cm de diámetro, se torna de color naranja y la superficie se vuelve polvosa. Si existen muchas lesiones o manchas, estas crecen hasta unirse unas con otras cubriendo toda la hoja y provocando su caída (B). Cuando las manchas de la roya envejecen, el polvo anaranjado se torna de un color naranja pálido y posteriormente en el centro de la lesión amarilla surge una mancha de color café marrón o negro de apariencia seca, que crece hasta cubrir toda la superficie de la lesión y donde no se producen esporas (C) [13].

Es posible observar alrededor de la mancha marrón, en muchas ocasiones, un borde de color amarillo, donde luego se producirán esporas de la roya si existen las condiciones de clima favorables para la esporulación (figura 2.3). Este tipo de lesiones representa la fuente de infección principal al inicio del siguiente periodo lluvioso [13].

2.3.3. Floración del banano

El principal factor que afecta la fenología de las plantas es la temperatura y se sabe que incrementos en la temperatura del aire pueden ser detectados fácilmente en los datos fenológicos [6].

Por su parte, el banano es una planta propia de los trópicos y subtrópicos, que requiere

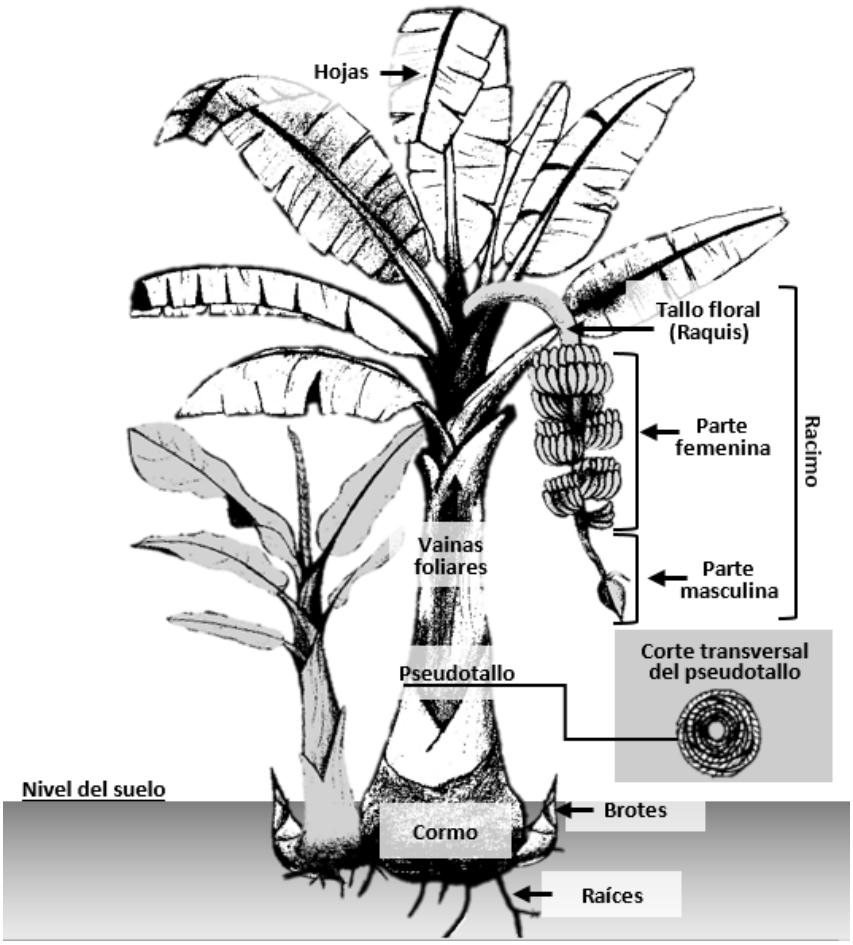


Figura 2.4: Esquema de una unidad productiva del cultivo de banano al momento de la floración.

(Tomado de Montero González [76])

un clima cálido y húmedo. El clima más adecuado es aquel con clima cálido y húmedo durante todo el año, sin vientos fuertes. La tasa de aparición de nuevas hojas y la tasa de crecimiento del fruto se rigen en gran medida por la temperatura. El desarrollo de la planta de banano se refleja al ritmo en que se producen las nuevas hojas. Si bien el suministro de nutrientes y agua puede influir en la tasa de aparición de nuevas hojas, el factor dominante de la conducción es la temperatura [91].

Por tal razón, es importante desde un punto de vista productivo, evaluar las frecuencias estacionales fenológicas en eventos tales como la emisión foliar, floración y crecimiento. La tasa de emisión foliar y el desarrollo de un cultivo están en función de las unidades térmicas o grados-día asociados con la producción [42].

La figura 2.4 presenta de manera esquemática una unidad productiva del cultivo de banano al momento de la floración. Esta figura es una adaptación realizada por [76] a la propuesta inicial de [23].

Expuestos estos tres procesos biológicos que serán utilizados en los casos de estudio más adelante (Capítulo 4), en la siguiente sección se exponen los criterios de evaluación ati-

nentes a la presente investigación.

2.4. Criterios de evaluación

Esta sección incluye cuatro conceptos:

1. Métricas para la comparación de resultados: Utilizadas en la propuesta para poder comparar los resultados que producen los diferentes experimentos.
2. Frente de Pareto: Utilizado en la optimización multiobjetivo para elegir el resultado óptimo considerando varias métricas.
3. Validación cruzada: Es uno de métodos utilizados en el entrenamiento y validación.
4. Diseño estadístico de experimentos: Utilizado en la propuesta para determinar si hay evidencia estadística de que los resultados obtenidos no son fruto de la aleatoriedad.

2.4.1. Métricas para la comparación de resultados

Si bien se podrían utilizar otras métricas en la estrategia propuesta, se presentan las utilizadas en los casos de estudio presentados del Capítulo 4, a saber: R^2 y $RMSE$.

Dados n registros y_i, $i = 1 \dots n$ del verdadero resultado de un proceso, el promedio $\bar{y}$ de los datos observados dado por:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

entonces, el error cuadrático medio (MSE) S_e^2 y la varianza explicada S_R^2 se estiman con [47]:

$$S_e^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad S_R^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

donde $\hat{y}_i$ es el valor predicho para y_i. La raíz del error cuadrático medio se define como $\text{RMSE} = \sqrt{S_e^2}$ y el coeficiente de determinación está dado por:

$$R^2 = \frac{S_R^2}{S_R^2 + S_e^2}$$

2.4.2. Frente de Pareto

En la sección anterior se presentaron dos métricas utilizadas para comparar los resultados obtenidos en los experimentos, pero en la toma de decisiones se requiere un paso más, sopesar el valor obtenido de dos o más métricas para decidir. Es por esto que la estrategia propuesta sugiere utilizar el concepto de frente de Pareto como medio para decidir considerando los resultados de las métricas: R^2 y $RMSE$.

Hernández [48] indica que el concepto de frente de Pareto, frontera de Pareto, óptimo de Pareto ó eficiencia de Pareto, es un método utilizado cuando se requiere optimizar más de una función (*optimización multiobjetivo*). Sea F un vector para el cual sus componentes son las distintas propiedades a optimizar. El planteamiento de la optimización multiobjetivo es [8]:

$$\begin{aligned} \text{minimizar} \quad & F(x) = |f_1(x),...,f_K(x)| \\ \text{sujeto a} \quad & \Big\{ x \in X \end{aligned} \tag{2.15}$$

donde K es el número de funciones objetivo y X es el espacio de soluciones factibles.

Se busca un vector de variables X tal que para $k = 1,...,K$:

$$f_k(X^*) = \min f_k(X) \tag{2.16}$$

sin embargo, lo anterior no suele producirse habitualmente, en cuyo caso se acude a los puntos definidos como óptimos de Pareto que son aquellos puntos X^P para los que no existe ningún punto X tal que para $k = 1,...,K$:

$$f_k(X) \leq f_k(X^P) \tag{2.17}$$

Y para al menos una función objetivo se cumple que:

$$f_k(X) < f_k(X^P) \tag{2.18}$$

La característica de este tipo de óptimalidad Pareto es que al disminuir el valor de alguna función objetivo, se incrementa al menos una de las restantes funciones objetivo [48].

En la figura 2.5, la solución "a" pertenece a un conjunto de soluciones óptimas (o soluciones no dominadas) dado que no puede encontrarse una solución "b", tal que mejore uno de los objetivos sin empeorar al menos uno de los otros. Por su parte, la solución "c" es dominada por "a" y por "b" [8].

2.4.3. Validación cruzada

Todo proceso de aprendizaje automático requiere una etapa de validación de sus resultados y como la presente propuesta se centra en casos en que no se disponen de grandes cantidades de datos (ver capítulo 1), es clave tener un método adecuado para determinar los conjuntos de entrenamiento y pruebas. Al respecto, Witten [116] recomienda en estos casos entrenar el modelo con los datos disponibles y utilizar subconjuntos de los datos para validar la predicción [116], este proceso se llama validación cruzada (*cross-validation*).

El proceso consiste en dividir el total de datos en 10 partes y correrlo diez veces, en cada una de las corridas nueve de esas partes se utilizan como conjunto de entrenamiento y la parte restante como conjunto de validación y se calcula el error de predicción. Terminadas las 10 corridas se promedian los valores obtenidos para la métrica y así se obtiene un indicador para ese modelo, este proceso es llamado validación cruzada de diez iteraciones (*ten-fold cross-validation*) [116].

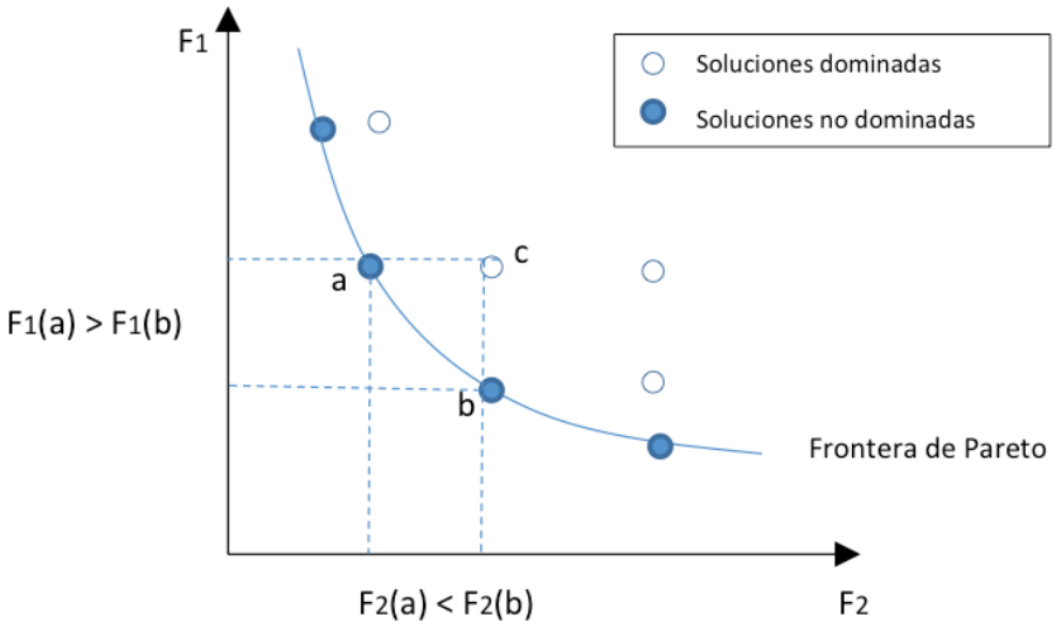


Figura 2.5: Ejemplo de optimización de una variable con dos funciones objetivo. (Tomado de Aranda Pinilla y Orjuela Castro [8])

2.4.4. Diseño estadístico de experimentos

La presente estrategia propone validar si se cumplen los supuestos para realizar un diseño estadístico de experimentos tal y como lo presenta [77]; esto con el fin de mostrar si las diferencias obtenidas son estadísticamente significativas.

En este sentido, [77] propone definir: factores, niveles, métricas y luego se realizan pruebas de significancia. En dichas pruebas se quiere demostrar si la diferencia de las medias en los resultados obtenidos de los diferentes niveles de cada factor son estadísticamente significativas y no son producto de la aleatoriedad.

En este punto se debe seleccionar entre dos tipos de métodos, los paramétricos, que son preferidos, y los no paramétricos. Entre los métodos paramétricos se encuentra la prueba ANOVA (1-way ANOVA [46]) que es una de las más robustas, pero que precisa de que las muestras cumplan tres requisitos: 1) normalidad, cada una de las muestras proviene de una población con distribución normal [30], 2) independencia, que las muestras sean independientes entre sí [100], y 3) homocedasticidad, que la varianza del error condicional a las variables explicativas es constante a lo largo de las observaciones [100]. En caso de no cumplirse uno o más de los tres requisitos, se propone realizar la prueba de hipótesis con un método no paramétrico. En el caso de esta propuesta, se propone utilizar el Kruskal-Wallis H-test [60] si las muestras son independientes entre sí y el Wilcoxon signed-rank test [115] si no lo son.

Presentados los criterios de evaluación, en la siguiente sección se comentan dos herramientas matemáticas que serán utilizadas en las etapas de la estrategia propuesta.

2.5. Herramientas matemáticas

La presente sección incluye dos herramientas: el coeficiente de variación multivariable y la técnica denominada t-Distributed stochastic neighbor embedding. La primera será utilizada para el aumento de datos y la segunda para la propuesta de aprendizaje por transferencia.

2.5.1. Coeficiente de variación multivariable

Una dificultad que se presenta al utilizar técnicas de aprendizaje automático, es no contar con suficientes datos para los experimentos, por lo que es conveniente buscar métodos para hacer aumento de datos (del inglés *data augmentation*). En esta propuesta se propone un método para generar muestras a partir de las series de datos existentes, pero para esta propuesta se requiere conocer qué tan variable es la serie de datos original. Es en ese punto que se propone utilizar el concepto de coeficiente de variación multivariable, que se procede a describir a continuación.

Albert y Zhang [3], a partir del coeficiente de variación (CV), el cual se utiliza para medir la variación relativa de una variable aleatoria respecto a su media, proponen extender el caso uni-variable al caso multi-variable, denominado coeficiente de variación multivariable (CV_m). Similar al CV, este nuevo coeficiente también es medido como un porcentaje, y entre mayor sea el valor del coeficiente, indica una variabilidad mayor del conjunto de variables. Los autores lo definen de la siguiente manera:

Sea $X = (X_1,...,X_p)^T$ un vector aleatorio normal p-dimensional con media $\mu \neq 0$ y con matriz de covarianza Σ. El CV_m de X se calcula de la siguiente manera:

$$CV_m = \left[\frac{\mu^T \Sigma \mu}{(\mu^T \mu)^2} \right]^{1/2} \tag{2.19}$$

De este coeficiente se destaca la propiedad de que las formas cuadráticas $\mu^T\Sigma\mu$ y $\mu^T\mu$ en (2.19) siempre existen y no se requiere invertir la matriz. Por lo tanto, el CV_m se puede calcular en toda generalidad.

2.5.2. t-Distributed stochastic neighbor embedding (t-SNE)

Como se indicó en el Capítulo 1, se desea que la propuesta permita un trabajo colaborativo entre agricultores, por lo que poder compartir el conocimiento aprendido es clave en el proceso. Para lograrlo, la presente propuesta incluye un proceso que permite buscar en el RCA si hay conocimiento aprendido que pueda ser utilizado para complementar otro conjunto de datos. Se propone utilizar un método para comparar conjuntos de datos y valorar su similitud relativa, de manera que sirvan para hacer aumento de datos. En este punto, se aprovecha la idea de [68], quienes propusieron una técnica llamada t-Distributed stochastic neighbor embedding (t-SNE) que visualiza datos de alta dimensión al asignar a cada punto de datos una ubicación en un mapa de dos o tres dimensiones. Esta técnica ha

recibido propuestas de mejora como la presentada en Accelerating t-SNE using Tree-Based Algorithms por [67].

t-SNE [67] minimiza la divergencia entre dos distribuciones: una distribución que mide similitudes por pares de los objetos de entrada y una distribución que mide similitudes por pares de los correspondientes puntos de baja dimensión en la incrustación. Dado un conjunto de datos de objetos de alta dimensión $D = \{x_1, x_2,....x_N\}$ y una función $d(x_i, x_j)$ que calcula la distancia entre un par de objetos, por ejemplo, la distancia Euclidiana $d(x_i,x_j) = \|x_i - x_j\|$. El objetivo es aprender una incrustación s-dimensional en la cual cada objeto es representado por un punto, $\mathcal{E} = \{y_1, y_2,...,y_N\}$ con $y_i \in \mathbb{R}^s$ (los valores típicos para s son 2 ó 3). Para este fin, t-SNE define probabilidades conjuntas p_{ij} que indican la similitud entre los objetos x_i y x_j por medio de la simetría de las dos probabilidades condicionales como sigue:

$$p_{j|i} = \frac{\exp\left(-d(x_i,x_j)^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-d(x_i,x_k)^2/2\sigma_i^2\right)} \tag{2.20}$$

y donde $p_{i|i} = 0$. Además:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{2.21}$$

En la ecuación anterior, el ancho de banda de los núcleos Gaussianos, σ_i, se establecen de tal manera que la perplejidad de la distribución condicional P_i sea igual a una perplejidad predefinida u. Como resultado, el valor óptimo de σ_i varía según el objeto: en las regiones del espacio de datos con una mayor densidad de datos, σ_i tiende a ser más pequeño que en las regiones del espacio de datos con menor densidad. El valor óptimo de σ_i para cada objeto de entrada se puede encontrar usando una búsqueda binaria simple o usando un método robusto de búsqueda de raíces. En la incrustación s-dimensional $\mathcal{E}$, las similitudes entre los dos puntos y_i y y_j (por ejemplo, los modelos de baja dimensión de x_i y x_j) se miden utilizando un núcleo normalizado de cola gruesa. Específicamente, la similitud incrustada q_{ij} entre los dos puntos y_i e y_j se calcula como un núcleo normalizado de t-Student con un solo grado de libertad:

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1 + \|y_k - y_l\|^2\right)^{-1}} \tag{2.22}$$

Las colas gruesas del núcleo t-Student normalizado permiten que los objetos de entrada disímiles x_i y x_j sean modelados por contra partes de baja dimensión y_i y y_j que están demasiado separadas. Esto es deseable porque crea más espacio para modelar con precisión las pequeñas distancias por pares (por ejemplo, la estructura de datos local) en la incrustación de baja dimensión. Las ubicaciones de los puntos de incrustación y_i se determinan minimizando la divergencia de Kullback-Leibler (divergencia-KL) entre las distribuciones conjuntas P y Q:

$$C(\mathcal{E}) = KL(P\|Q) = \sum_{i\neq j} p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{2.23}$$

Debido a la asimetría de la divergencia de *Kullback-Leibler*, la función objetivo se centra en modelar valores altos de objetos similares p_{ij} (objetos similares) por valores altos de q_{ij} (puntos cercanos en el espacio de incrustación). La función objetivo no es convexa en la incrustación $\mathcal{E}$. Por lo general, se minimiza por descenso de gradiente:

$$\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij}) q_{ij} \mathcal{Z} (y_i - y_j) \tag{2.24}$$

donde se define el término de normalización $\mathcal{Z} = \sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}$.
Se aprecia que la evaluación de las distribuciones conjuntas P y Q es $\mathcal{O}(N^2)$, porque ambas distribuciones implican un término de normalización que suma a todos los $N(N-1)$ pares de objetos únicos. Dado que *t-SNE* se escala de forma cuadrática en el número de objetos N, su aplicabilidad se limita a conjuntos de datos con solo unos pocos miles de objetos de entrada; más allá de eso, el aprendizaje se vuelve demasiado lento para ser práctico (y los requisitos de memoria principal son grandes) [67].

Expuestos los anteriores conceptos con miras a la comprensión de la propuesta, en la siguiente sección se presenta el estado del arte del campo de estudio.

2.6. Estado del arte

En esta sección se presenta un resumen de las principales investigaciones en el campo de estudio en cuestión, teniendo en mente que en este punto no se hará una comparación detallada de los resultados obtenidos, pues será en las secciones del Capítulo 4, Resultados y análisis, que se efectuarán.

Varias propuestas se han realizado para aplicar técnicas del aprendizaje automático al descubrimiento automático de relaciones entre variables ambientales e indicadores numéricos de interés agrícola.

Holloway y Mengersen [49] presentan una revisión de la literatura respecto a los métodos del aprendizaje automático estadístico aplicados a los datos recabados con sensores remotos. El artículo se centra en los objetivos del Desarrollo Sostenible del Banco Mundial de las Naciones Unidas, incluyendo agricultura, bosques y agua. Luego de presentar varias tablas de resultados experimentales, concluyen que la selección del método de análisis de los datos provenientes de sensores remotos depende de varios factores: la naturaleza y cantidad de los datos de entrenamiento, la cantidad de datos de referencia (*ground truth*), el tipo de estimaciones y las inferencias requeridas, y la disponibilidad de software y poder de procesamiento para modelar. Esta conclusión es congruente con la delimitación del alcance que se realizó en la sección 1.3.

Konstantinos, Busato, Moshou y col. [59] realizan una revisión de la investigación dedicada a las aplicaciones del aprendizaje automático en sistemas de producción agrícola. Los autores categorizaron los trabajos en: a) Manejo de cultivos, incluidas las aplicaciones

de predicción de rendimiento, detección de enfermedades, detección de malezas, calidad de cultivos y reconocimiento de especies, b) manejo de ganado, incluyendo aplicaciones en bienestar animal y producción ganadera, c) gestión del agua, y d) manejo del suelo. Los autores resumen 40 artículos relacionados al tema y una de sus conclusiones es que lo publicado se concentra en el manejo de cultivos con un 61 %, le sigue la predicción de rendimiento 20 % y la detección de enfermedades 22 %. Las investigaciones analizadas prefieren el uso de imágenes (espectral, hiper-espectral, infrarrojo cercano (*NIR*), etc.) y la disponibilidad de grandes cantidades de datos, lo cual es ideal si se disponen de ellas, pero el reto sigue siendo aportar una estrategia que pueda dar resultados sin tener inicialmente ese volumen de datos, que es precisamente uno de los principales aportes de la presente propuesta.

Otras aplicaciones de los métodos de aprendizaje automático en la agricultura de precisión incluyen el uso de máquinas de soporte vectorial para predecir el peso del ganado de carne antes de la matanza [5], las evaluaciones de aprendizaje automático del secado del suelo para la planificación agrícola [27], la detección temprana y clasificación de enfermedades [93], el desarrollo de técnicas de *soft computing* en ingeniería agrícola y biológica, especialmente en el suelo y el agua, esto con fines de gestión de la siembra y el soporte a la toma de decisiones en agricultura de precisión [50], métodos de predicción para plagas de cultivos utilizando métodos de aprendizaje automático [58].

Además, se han propuesto herramientas de software para estos fines. Por ejemplo, Camargo, Molina, Cadena-Torres y col. [22] presentaron un sistema de información para la evaluación de trastornos de plantas, *Isacrodi* y mostraron que los expertos humanos logran una evaluación más precisa que el clasificador Isacrodi, particularmente cuando se les proporcionan muestras del cultivo afectado. Sin embargo, en aquellos casos en que no se dispone de dicha experiencia, los autores sugieren que Isacrodi aún proporciona información valiosa para apoyar a los agricultores. Isacrodi incluye 15 trastornos de cultivos y el proceso de predicción se basa en máquinas de soporte vectorial para múltiples clases.

Huang, Lan, Thomson y col. [50] resumen en su estudio el desarrollo de técnicas de *soft computing* en agricultura e ingeniería biológica, especialmente en el contexto del suelo y el agua para el manejo de cultivos y apoyo a la toma de decisiones en agricultura de precisión. Aunque ellos presentan varias técnicas como lógica difusa, redes neuronales artificiales, algoritmos genéticos, inferencia bayesiana y árboles de decisión, no presentan los resultados cuantitativos de cada trabajo, sino que se enfocan en presentar las ideas principales.

De manera similar, Kim, Yoo, Gu y col. [58] hacen un estudio de técnicas para el pronóstico de plagas en cultivos utilizando métodos de aprendizaje automático, incluida la regresión.

Glezakos, Moschopoulou, Tsiligiridis y col. [41] utilizaron algoritmos genéticos (GA) y redes neuronales artificiales (ANN) para identificar un virus del tabaco (TRV) y un virus del pepino (CGMMV). El método fue probado contra algunos de los clasificadores más utilizados en el aprendizaje automático (clasificadores de Bayes, árboles de decisión y k vecinos más cercanos) vía validación cruzada. Los resultados mostraron su aplicabilidad a este tipo de problemas. Estos autores no probaron sus métodos en la Sigatoka negra y

en lugar de hacer regresión como en el presente trabajo, ellos tomaron el enfoque de la clasificación.

En un contexto más amplio que el campo agrícola, el aprendizaje automático también ha alcanzado áreas de conocimiento como la ecología. Al respecto se puede destacar a Humphries, Magness y Huettmann [51], quienes proponen a los ecólogos el uso del aprendizaje automático en tres campos: 1) exploración de datos para obtener conocimiento del sistema y generar nuevas hipótesis, 2) predecir patrones ecológicos en el espacio y el tiempo, y 3) reconocimiento de patrones para muestreo ecológico.

Por completitud del contexto, aunque en la presente investigación no se privilegia el uso de imágenes, seguidamente se resaltan los siguientes trabajos en esa línea de investigación. Mayuri y Vani Priya [73] presentan un estudio de metodologías aplicadas al procesamiento de imágenes y cómo los enfoques del aprendizaje automático aumentan la productividad en diversos cultivos, considerando las siguientes medidas: detección temprana, reconocimiento de enfermedades en cultivos, métodos de diagnóstico y métodos de selección de cultivos para la predicción del rendimiento.

Lobet [64] presenta una aplicación denominada Plantix, la cual, utilizando procesamiento de imágenes, ayuda a los agricultores en el reconocimiento de enfermedades. A diferencia de la presente propuesta, esta herramienta identifica la enfermedad, pero no aporta en la determinación del progreso de la misma.

Finalmente, Singha y Misrab [98] proponen un algoritmo para la segmentación de imágenes en la detección y clasificación automática de enfermedades de las hojas de las plantas, a la vez, enumeran diferentes técnicas de clasificación de enfermedades que se pueden utilizar para este mismo objetivo.

Respecto a los procesos biológicos que se utilizan como casos de estudio en esta investigación, vale resaltar los siguientes trabajos.

Romero Calderón [92] se basó en modelos de regresión usando un procedimiento paso a paso para pronosticar los períodos de incubación y latencia de la Sigatoka negra. Romero recolectó datos ambientales de dos fincas diferentes en Costa Rica, entre diciembre de 1993 y agosto de 1995. Los modelos de predicción alcanzaron un R^2 de 69 % a 78 % en los datos observados para los periodos de incubación y latencia de la enfermedad, respectivamente; sin embargo, la validación cruzada en los conjuntos de datos independientes fallaron. A diferencia del trabajo de Romero Calderón, quien seleccionó las variables a incluir en el modelo basado solamente en el criterio experto, el presente trabajo tiene el enfoque de aprendizaje automático, donde la estrategia misma colabora en la determinación de las variables a incluir en el modelo e incluso la periodicidad de las mismas es aprendida.

Bendini, Moraes, S. Silva y col. [14] presentan un estudio sobre el análisis de riesgo de aparición de Sigatoka negra basado en modelos polinomiales. Desarrollaron un estudio de caso en una plantación comercial de banano ubicada en Jacupiranga, Brasil, la cual fue monitoreada semanalmente desde febrero hasta diciembre, ambos de año 2005. Los datos

incluyeron observaciones semanales del estado de evolución de la Sigatoka negra, series de tiempo de datos meteorológicos y datos de sensado remoto. Ellos obtuvieron un modelo para estimar la evolución de la enfermedad a partir de imágenes de satélite. Este modelo relaciona los niveles de gris (NC) de la banda 2 de las imágenes satelitales del Landsat-5, con el estado de progreso de la enfermedad. Los autores indican que alcanzaron un R^2 de 90 %. Estos autores utilizan imágenes para mejorar la predicción, predeterminan las variables climatológicas a usar, los periodos, y las imágenes pasan por un proceso de pre-procesamiento no automático, sino con criterio experto, lo cual difiere de la presente propuesta que, utilizando aprendizaje automático, colabora con el experto del dominio en la selección de dichas variables y no exige contar con imágenes para iniciar con la aplicación de la estrategia.

Con respecto a la aplicación del aprendizaje automático al cultivo del café y las enfermedades que lo afectan, se han realizado varias investigaciones, entre otras: multi-clasificador para la detección de la roya del café en cosechas colombianas [29], gráficos de patrones como representación de reglas extraídas de árboles de decisión para la detección de la roya del café [61]. En cuanto a la detección, los trabajos se centran en el procesamiento de imágenes. Por ejemplo se pueden citar los trabajos de Lasso, Thamada, Alves y col. [61], Triantakonstantis y Barr [108] y Singha y Misrab [98]; los cuales difieren de la presente propuesta en que su objetivo es detectar un grado de avance de la enfermedad por medio de imágenes, más que la predicción de la incidencia. Por otro lado, respecto a la predicción del avance de la enfermedad, tal y como se hace en el presente estudio, se encuentran aportes como los de: Perez-Ariza, Nicholson y Flores [87], Kim, Yoo, Gu y col. [58], Ahamed, Mahmood, Hossain y col. [1], Thamada, Rodrigues y Meira [105] y Luaces, Rodrigues, Alves Meira y col. [66]; los cuales pretenden un objetivo más cercano al de este trabajo, pues se centran en la predicción de la incidencia de la roya. A manera de ejemplo, Luaces, Rodrigues, Alves Meira y col. [66] utilizan técnicas de regresión y clasificación y proponen un sistema de alarma basado en un umbral definido por el equipo investigador. Además de las variables meteorológicas y la incidencia de la roya, incluyen en su modelo variables tales como: el espaciado entre plantas y la carga de fruta en la plantación.

En cuanto a la floración del banano, K P y CH [57] presentan una revisión de trabajos en esta área, a diferencia de la presente propuesta, dan énfasis al uso del procesamiento de imágenes y no consideran la floración del banano entre los productos presentados. Además, los trabajos los clasifican en regresión y clasificación, pero no detallan sobre la reducción de atributos, como se hace en esta investigación. El trabajo de Ahamed, Mahmood, Hossain y col. [1], se centra en procesos de clasificación de la producción de varios productos en Bangadesh. Aunque no incluyen el banano, presentan resultados para varios productos, tales como: arroz, tomate y trigo. Hacia el final del artículo, presentan los $RMSE$ obtenidos, pero no se cuenta con los conjuntos de datos para poder comparar. Además no indican procesos de reducción de atributos ni experimentación sobre el efecto en las métricas de la cantidad de periodos previos y periodos adelante utilizados.

Presentado el estado del arte del campo de estudio, en el siguiente Capítulo se expone en detalle la estrategia propuesta.

Capítulo 3

Estrategia propuesta

La presente propuesta se diferencia respecto a otras opciones en que permite la optimización multiobjetivo, promueve la transferencia del aprendizaje adquirido, contribuye a la selección de atributos, no requiere predecir variables climatológicas y todo esto de una manera esquemática que favorece la repetibilidad del proceso.

La propuesta gira en torno a dos elementos principales: el Repositorio de Conocimiento Aprendido (RCA) y las etapas para llevar a cabo la estrategia.

La estructura del RCA se muestra en la figura 3.1 y las etapas de la estrategia propuesta se esquematizan en la figura 3.2. En las siguientes secciones se procede a detallar ambos elementos.

Nota: Aunado a que en el capítulo 4 se presentan casos de aplicación de la estrategia propuesta, en el apéndice A se muestra una iteración de la estrategia y en los apéndices B, C y D se detallan los resultados de varios experimentos, a lo largo del presente capítulo se incluirán ejemplificaciones en los casos que se considera oportuno.

3.1. Etapa preliminar

La etapa preliminar incluye la delimitación de uso de la propuesta y la comprensión de cómo se estructura el RCA.

3.1.1. Delimitación de uso

Como se ha mencionado a lo largo del documento, el pronóstico en el mundo agrícola, además de ser de utilidad para los agricultores, debe ser encarado tomando en cuenta el tipo y volumen de datos con que se cuenta para servir como insumo en el proceso del aprendizaje automático. Es por lo anterior, que seguidamente se procede a delimitar el campo de aplicación de la presente estrategia:

- Se realiza aprendizaje supervisado de tipo regresión.
- Como criterio de selección se utiliza optimización multiobjetivo, en particular los

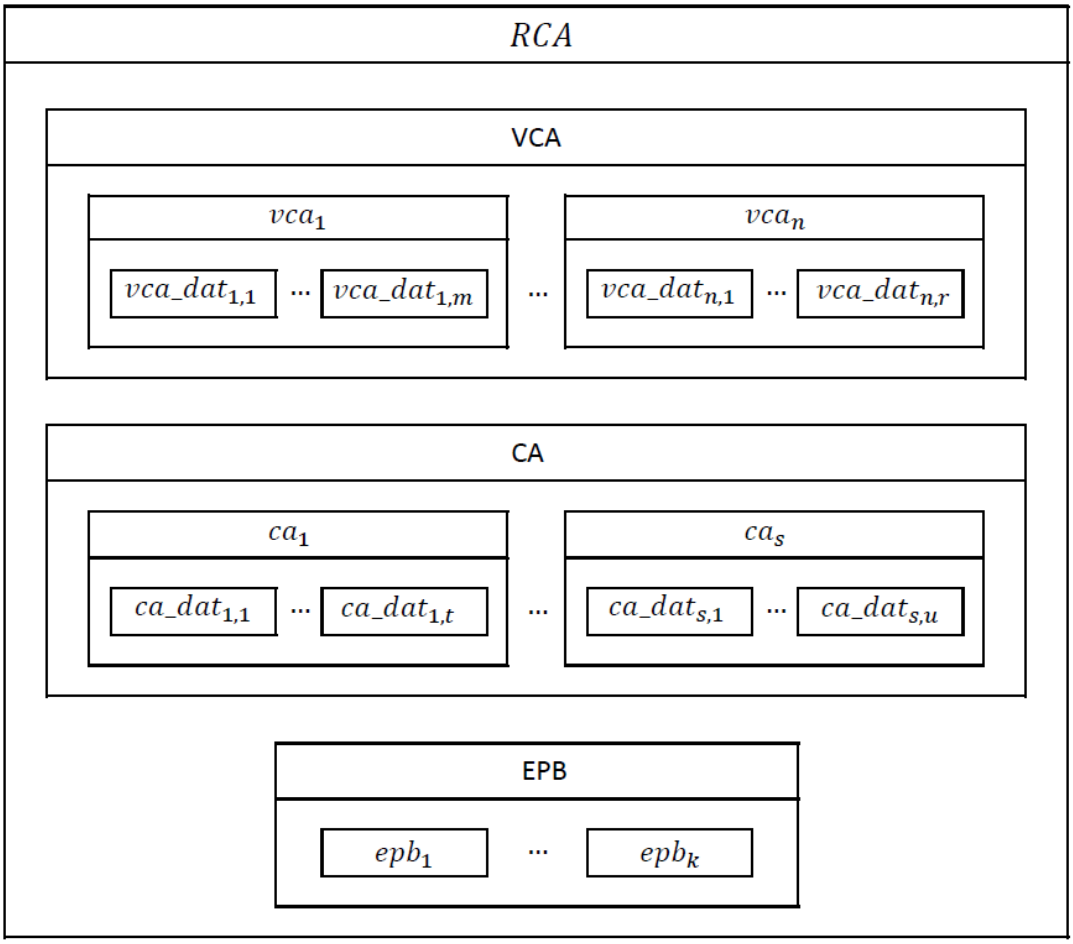


Figura 3.1: Estructura del *RCA*

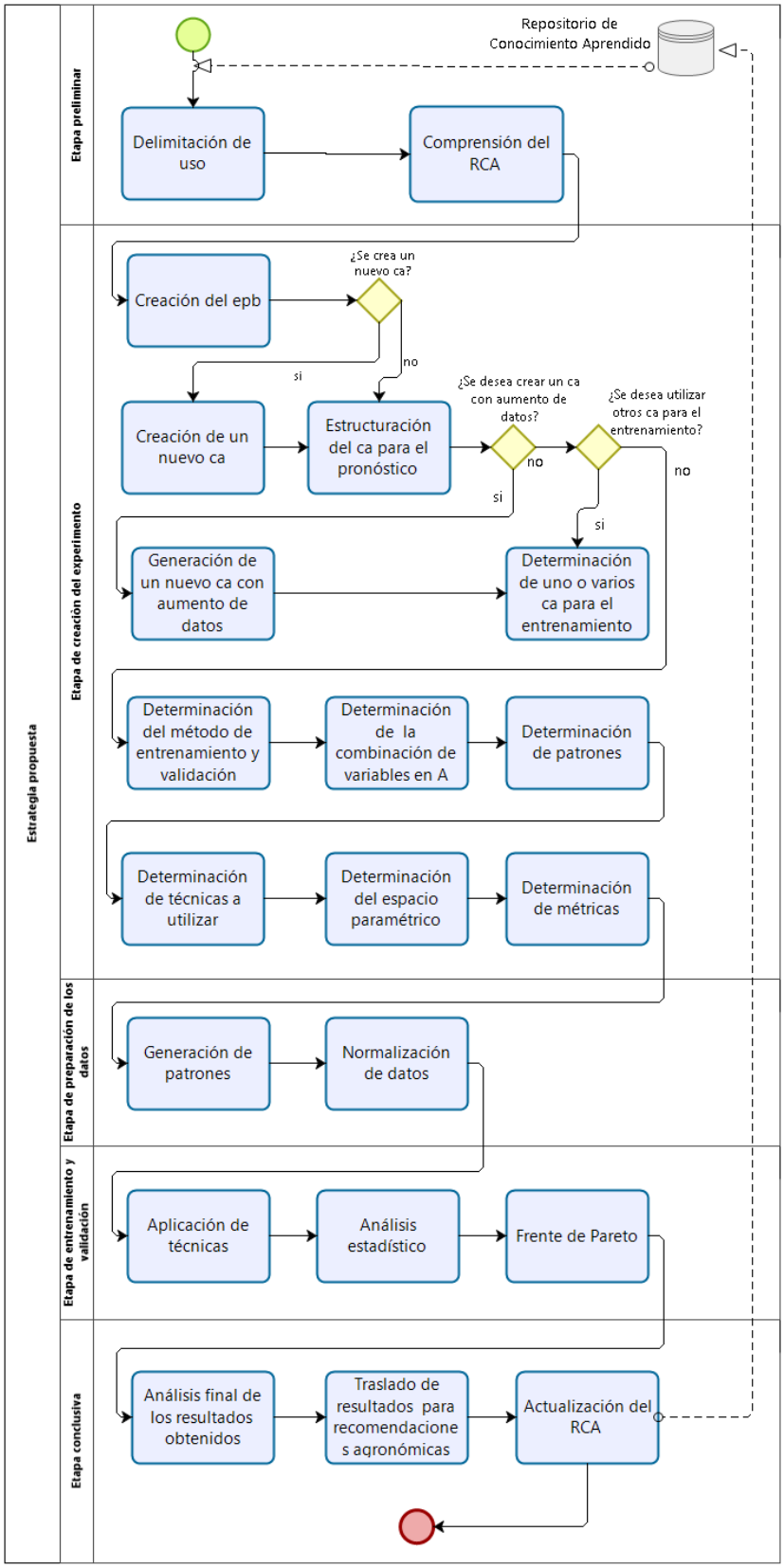


Figura 3.2: Esquema de la estrategia propuesta

elementos que conforman el frente de Pareto al calcular las métricas elegidas.
- La variable a pronosticar representa un proceso biológico en el campo agrícola.
- Aunque como parte del modelo se permiten variables de entrada que representan fenómenos climáticos, no es el objetivo de la presente propuesta pronosticar variables de este tipo.
- Sobre la cantidad de elementos de la variable a pronosticar, si bien no se puede asegurar que el nivel de predicción es óptimo dentro de un rango particular, pues esto depende en gran medida del comportamiento del proceso biológico en particular y de cuánta información para el pronóstico esté contenida en las variables incluidas en el modelo, sí se puede afirmar que la presente propuesta está diseñada para poder iniciar la experimentación cuando la cantidad de elementos de la variable a pronosticar se encuentra en el rango de las decenas, centenas o unos cuantos miles (no impide mayor cantidad de elementos). Esta delimitación de uso responde a un tipo particular de problemática presente en el campo agrícola, en la que es común tener varios años de muestras de un proceso biológico que por su naturaleza toma sentido registrarlo semanal, quincenal o mensualmente, pues el registro con mayor periodicidad no hace un aporte significativo pues el fenómeno no muestra cambios sensibles en frecuencias altas. Por ejemplo, medir el grado de la enfermedad denominada Sigatoka negra cada segundo podría generar hasta 525.600 registros al año, pero el cambio que puede mostrar el proceso biológico de un segundo a otro, no aporta información significativa para el pronóstico, esto aunado al aumento en el costo para medir el proceso biológico a tal frecuencia o la imposibilidad de hacerlo, por ejemplo cuando la toma de un registro dura mucho más tiempo que la frecuencia de toma del dato.
- Las variables a incluir en el modelo pueden ser discretas o continuas. Otro tipo de variables requeriría pre-procesamiento.

3.1.2. Comprensión del RCA

El Repositorio de Conocimiento Aprendido (RCA) será un espacio centralizado en el que se almacena información de los experimentos que se van realizando, de manera que pueda servir para futuras iteraciones de la estrategia.

El RCA se estructura de la siguiente manera:

- VCA: Variable de Conocimiento Aprendido, corresponderá al conjunto de objetos que representan variables que miden algún fenómeno, sea físico o biológico. Cada objeto en el VCA, llamado vca, tendrá la siguiente estructura:
  - id_vca: Identificador único, será una hilera de caracteres.
  - $descripcion$: Texto explicativo del tipo de fenómeno que refleja la variable.
  - $mostrar\_como$: Texto utilizado al mostrar el id_vca.
  - $unidad$: La unidad en que se mide la variable.

- *origenes*: Cada *vca* podrá contener cero o más *vca_dat*. Cada *vca_dat* será un objeto con la siguiente estructura.
  vca_dat:
  - *id_vca_dat*: Identificador único, será una hilera de caracteres.
  - *id_vca*: Identificador del *vca* al que pertenece.
  - *origen*: Descripción de donde fueron tomados los datos.
  - *datos*: Matriz $\in \mathbb{R}^{n\times 2}$, donde para cada fila i, $i \in \{1,2,3,...,n\}$, la columna uno será el valor de la variable y la columna dos será la marca temporal cuando se tomó el valor de la variable.
- $filtro$: Filtro recomendado a aplicar en caso de requerirse para unificar frecuencias. Será una hilera y es determinado por el experto en el dominio. Se consideran las siguientes opciones, pero con criterio experto podrán definirse adicionales:
  - *suma*: Suma de todos los valores en el rango.
  - *promedio*: Media aritmética de los valores en el rango.
  - *mediana*: Es el valor que se ubica en la posición central al ordenar de menor a mayor los datos en el rango, de existir varios valores que cumplen esta característica, se toma el de mayor valor numérico.
  - *moda*: Es el valor que más se repite en el rango a filtrar, de existir varios valores que cumplen esta característica, se toma el de mayor valor numérico de entre ellos.
  - *maximo*: Es el mayor valor de todos los valores en el rango.
  - *minimo*: Es el menor valor de todos los valores en el rango.
- *metodo_imputacion*: En diálogo con los expertos del área se recomienda el método de imputación de faltantes.

- EPB: Experimento de Proceso Biológico, corresponderá al conjunto de objetos que representan un experimento. Cada objeto en el EPB, llamado *epb*, tendrá la siguiente estructura:
  - *id_epb*: Identificador único, será una hilera de caracteres.
  - *descripcion*: Texto explicativo del objetivo del experimento.
  - *variables*: Vector que contendrá los *id_vca* de los *vca* que conforman el experimento y donde el último *id_vca* en el vector corresponderá a la variable a predecir.
  - *ca_estudio*: *id_ca* del *ca* que se estudia en el experimento.
  - C: Vector con todos los *id_ca* de los *ca* utilizados como entrenamiento en el experimento.
  - Pat: Vector que contendrá todos los patrones a experimentar.
  - T: Vector que contendrá las técnicas a aplicar.

- T': Vector que contendrá las técnicas en T que tienen uno o más parámetros y que por tanto requieren que se determine su espacio paramétrico.
- E: Matriz $\in \mathbb{R}^{n \times m}$, para $i \in$ {1,2,3,...,n} y $j \in$ {1,2,3,...,m}, donde $e_{i,j}$ corresponderá al espacio paramétrico de la i-ésima técnica en T', para su j-ésimo parámetro.
- O: Matriz $\in \mathbb{R}^{n \times m}$, para $i \in$ {1,2,3,...,n} y $j \in$ {1,2,3,...,m}, donde $o_{i,j}$ contendrá los valores seleccionados por el proceso de optimización heurística para la i-ésima técnica en T', en su j-ésimo parámetro.
- M: Vector que contendrá las métricas a calcular.
- MEV: Método de entrenamiento y validación. Será una hilera de caracteres.
- S: Objeto que contendrá todas las matrices con patrones generadas.
- R: Matriz $\in \mathbb{R}^{n \times m \times p}$, para $i \in$ {1,2,3,...,n}, $j \in$ {1,2,3,...,m} y $k \in$ {1,2,3,...,p}, que contendrá los resultados obtenidos en el proceso de entrenamiento y validación, y donde $r_{i,j,k}$ corresponderá al resultado de la i-ésima matriz de patrones en S, para la j-ésima técnica en T y para la k-ésima métrica en M.
- AE: Objeto que contendrá los resultados de realizar el análisis de varianza a los resultados en R. Podrán ser documentos de texto, gráficos, tablas de datos, entre otros.
- U: Objeto que contendrá los resultados en R que pertenecen al frente de Pareto determinado en el experimento.
- *observaciones*: Objeto que contendrá las observaciones que el equipo investigador considera deben ser guardadas respecto al experimento. Podrán ser documentos de texto, gráficos, videos, audios, tablas de datos, entre otros.

- CA: Conocimiento para el Aprendizaje, corresponderá al conjunto de objetos que representan datos para el aprendizaje. Cada objeto en el CA, llamado *ca*, tendrá la siguiente estructura:
  - *id_ca*: Identificador único, será una hilera de caracteres.
  - *id_epb*: Identificador del *epb* que le dio origen.
  - *tipo_aumento_datos*: Texto que indicará el tipo de aumento de datos utilizado. Si el *ca* no proviene de aumento de datos, este atributo contendrá una hilera nula.
  - A: Atributos que se incluirán en el *ca*. Se representa como un vector que contendrá los *id_vca* del *vca* que lo conforman. a_i será el i-ésimo atributo en A, para $i \in$ {1,2,3,...m}. Los primeros $m-1$ elementos corresponden a los atributos independientes y a_m corresponde al atributo a predecir
  - N: Serán los atributos para los que desean probar sus combinaciones. N se representará como un vector conformado por un subconjunto propio de los primeros $m-1$ elementos en A (Subconjunto propio de los atributos independientes).

- X: Matriz $\in \mathbb{R}^{n \times m-1}$, para $i \in \{1,2,3,...,n\}$ y $j \in \{1,2,3,...,m-1\}$, donde $x_{i,j}$ corresponderá al i-ésimo vector en su j-ésimo atributo.
- y: Vector columna de datos numéricos con n elementos, en donde y_i contendrá el valor del atributo a_m para la i-ésima fila en X. Atributo independiente.
- *detalles*: Para cada uno de los *id_vca* en el vector A, se tendrá un *ca_dat*, que será un objeto con la siguiente estructura.
  ca_dat:
  - *ca_id_vca*: Identificador único.
  - *periodicidad*: Intervalo temporal entre dato y dato.
  - *marca_temporal_inicio*: Indicación cronológica del dato más antiguo utilizado.
  - *marca_temporal_fin*: Indicación cronológica del dato más reciente utilizado.
  - *maximo*: Valor máximo permitido. Determinado por el experto en el dominio, el cual sirve para detectar valores atípicos.
  - *minimo*: Valor mínimo permitido. Determinado por el experto en el dominio, el cual sirve para detectar valores atípicos.

3.2. Etapa de creación del experimento

En conjunto con los expertos del dominio, se crea el experimento para aplicar la estrategia al proceso biológico en estudio. En las siguientes secciones se detalla esta etapa.

3.2.1. Creación del *epb*

Este proceso se compone de varios pasos:

1. Se incluye en el EPB un nuevo *epb* para el actual experimento.
2. Los atributos: *id_epb* y *descripcion*, se completan como se indica en la sección 3.1.2.
3. Los atributos: C, Pat, T, T', E, O, M, MEV, S, R, AE, U y *observaciones* quedan pendientes de completar más adelante en esta estrategia.
4. En cuanto a los atributos *ca_estudio* y *variables*, se debe revisar en CA si existe un *ca* con un *ca_dat* que coincida con el requerido en este experimento; de ser así, se pasa a la sección 3.2.3; de lo contrario, se continúa con la sección 3.2.2.

3.2.2. Creación de un nuevo *ca*

El proceso a seguir es el siguiente:

- Primero se debe determinar si en VCA existen los *vca* con los *vca_dat* que se requieren como base para el experimento en cuestión. De faltar alguno, se incluyen en VCA los *vca* y *vca_dat* necesarios; los cuales deben pasar por un proceso previo de limpieza de datos para velar porque cumplan lo que se indica en la sección 3.1.2.

- Contando ya con los *vca* y *vca_dat* requeridos, se procede a crear un nuevo *ca* con la estructura indicada en la sección 3.1.2, considerando:

  - *id_ca*: será un identificador único.
  - *id_epb*: será el identificador del *epb* en desarrollo.
  - *tipo_aumento_datos*: será una hilera de caracteres vacía (pues en este punto no se trata de aumento de datos).
  - A: será un vector con m elementos, donde cada $a_i \in A$, para $i \in \{1,2,3,...,m\}$ será un *id_vca* de los *vca* que conforman el *ca*. Los primeros $m-1$ elementos corresponden a los atributos independientes y a_m corresponde al atributo a predecir.
  - N: será un vector que contiene un subconjunto propio de los primeros $m-1$ elementos en A. N contendrá los *id_vca* de las variables que se desea probar su combinatoria. En este punto queda pendiente su asignación.
  - Con los datos provenientes de los *vca_dat* seleccionados, se completarán los siguientes atributos, para lo cual se puede usar la totalidad de datos del *vca_dat*, un subconjunto de los datos o aplicar algún filtro a los datos para cambiar la periodicidad:
    - X: será una matriz de datos numéricos con n filas por $m-1$ columnas. Corresponde a los atributos independientes.
    - y: será un vector de datos numéricos con n elementos (por tanto de la forma n filas por 1 columna), es el atributo dependiente o a predecir.
    - *detalles*: Para cada uno de los *id_vca* en el vector A, se tendrá un *ca_dat*, que contendrá:
      - *ca_id_vca*: Identificador único.
      - *periodicidad*: Intervalo temporal entre dato y dato.
      - *marca_temporal_inicio*: Indicación cronológica del dato más antiguo utilizado.
      - *marca_temporal_fin*: Indicación cronológica del dato más reciente utilizado.
      - *maximo*: Valor máximo permitido. Determinado por el experto en el dominio.
      - *minimo*: Valor mínimo permitido. Determinado por el experto en el dominio.

3.2.3. Estructuración del *ca* para el pronóstico

Teniendo ya el *ca*, se completan los atributos *ca_estudio* y *variables* del *epb* en proceso, como se indica en la sección 3.1.2.

En este punto se realizará un análisis estadístico de tipo descriptivo del contenido de X e y.

Además, D será la concatenación vertical entre X e y. Por tanto, D será una matriz de n filas por m columnas.

Ejemplificación:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,m-1} \\ x_{2,1} & x_{2,2} & \dots & x_{2,m-1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,m-1} \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$D = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,m-1} & y_1 \\ x_{2,1} & x_{2,2} & \dots & x_{2,m-1} & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,m-1} & y_n \end{bmatrix}$$

3.2.4. Generación de un nuevo *ca* con aumento de datos

Si se va a crear un nuevo *ca* con aumento de datos, se procede como se describe a continuación, en caso contrario, se continúa con la sección 3.2.5.

Se utilizará el coeficiente de variación multidimensional (ver sección 2.5.1) para producir un nuevo *ca* a partir de *ca_estudio*. Los pasos a seguir son:

- *cvm* será el coeficiente de variación multidimensional de D según lo explicado en la sección 2.5.1. Por lo que: $cvm \in \mathbb{R}$.
- q será el resultado de multiplicar *cvm* por 100 y tomar la parte entera. Por lo que: $q \in \mathbb{N}$.
- Según se indicó en la sección 3.2.3, D es una matriz $\in \mathbb{R}^{n \times m}$, y a partir de ella se calculará la variación que hay entre cada fila de D y su fila precedente, esto a partir de la segunda fila, pues la primera fila no tiene fila precedente.
  variacion será una matriz $\in \mathbb{R}^{n-1 \times m}$, donde para cada fila (vector) $variacion_k$, $k \in \{1,2,3,..,n-1\}$ se tendrá que: $variacion_k = D_{k+1} - D_k$
- $aleatorio(a,b,inicio,fin)$ será una función que retorna una matriz $\in \mathbb{R}^{a \times b}$ y donde cada elemento de la matriz es un número aleatorio en el rango $[inicio,fin]$.
- El operador $\circ$ corresponde al producto Hadamard entre matrices de igual dimensión (conocido también como multiplicación matricial por elementos).
- Se generará un nuevo *ca* de la siguiente manera:
  DF será una matriz $\in \mathbb{R}^{n-1 \times m}$ que contendrá los mismos valores de las primeras

$n-1$ filas de D.
$DFcs$ será una matriz $\in \mathbb{R}^{(n+(q*(n-1)))\times m}$ y consiste en la concatenación vertical de j matrices, para $j \in \{1,2,3,...,(q+1)\}$, tal que:

- Si $j = 1$, D.
- Si $j \neq 1$, $DF + (variacion \circ aleatorio(n-1,m,0,1))$.

- Finalmente, a partir de $DFcs$ se creará un nuevo *ca* y se incorporará a CA con un *id_ca* diferente al *ca_estudio*. El atributo *tipo_aumento_datos* contendrá la hilera de caracteres "-CS", los atributos A y N serán los mismos que del experimento en proceso, X será una matriz de datos numéricos con todas las filas de $DFcs$ y las primeras $m-1$ columnas de $DFcs$, e y será un vector de datos numéricos con la última columna de $DFcs$. Los atributos *detalles* e *id_epb* serán iguales a los de *ca_estudio*.

Ejemplificación:

$cvm \in \mathbb{R}$

$q \in \mathbb{N}$

$D \in \mathbb{R}^{n\times m}$

$$D = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,m} \\ a_{2,1} & a_{2,2} & \dots & a_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,m} \end{bmatrix}$$

$variacion \in \mathbb{R}^{n-1\times m}$

$$variacion = \begin{bmatrix} - & (D_{2,:} - D_{1,:}) & - \\ - & (D_{3,:} - D_{2,:}) & - \\ & \vdots & \\ - & (D_{n,:} - D_{n-1,:}) & - \end{bmatrix}$$

$aleatorio(a,b,inicio,fin) \in \mathbb{R}^{a\times b}$

$e_{i,j}$ será un número aleatorio en el rango $[inicio,fin]$

$$aleatorio\ retorna = \begin{bmatrix} e_{1,1} & e_{1,2} & \dots & e_{1,b} \\ e_{2,1} & e_{2,2} & \dots & e_{2,b} \\ \vdots & \vdots & \vdots & \vdots \\ e_{a,1} & e_{a,2} & \dots & e_{a,b} \end{bmatrix}$$

$DF \in \mathbb{R}^{n-1\times m}$

$$DF = \begin{bmatrix} - & D_{1,:} & - \\ - & D_{2,:} & - \\ & \vdots & \\ - & D_{n-1,:} & - \end{bmatrix}$$

$DFcs \in \mathbb{R}^{(n+(q*(n-1)))\times m}$

$DFcs$ es la concatenación vertical de:

Para $j \in \{1,2,3,...,(q+1)\}$

$$\begin{cases} D & si \quad j = 1 \\ DF + (variacion \circ aleatorio(n-1,m,0,1)) & si \quad j \neq 1 \end{cases}$$

$$DFcs = \begin{bmatrix} D \\ DF + (variacion \circ aleatorio(n-1,m,0,1)) \\ \vdots \\ DF + (variacion \circ aleatorio(n-1,m,0,1)) \end{bmatrix}$$

3.2.5. Determinación de uno o varios *ca* para el entrenamiento

Si lo que se desea es hacer validación cruzada en el mismo *ca*, se continúa con la sección 3.2.6. En caso contrario, en este punto de la estrategia se tendrán dos opciones:

1. Determinación de similitud con otros *ca*: se busca en CA, otros *ca* que compartan el mismo dominio para aplicar t-SNE (ver sección 2.5.2), en cuyo caso se debe considerar:

   - Es recomendable que se utilicen los elementos de las serie que posean la misma marca temporal (es decir, alinear las series según las fechas).
   - Como los *ca* no siempre incluyen las mismas variables, se podrán seleccionar solo las variables que están en todos los *ca* seleccionados, o seleccionar sólo las variables que están presentes en todos los elementos del frente de Pareto de *ca_estudio*, y una tercera opción podrá ser una selección de variables según criterio de los expertos del dominio.
   - *Selca* será un vector conformado por cada uno de los *id_ca* de los *ca* seleccionado para procesar. Además, *nSelca* será la cantidad de elementos en *Selca*.
   - $DtSNE$ será una matriz $\in \mathbb{R}^{r \times s}$, donde r corresponde al total de muestras seleccionadas y s al total de atributos seleccionados. $DtSNE$ consiste en la concatenación vertical de los vectores fila en X e y, correspondientes a los atributos y series seleccionadas en los pasos anteriores. $DtSNE$ será normalizada.
   - *Eca* será un vector columna con r elementos, correspondientes al *id_ca* de cada una de las r filas en $DtSNE$. El objetivo es recordar a qué *ca* corresponde cada fila en $DtSNE$.
   - La implementación de t-SNE que se utiliza, propuesta en [86], tiene como salida para cada grupo de muestras en estudio: 1) la divergencia-KL obtenida, 2) el número de iteraciones realizadas y 3) una matriz $\in \mathbb{R}^{a \times b}$, donde a corresponde al número de muestras y b a la dimensión del espacio embebido, el cual es indicado por el parámetro denominado: número de componentes. Cada fila de esta matriz corresponde al vector en el espacio embebido de la muestra respectiva. Para efectos del resto de la sección, esta matriz se llamará $MtSNE$.
   - En [112] se explica que la gran variabilidad paramétrica de la técnica t-SNE podría llegar a producir valores bajos de divergencia-KL entre conjuntos de

datos que no son similares; esto debido a la terminación temprana de la técnica ocasionado por alcanzar un umbral predefinido. Por lo indicado anteriormente, no es suficiente utilizar la divergencia-KL entre pares de *ca* como criterio de selección de similitud, por lo que en la presente estrategia se proponen los siguientes criterios:

- Analizar el gráfico en tres dimensiones ($3D$): Aplicar t-SNE a $DtSNE$ con tres como dimensión del espacio embebido. Luego, utilizando Eca y la matriz $MtSNE$, graficar en tres dimensiones con el fin de poder ver si existen sobre-posiciones de elementos de diferentes *ca*, esto por cuanto de graficarse en dos dimensiones, este detalle no se podría observar.
- Calcular un valor que represente cada *ca* y su ubicación en el espacio en dos dimensiones ($2D$) y un indicador de la distancia entre cada par de *ca*, tal y como se describe a continuación:
  - La implementación utilizada [86], retorna una $MtSNE$ con valores centrados en cero, por lo cual incluye valores positivos y negativos. Para el cálculo propuesto, se desplazan todos los valores de manera que sean positivos, y no se afecten las restas que se realizarán más adelante.
  - $MtSNE_{j_{min}}$ será el valor absoluto del menor valor presente en la j-ésima columna de $MtSNE$, $j \in \{1,2,3,..,s\}$. $MtSNE_{j_{min}}$ será un escalar.
  - Para $j \in \{1,2,3,..,s\}$, se le sumará a cada elemento en el vector columna $MtSNE_j$ el valor $MtSNE_{j_{min}}$.
  - Se procederá a obtener la norma euclidiana para cada una de las matrices conformadas por los vectores fila de un mismo *id_vca* en $MtSNE$:
    - Para $k \in \{1,2,3,..,nSelca\}$ tenemos que:
      nca_k será la norma euclidiana de la matriz conformada por todos los vectores fila en $MtSNE$ que corresponden al k-ésimo *id_ca* en $Selca$.
    - Recordando que Eca indica a qué *ca* pertenece cada fila en $MtSNE$.
  - $Comb$ será la combinatoria sin repetición de 2 elementos entre los $nSelca$ diferentes *ca* en estudio. Por tanto:
    $Comb = \binom{nSelca}{2}$
  - Denotamos a $nComb$ como la cantidad de elementos en $Comb$.
  - Se procederá a obtener la diferencia en valor absoluto entre cada par de *ca* respecto a su norma euclidiana:
    - Para $t \in \{1,2,3,..,nComb\}$
      $Comb_t$ tendrá dos elementos $Comb_{t_1}$ y $Comb_{t_2}$, correspondientes a dos *id_ca* de la t-ésima combinación.
    - $dnca_{Comb_{t_1},Comb_{t_2}} = \left|nca_{Comb_{t_1}} - nca_{Comb_{t_2}}\right|$.
  - Finalmente se tendrá el siguiente indicador entre cada par de *ca*:
    - Para $t \in \{1,2,3,..,nComb\}$
    - $pnca_{Comb_{t_1},Comb_{t_2}} = \frac{dnca_{Comb_{t_1},Comb_{t_2}}}{nca_{Comb_{t_1}} + nca_{Comb_{t_2}}}$

- En conjunto con los expertos del dominio se decide cuáles *ca* pueden ser utilizados para probar si su inclusión en el entrenamiento mejora o no el nivel de pronóstico. Considerando los siguientes criterios:
  - Preferir los valores más bajos de *pnca* entre el *ca_estudio* y otro *ca*.
  - Que en el gráfico $3D$, las marcas que representan a los *ca* seleccionados en el punto anterior, estén espacialmente cercanos.

2. Selección específica e intencional de uno o varios *ca* por parte del equipo investigador.

Nota sobre la configuración paramétrica: Aunque los gráficos de t-SNE son utilizados para visualizar datos de alta dimensión, no siempre son de fácil interpretación, particularmente debido a los cambios en los resultados producto de la variación en la configuración paramétrica [112]. Los parámetros a considerar serán [86]:

1. *Perplejidad:* La perplejidad está relacionada con el número de vecinos más cercanos que se utiliza.
2. *Número de componentes:* Se refiere a la dimensión del espacio embebido.
3. *Mínimo de la norma del gradiente:* Si la norma del gradiente está por debajo de este umbral, la optimización se detiene.
4. *Número de iteraciones:* Se refiere al número máximo de iteraciones para la optimización.
5. *Exageración temprana:* Controla qué tan cercanos están los grupos originales en el espacio embebido y cuánto espacio hay entre ellos.
6. *Tasa de aprendizaje:* Incremento del aprendizaje entre iteraciones.
7. *Métrica:* Se refiere a la métrica a utilizar para calcular la distancia entre las instancias del vector de características.

En este punto se actualiza el atributo C del *epb* en proceso. C contendrá los *id_vca* de los *ca* seleccionados para ser incluidos en el entrenamiento. En caso de no seleccionarse ningún *ca* para el entrenamiento, se tendrá que $C = []$.

Ejemplificación:

Para efectos del ejemplo, se tendrán tres *ca*, a saber, ca_1, ca_2 y ca_3. En cuanto a los atributos, se considerarán cuatro: id_vca_1, id_vca_2, id_vca_3 e id_vca_4. Finalmente, para cada *ca* se tendrán tres series, por tanto nueve muestras (3×3).

$Selca$: [id_ca_1, id_ca_2, id_vca_3]

$nSelca$: 3

$DtSNE \in \mathbb{R}^{9\times 2}$

$$DtSNE = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ a_{9,1} & a_{9,2} & a_{9,3} & a_{9,4} \end{bmatrix}$$

$$Eca = \begin{bmatrix} id\_vca_1 \\ id\_vca_1 \\ id\_vca_1 \\ id\_vca_2 \\ id\_vca_2 \\ id\_vca_2 \\ id\_vca_3 \\ id\_vca_3 \\ id\_vca_3 \end{bmatrix}$$

En este punto se aplica tSNE y se obtiene $MtSNE$. Se muestra el caso cuando el parámetro número de componentes se establece en dos.

$$MtSNE = \begin{bmatrix} e_{1,1} & e_{1,2} \\ e_{2,1} & e_{2,2} \\ \vdots & \vdots \\ e_{9,1} & e_{9,2} \end{bmatrix}$$

Se obtienen los $MtSNE_{j_{min}}$ y se le suman a su respectivo vector columna $MtSNE_j$.

Se calculan las normas (nca), para lo cual se debe tener presente Eca, el cual indica a qué ca corresponde cada fila en $MtSNE$.

La norma nca_1 se calcula sobre:

$$\begin{bmatrix} e_{1,1} & e_{1,2} \\ e_{2,1} & e_{2,2} \\ e_{3,1} & e_{2,2} \end{bmatrix}$$

La norma nca_2 se calcula sobre:

$$\begin{bmatrix} e_{4,1} & e_{4,2} \\ e_{5,1} & e_{5,2} \\ e_{6,1} & e_{6,2} \end{bmatrix}$$

La norma nca_3 se calcula sobre:

$$\begin{bmatrix} e_{7,1} & e_{7,2} \\ e_{8,1} & e_{8,2} \\ e_{9,1} & e_{9,2} \end{bmatrix}$$

Se obtiene $Comb = \binom{3}{2}$.

$Comb$ = [[id_ca_1, id_ca_2], [id_ca_1, id_ca_3], [id_ca_2, id_ca_3]]

Finalmente se calculan $dnca$ y $pnca$ para cada par de ca en $Comb$.

$dnca_{id\_ca_1,id\_ca_2} = |nca_{id\_ca_1} - nca_{id\_ca_2}|$

$dnca_{id\_ca_1,id\_ca_3} = |nca_{id\_ca_1} - nca_{id\_ca_3}|$

$dnca_{id\_ca_2,id\_ca_3} = |nca_{id\_ca_2} - nca_{id\_ca_3}|$

$$pnca_{id\_ca_1,id\_ca_2} = \frac{dnca_{id\_ca_1,id\_ca_2}}{nca_{id\_ca_1}+nca_{id\_ca_2}}$$
$$pnca_{id\_ca_1,id\_ca_3} = \frac{dnca_{id\_ca_1,id\_ca_3}}{nca_{id\_ca_1}+nca_{id\_ca_3}}$$
$$pnca_{id\_ca_2,id\_ca_3} = \frac{dnca_{id\_ca_2,id\_ca_3}}{nca_{id\_ca_2}+nca_{id\_ca_3}}$$

3.2.6. Determinación del método de entrenamiento y validación

Se pueden utilizar dos opciones para este proceso: la validación cruzada, como se expuso en la sección 2.4.3, o Entrenamiento/Pruebas, que consiste en tomar dos conjuntos independientes de uno o más *ca*, un conjunto para la etapa de entrenamiento (*training set*) y el otro conjunto para la etapa de pruebas (*test set*).

Por las características de los conjuntos de datos a los que va orientada la presente estrategia, se utilizará regularmente la validación cruzada pues no se cuenta con un gran número de observaciones. Ahora bien, cuando el interés es entrenar con uno o varios *ca* ya presentes en CA y validar el *ca_estudio*, se utilizará el método Entrenamiento/Pruebas.

MEV será el método de entrenamiento y validación seleccionado, tomará uno de dos valores, "ValidacionCruzada"ó "Entrenamiento/Pruebas".

3.2.7. Determinación de la combinación de variables en A

Parte de la estrategia es probar varias combinaciones de variables en A, de manera que se puedan sacar conclusiones sobre el nivel de predicción alcanzado con dichas combinaciones. El objetivo es tratar de descubrir si se requieren todas las variables para lograr el mejor resultado en el frente de Pareto, o si con un subconjunto de variables de A se puede obtener el mismo resultado, o incluso hasta mejor (lo cual puede suceder si una o más variables incluidas en el modelo, en lugar de aportar información, generan ruido).

En N quedarán los atributos para los que desean probar sus combinaciones. N será definido en diálogo con los expertos del dominio, u otra forma de determinar N, es utilizar alguna técnica de selección de variables como la propuesta en [20], en donde se utilizan los conceptos de ganancia de información y conjuntos aproximados para su determinación.

3.2.8. Determinación de patrones

Desde el punto de vista de aprendizaje automático, la presente investigación trata con aprendizaje supervisado, específicamente se enfrenta un problema de regresión, por lo que y es un número real.

En estos casos se tiene:

$$y = f(X) \tag{3.1}$$

Se utiliza el concepto de ventanas móviles para generar los diferentes patrones [78].

Para trabajar con X e y desde el punto de vista de aprendizaje supervisado y para superar el hecho que no todos los algoritmos suponen que X e y sean series de tiempo, se procede

de la siguiente manera:

X, tiene n filas y $m-1$ columnas, por lo que $x_{i,j}$ será el valor del j-ésimo atributo de entrada en X en el tiempo i (fila), para $i \in \{1,2,3,...,n\}$ y $j \in \{1,2,3,...,m-1\}$. Además, y_i será la variable dependiente en el tiempo i (fila).

Como se desea determinar cuántos periodos previos de información son requeridos para alcanzar un determinado nivel de pronóstico, éste medido por las métricas seleccionadas, se tiene que p será el número de periodos observados requeridos para realizar el pronóstico y donde p es un número entero y cumple que: $p \geq 1$. Además, a será el número de periodos adelante en el pronóstico y donde a es un número entero y cumple que: $a \geq 1$.

Por tanto, si nos encontramos en el tiempo t (último periodo observado) y se desea pronosticar a periodos adelante, utilizando p periodos previos observados, se tendrá que:

$$y_{t+a} = f(x_{t,1}, x_{t,2}, ..., x_{t,j}, y_t, x_{t-1,1}, x_{t-1,2}, ..., x_{t-1,j}, y_{t-1}, x_{t-p-1,1}, x_{t-p-1,2}, ..., x_{t-p-1,j}, y_{t-p-1}) \tag{3.2}$$

La tabla 3.1 presenta como se estructuran X e y si se desea pronosticar el valor de y un periodo adelante ($a = 1$) y considerando *dos* periodos previos observados ($p = 2$). Por su parte, la tabla 3.2 muestra el caso para $p = 4$ y $a = 2$.

Tabla 3.1: Patrones con $p = 2$ y $a = 1$

id	***Variables independientes***	***Variable dependiente***
1	x_1, y_1, x_2, y_2	y_3
2	x_2, y_2, x_3, y_3	y_4
...	...	...
$n-1$	$x_{n-2}, y_{n-2}, x_{n-1}, y_{n-1}$	y_n

Tabla 3.2: Patrones con $p = 4$ y $a = 2$

id	***Variables independientes***	***Variable dependiente***
1	$x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4$	y_6
2	$x_2, y_2, x_3, y_3, x_4, y_4, x_5, y_5$	y_7
...	...	...
$n-2$	$x_{n-5}, y_{n-5}, x_{n-4}, y_{n-4}, x_{n-3}, y_{n-3}, x_{n-2}, y_{n-2}$	y_n

En este punto del proceso se tienen dos opciones:

1. Que se tenga un conjunto predeterminado de patrones que se deseen probar.
2. Que se desee experimentar con un grupo más amplio de patrones para investigar la mejor configuración:

   En el segundo caso se introduce un elemento más:

   inc será el valor del incremento a considerar al generar los patrones, tanto para los periodos observados requeridos (p), como para los periodos adelante en el pronóstico (a). Y donde *inc* es un número entero y cumple que: $inc \geq 1$.

Pat será un vector que contiene los diferentes patrones obtenidos del siguiente proceso:

2.1. $Prev$ será un vector con los periodos previos deseados, y que contiene a los periodos con índices $[1+(0*inc), 1+(1*inc), 1+(2*inc), 1+(3*inc), ..., (1+(g*inc)]$ mientras que $(1+(g*inc))$ sea $\leq p$.

2.2. $Adel$ será un vector con los periodos adelante deseados, y que contiene a los periodos con índices $[1+(0*inc), 1+(1*inc), 1+(2*inc), 1+(3*inc), ..., (1+(h*inc)]$ mientras que $(1+(h*inc))$ sea $\leq a$.

2.3. Si definimos el producto cartesiano de dos vectores, A y B, denotado como $A \times B$, como el vector que contiene todos los posibles vectores resultantes, tenemos que:
$Pat = Prev \times Adel$
donde $Pat_{i,j}$ denotará el patrón obtenido con i periodos previos y j periodos adelante.

Para ejemplarizar lo anterior, si $p = 4$, $a = 2$ e $inc = 1$, tenemos:

- $Prev$ = [1,2,3,4]
- $Adel$ = [1,2]
- Pat = [[1,1], [1,2], [2,1], [2,2], [3,1], [3,2], [4,1], [4,2]]

Y en el caso de que $p = 9$, $a = 4$ y $inc = 2$, tenemos:

- $Prev$ = [1,3,5,7,9]
- $Adel$ = [1,3]
- Pat = [[1,1], [1,3], [3,1], [3,3], [5,1], [5,3], [7,1], [7,3], [9,1], [9,3]]

En este punto se determinan los valores de p, a e inc a utilizar en la aplicación de la estrategia.

3.2.9. Determinación de técnicas a utilizar

T será un vector que contiene las técnicas a aplicar y t la cantidad de elementos en T.

Aunque no se limita a las siguientes, para efectos de mostrar el uso de la estrategia, se considerarán las siguientes (ver detalle de cada una en la sección 2.2):

- SVR con kernel lineal: SVR/L.
- SVR con kernel gaussiano: SVR/G.
- SVR con kernel sigmoidal: SVR/S.
- SVR con kernel polinomial: SVR/P.
- Regresión de mínimos cuadrados ordinarios: $OLSR$.
- Regresión elasticNet: ENR.

3.2.10. Determinación del espacio paramétrico

Esta determinación depende del estudio de cada una de las técnicas y de las características del dominio de cada uno de los parámetros.

T' será un vector que incluye todas las técnicas en T que tienen uno o más parámetros y que por tanto requieren que se determine su espacio paramétrico. Además t' será la cantidad de elementos en T'.

E será un vector con t' elementos, donde el elemento $e_i \in E$, será un vector que contiene los parámetros a determinar para la i-ésima técnica en T' y donde $e_{i,j}$ corresponde al espacio paramétrico del j-ésimo parámetro de la técnica ${T'}_i$.

3.2.11. Determinación de métricas

Se sugiere utilizar el criterio del frente de Pareto (sección 2.4.2) entre el coeficiente de determinación (R^2) y la raíz del error cuadrado medio ($RMSE$). Esta decisión también está soportada por el amplio uso del primer indicador en las investigaciones en el campo agrícola y el segundo en el campo del aprendizaje automático [32], [53], [101], [102].

M será un vector que contiene las métricas a calcular. Se considerará la siguiente simbología en cuanto a las métricas:

- Coeficiente de determinación: R^2.
- Raíz del error cuadrado medio: $RMSE$.

3.3. Etapa de preparación de los datos

Esta etapa tiene como fin preparar los *ca* para poder ejecutar los experimentos.

En las siguientes secciones se detalla esta etapa.

3.3.1. Generación de patrones

Considerando lo establecido en las Secciones: 3.2.3 (Estructuración del *ca* para el pronóstico), 3.2.7 (Determinación de la combinación de variables en A) y 3.2.8 (Determinación de patrones), se generarán las matrices con patrones a procesar de la siguiente manera:

- S denotará la totalidad de matrices con patrones generadas en esta sección.
- Para cada *ca* a procesar $\in [ca\_estudio, C]$:
  - u será la cantidad de elementos de N en el *ca*.
  - N_j será el j-ésimo atributo en N, para $j \in \{1,2,3,...,u\}$.
  - Recordar que en el *ca*, X es una matriz $\in \mathbb{R}^{n \times m-1}$ e y es un vector columna con n elementos.
  - X_j será el vector columna correspondiente al atributo N_j de X.

- $\binom{u}{i}$ será la combinatoria sin repetición de i elementos entre un total de u elementos.
- Para $i \in \{1,2,3,...,u\}$:
  - $K_i = \binom{u}{i}$
  - k_i será la cantidad de combinaciones en K_i.
  - $K_{i,j}$ representa la j-ésima combinatoria en K_i, para $j \in \{1,2,3,...,k_i\}$.
  - Para cada $K_{i,j}$:
    * $F_{i,j}$ será una matriz $\in \mathbb{R}^{n \times i+1}$, que concatena verticalmente los vectores columna de los atributos en $K_{i,j}$ en X más el vector columna y.
- F representará la totalidad de matrices, $F_{i,j}$, obtenidas en el paso anterior.
- Cuando N no incluya la totalidad de los $m-1$ atributos independientes, se tendrá que agregar a F la matriz D, que como se definió en la sección 3.2.3, consiste en la concatenación vertical de X e y, del *ca*.
- Dados a, p y inc, según se determinó en la sección 3.2.8, se generarán los patrones para cada una de las matrices en F, esto de la manera expuesta en la sección 3.2.8.
- Finalmente, todos los patrones generados en el punto anterior son agregados a S, recibiendo el nombre de matrices con patrones.

Ejemplificación:

Ejemplo con un *ca*, tres *vca*, N incluye los $m-1$ primeros atributos de A, se considerarán cuatro registros para cada *vca*. Por tanto, X del $ca \in \mathbb{R}^{4\times2}$ e y es un vector columna con cuatro elementos. Para los patrones se considerarán dos periodos previos ($p = 2$), un periodo adelante ($a = 1$) y con incrementos de uno en uno ($inc = 1$)

C: [id_ca_1]

A: [id_vca_1, id_vca_2, id_vca_3]

N: [id_vca_1, id_vca_2]

u: 2

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ x_{3,1} & x_{3,2} \\ x_{4,1} & x_{4,2} \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

Se realizan dos combinatorias: $K_1 = \binom{2}{1}$, $K_2 = \binom{2}{2}$, $k_1 = 2$, $k_2 = 1$

K_1: [[id_vca_1], [id_vca_2]]

K_2: [[id_vca_1, id_vca_2]]

K: [[id_vca_1], [id_vca_2], [id_vca_1, id_vca_2]]

$$F_{1,1} : \begin{bmatrix} x_{1,1} & y_1 \\ x_{2,1} & y_2 \\ x_{3,1} & y_3 \\ x_{4,1} & y_4 \end{bmatrix}$$

$$F_{1,2}: \begin{bmatrix} x_{1,2} & y_1 \\ x_{2,2} & y_2 \\ x_{3,2} & y_3 \\ x_{4,2} & y_4 \end{bmatrix}$$

$$F_{2,1}: \begin{bmatrix} x_{1,1} & x_{1,2} & y_1 \\ x_{2,1} & x_{2,2} & y_2 \\ x_{3,1} & x_{3,2} & y_3 \\ x_{4,1} & x_{4,2} & y_4 \end{bmatrix}$$

No se incluye D a F porque N incluye los $m-1$ primeros atributos en A

F contiene a: $F_{1,1}$, $F_{1,2}$, y $F_{2,1}$

Se procede a generar S a partir de F, $p=2$, $a=1$ y $inc=1$ como se indicó en la sección 3.2.8

- Con: $F_{1,1}$
  - Con: $p=2$, $a=1$

    S_1: $\begin{bmatrix} x_{1,1} & y_1 & x_{2,1} & y_2 & y_3 \\ x_{2,1} & y_2 & x_{3,1} & y_3 & y_4 \end{bmatrix}$
  - Con: $p=1$, $a=1$

    S_2: $\begin{bmatrix} x_{1,1} & y_1 & y_2 \\ x_{2,1} & y_2 & y_3 \\ x_{3,1} & y_3 & y_4 \end{bmatrix}$
- Con: $F_{1,2}$
  - Con: $p=2$, $a=1$

    S_3: $\begin{bmatrix} x_{1,2} & y_1 & x_{2,2} & y_2 & y_3 \\ x_{2,2} & y_2 & x_{3,2} & y_3 & y_4 \end{bmatrix}$
  - Con: $p=1$, $a=1$

    S_4: $\begin{bmatrix} x_{1,2} & y_1 \\ x_{2,2} & y_2 \\ x_{3,2} & y_3 \\ x_{4,2} & y_4 \end{bmatrix}$
- Con: $F_{2,1}$
  - Con: $p=2$, $a=1$

    S_5: $\begin{bmatrix} x_{1,1} & x_{1,2} & y_1 & x_{2,1} & x_{2,2} & y_2 & y_3 \\ x_{2,1} & x_{2,2} & y_2 & x_{3,1} & x_{3,2} & y_3 & y_4 \end{bmatrix}$
  - Con: $p=1$, $a=1$

    S_6: $\begin{bmatrix} x_{1,1} & x_{1,2} & y_1 & y_2 \\ x_{2,1} & x_{2,2} & y_2 & y_3 \\ x_{3,1} & x_{3,2} & y_3 & y_4 \end{bmatrix}$

3.3.2. Normalización de datos

- Se recomienda utilizar un método de normalización que sea robusto a valores atípicos, tal como RobustScaler [110]. Para este método, x_j denotará la j-ésima columna en X (que corresponde al j-ésimo atributo). Este método calcula el valor de escala

para cada atributo de la siguiente manera:

$$x_{j_{escalado}} = \frac{x_j - Q1(x_j)}{Q3(x_j) - Q1(x_j)} \tag{3.3}$$

Donde $Q1(x_j)$ corresponde al cuartil uno (25 %) del atributo x_j y $Q3(x_j)$ al tercer cuartil (75 %). Los cuartiles de x_j son calculados considerando el valor del atributo j-ésimo de X en todos los *ca* en proceso (*ca_estudio* y los *ca* en C).

- Se procede a normalizar todas las matrices con patrones en S.
  Si S_k representa la k-ésima matriz con patrones en S, para $k \in \{1,2,3,...,s\}$ y donde s representa la cantidad de elementos en S, Para cada S_k:
  - $S_k \in \mathbb{R}^{n \times m}$.
  - X serán las primeras $m-1$ columnas de S_k e y será el último vector columna de S_k.
  - Se procede a normalizar todo X.
  - En cuanto a y, esta no es normalizada.
  - Se incluye la columna de *bias* con todos los valores igual a 1.

3.4. Etapa de entrenamiento y validación

Esta etapa tiene como fin ejecutar el diseño experimental definido.

3.4.1. Aplicación de técnicas

Se realiza el siguiente proceso:

1. Cuando MEV es ValidacionCruzada, el siguiente paso se aplica con las matrices con patrones en S del *ca_estudio*, pero cuando MEV es Entrenamiento/Pruebas, el proceso siguiente se realiza con las matrices con patrones en S correspondientes a C:
   - Para cada una de las matrices con patrones en S, generadas según la sección 3.3.1:
     - Para cada una de las técnicas en T' (sección 3.2.9), se realiza la optimización heurística de sus parámetros, para lo cual se utiliza la técnica de algoritmos genéticos (en esta investigación se utilizó y configuró la propuesta de [34]). Se analiza el respectivo espacio paramétrico (sección 3.2.10), definido como $E_{i,j}$.
     - En este punto se completará el atributo O del *epb* en experimentación.

2. Luego, si MEV es ValidacionCruzada, se realiza la validación cruzada con las matrices con patrones en S del *ca_estudio*. Pero si MEV es Entrenamiento/Pruebas, se realiza el entrenamiento con las matrices con patrones en S de C y luego se realiza el proceso de pruebas con las matrices con patrones S del *ca_estudio*.

   2.1. Al aplicar las técnicas en T, considere la selección paramétrica en O.

   2.2. Durante la aplicación de las técnicas, se calculan las métricas en M.

   2.3. Con los resultados obtenidos en el proceso de entrenamiento y validación, se generará R, donde $r_{i,j,k}$ corresponde al resultado de la i-ésima matriz con patrones en S, para la j-ésima técnica en T y para la k-ésima métrica en M.

3.4.2. Análisis estadístico

Se realiza el análisis estadístico de los resultados en R según lo indicado en la sección 2.4.4. Se debe establecer el nivel de confianza para las pruebas.

AE será el resultado de realizar el análisis de varianza a los resultados en R.

3.4.3. Frente de Pareto

Para R se calcula el frente de Pareto de las métricas M (ver sección 2.4.2).

U será el conjunto de resultados en R que pertenecen al frente de Pareto determinado.

3.5. Etapa conclusiva

En esta etapa se realizan las conclusiones del experimento, se trasladan los resultados a los expertos del dominio para las decisiones agronómicas que correspondan y se actualiza el RCA con el conocimiento aprendido.

3.5.1. Análisis final de los resultados obtenidos

Los resultados obtenidos en R del *epb* son analizados por el equipo investigador con el fin de revisar la utilidad de sus resultados y la consistencia de los resultados con lo esperado por los expertos del dominio. En este punto se tratan de explicar los resultados numéricos con la realidad biológica estudiada para sacar conclusiones.

Adicionalmente, se podrá realizar un análisis de sobreajuste, subajuste, la comparación de los valores reales versus los predichos y comparar los resultados obtenidos con otras iteraciones de la presente estrategia propuesta.

El atributo *observaciones* del *epb* será modificado para respaldar cualquier aporte presentado en esta sección.

3.5.2. Traslado de resultados para recomendaciones agronómicas

Los resultados obtenidos y analizados son trasladados a los expertos del dominio para proceder con las recomendaciones agronómicas respectivas, las cuales van más allá del alcance de la presente estrategia.

Se podrá actualizar el atributo *observaciones* del *epb* en proceso con cualquier recomendación agronómica que indiquen los expertos del dominio.

3.5.3. Actualización del RCA

El RCA será modificado, ya sea agregando, modificando o eliminando, alguno de sus componentes, a partir del experimento realizado.

Capítulo 4

Resultados y análisis

Presentada la estrategia propuesta en el capítulo anterior, en este capítulo se muestran, a partir de tres casos de estudio, los principales aportes de la presente investigación al estado del arte. Como el fin primordial de los casos de estudio es resaltar algunos aspectos particulares de la estrategia, si se desean conocer en detalle los resultados numéricos de aplicar la estrategia a un proceso biológico en particular, en el Apéndice A se muestra un ejemplo detallado de la aplicación de la estrategia y en los Apéndices B, C y D, se muestra información pormenorizada de los resultados obtenidos en cada caso de estudio.

Los casos de estudio son:

1. Características generales de la estrategia: Se aplica la estrategia a la predicción del estado de evolución de la enfermedad del banano denominada Sigatoka negra.

2. Propuesta en el proceso de aprendizaje por transferencia: Se aplica la estrategia a la predicción del nivel de incidencia de la enfermedad del cafeto denominada roya.

3. Propuesta en el proceso de reducción de atributos: Se aplica la estrategia a la predicción de la floración del banano medida por medio del peso del racimo.

La presentación de cada uno de los casos se dividirá en dos partes: primeramente se resumen los materiales y métodos utilizados, luego se presentan los principales resultados obtenidos, se realiza el análisis de los mismos y a la vez se resaltan y contrastan contra el estado del arte.

4.1. Características generales de la estrategia

El proceso biológico en estudio, enfermedad del banano denominada Sigatoka negra, fue descrito en la subsección 2.3.1 y en el Apéndice B se detallan los resultados obtenidos.

4.1.1. Materiales y métodos

Los datos utilizados fueron adquiridos de dos fincas de investigación de Corbana: *28 Millas*, localizada en Siquirres, y *La Rita*, localizada en Pococí, ambas en la provincia de Limón, Costa Rica. Estas fincas producen banano tipo *Musa* sp. AAA grupo Grande Naine (subgrupo Cavendish). Las variables utilizadas se resumen en la tabla 4.1.

Tabla 4.1: Variables disponibles (Caso: Sigatoka negra)

Símbolo	Descripción	Unidades
$T_{a_{min}}$	Temperatura del aire mínima	[°C]
$\overline{T}_a$	Temperatura del aire promedio	[°C]
$T_{a_{max}}$	Temperatura del aire máxima	[°C]
H_{min}	Humedad relativa mínima	[%]
$\overline{H}$	Humedad relativa promedio	[%]
H_{max}	Humedad relativa máxima	[%]
$\overline{R}$	Radiación solar promedio	[W/m^2]
P	Precipitación acumulada	[mm]
$\overline{W}$	Velocidad del viento promedio	[m/s]
W_{max}	Velocidad del viento máxima	[m/s]
E_s	Estado de evolución	—

Los datos fueron tomados para *La Rita* entre el año 2002 y el 2015, y para *28 Millas* entre el 2003 y el 2015. La variable a pronosticar es el Estado de evolución (E_s), medida semanalmente y aunque las estaciones meteorológicas de Corbana adquieren los datos cada cinco minutos, se utilizan valores semanales en concordancia con la variable de salida. Se utiliza periodicidad semanal, por lo que se cuenta con 676 observaciones para La Rita y 634 para 28 Millas.

Las tablas 4.2 y 4.3 muestran estadísticas descriptivas de los conjuntos de datos utilizados.

Tabla 4.2: Estadísticas del conjunto de datos: La Rita

Métrica	$T_{a_{max}}$	$T_{a_{min}}$	$\overline{T}_a$	$\overline{H}$	H_{min}	H_{max}	$\overline{Sr}$	P	W_{max}	$\overline{W}$	E_e
Cardinalidad	676	676	676	676	676	676	676	676	676	676	676
Promedio	31.47	19.85	24.62	89.51	58.45	99.32	280.75	73.71	3.99	0.54	5464.08
Mediana	31.5	20.1	24.72	90.08	59.0	100.0	286.5	56.65	3.2	0.37	5494.5
Desviación estándar	1.39	1.65	1.05	4.79	9.85	1.16	85.75	65.41	2.68	0.48	1229.76
Valor mínimo	25.1	0.0	20.82	56.34	0.0	96.0	0.0	0.0	0.0	0.0	1042.48
Valor máximo	39.8	23.72	27.09	98.88	81.1	100.0	511.0	376.67	17.7	2.61	9936.79
Rango	14.7	23.72	6.27	42.54	81.1	4.0	511.0	376.67	17.7	2.61	8894.31
Coeficiente de variación	0.04	0.08	0.04	0.05	0.17	0.01	0.31	0.89	0.67	0.89	0.23

El conjunto de técnicas utilizadas (T) fue: $\{SVR/L$, SVR/G, SVR/S, SVR/P, ENR, $OLSR\}$. En cuanto a los patrones, se analizaron hasta 12 semanas previas y 3 semanas adelante, con incrementos de 1 ($p = 12$, $a = 3$, $inc = 1$).

Con el fin de acortar los nombres en las tablas a presentar, se utilizan las siguientes abreviaturas: 28 Millas: **28**, La Rita: **LR**, datos sin aumento de datos: **SS**, datos con

Tabla 4.3: Estadísticas del conjunto de datos: 28 Millas

Métrica	$T_{a_{max}}$	$T_{a_{min}}$	$\overline{T}_a$	$\overline{H}$	H_{min}	H_{max}	$\overline{Sr}$	P	W_{max}	$\overline{W}$	E_e
Cardinalidad	634	634	634	634	634	634	634	634	634	634	634
Promedio	31.7	20.74	25.22	90.82	60.55	99.37	274.48	63.34	8.34	1.31	5142.09
Mediana	31.8	20.92	25.35	91.25	61.0	100.0	281.0	40.4	8.0	1.31	5176.86
Desviación estándar	1.23	1.39	1.0	3.87	7.16	1.73	73.12	73.8	3.04	0.49	612.41
Valor mínimo	26.1	16.3	21.38	62.48	0.0	93.7	0.0	0.0	0.0	0.0	3047.91
Valor máximo	37.9	24.4	27.43	99.76	93.0	100.0	480.0	502.0	20.9	2.99	6808.9
Rango	11.8	8.1	6.05	37.28	93.0	6.3	480.0	502.0	20.9	2.99	3760.99
Coeficiente de variación	0.04	0.07	0.04	0.04	0.12	0.02	0.27	1.17	0.36	0.37	0.12

aumento de datos: **CS**, conjunto de datos del entrenamiento: **Tr**, y conjunto de datos de Prueba: **Te**.

La estrategia propuesta fue aplicada en seis iteraciones (*Experimentos*), a saber:

1. *ten-fold-cross-validation* con los datos de 28 Millas (**Validación cruzada en 28 Millas**).
2. *ten-fold-cross-validation* con los datos de La Rita (**Validación cruzada en La Rita**).
3. Entrenamiento con 28 Millas y pruebas con La Rita (**Tr:28-SS / Te:LR-SS**), utilizando las configuraciones presentes en el frente de Pareto de 28 Millas.
4. Entrenamiento con La Rita y pruebas con 28 Millas (**Tr:LR-SS / Te:28-SS**), utilizando las configuraciones presentes en el frente de Pareto de La Rita.
5. Entrenamiento con 28 Millas luego de aplicarle el aumento de datos y pruebas con La Rita sin aumento de datos (**Tr:28-CS / Te:LR-SS**), utilizando las configuraciones presentes en el Frente de Pareto de 28 Millas.
6. Entrenamiento con La Rita luego de aplicarle el aumento de datos y pruebas con 28 Millas sin aumento de datos (**Tr:LR-CS / Te:28-SS**), utilizando las configuraciones presentes en el frente de Pareto de La Rita.

4.1.2. Resultados y análisis

Ejecutados los experimentos, se obtuvieron 3456 resultados en la validación cruzada para La Rita, y la misma cantidad para 28 Millas. El frente de Pareto de 28 Millas lo componen 5 configuraciones, y el de La Rita 9 configuraciones. A partir de estas configuraciones se realizaron los experimentos de Entrenamiento/Pruebas, como se indicó en la subsección 4.1.1. La tabla 4.4 muestra los primeros 7 resultados según el frente de Pareto para 28 Millas y la tabla 4.5 los 12 primeros resultados para La Rita. Ambas tablas están ordenadas según el $RMSE$ (de menor a mayor) y para los resultados que formen parte del frente de Pareto respectivo.

En los experimentos relativos a la validación cruzada, los valores de R^2 para *La Rita* fueron superiores en comparación con los obtenidos para *28 Millas*. En cuanto al $RMSE$, esta última finca tuvo valores menores que la primera. Este comportamiento del $RMSE$ es proporcional al valor absoluto de los estados de evolución predichos (E_s), los cuales

Tabla 4.4: Primeros 7 resultados del Frente de Pareto entre el R^2 y $RMSE$ para el estado de evolución de la Sigatoka negra (28 Millas)

Variables	$p \rightarrow a$	**Técnica**	**RMSE**	R^2	**Experimento**
Ta-H-W-P	$10 \rightarrow 1$	ENR	397,69	57,69 %	Validación cruzada
$\overline{T}_a$ $\overline{H}$ P	$7 \rightarrow 1$	SVR/S	397,95	58,61 %	Validación cruzada
	$9 \rightarrow 1$	SVR/S	398,01	59,73 %	Validación cruzada
	$9 \rightarrow 1$	SVR/L	399,2	59,84 %	Validación cruzada
$\overline{T}_a$ $\overline{H}$	$4 \rightarrow 1$	SVR/G	405,46	60,18 %	Validación cruzada
$\overline{H}$	$9 \rightarrow 1$	SVR/G	414,12	59,65 %	Tr:LR-SS / Te:28-SS
P	$4 \rightarrow 1$	SVR/S	418,34	60,54 %	Tr:LR-SS / Te:28-SS

Tabla 4.5: Primeros 12 resultados del Frente de Pareto entre el R^2 y $RMSE$ para el estado de evolución de la Sigatoka negra (La Rita)

Variables	$p \rightarrow a$	**Técnica**	**RMSE**	R^2	**Experimento**
$\overline{H}$	$9 \rightarrow 1$	SVR/G	679,32	68,79 %	Validación cruzada
$\overline{T}_a$ P $\overline{W}$	$6 \rightarrow 1$	ENR	679,5	68,8 %	Validación cruzada
$\overline{T}_a$ $\overline{H}$	$5 \rightarrow 1$	OLSR	681,48	69,18 %	Validación cruzada
$\overline{T}_a$ $\overline{H}$ $\overline{W}$	$5 \rightarrow 1$	SVR/S	682,03	69,31 %	Validación cruzada
$\overline{T}_a$	$6 \rightarrow 1$	SVR/L	682,14	69,69 %	Validación cruzada
P	$4 \rightarrow 1$	SVR/S	682,82	69,99 %	Validación cruzada
$\overline{T}_a$ $\overline{H}$	$4 \rightarrow 1$	SVR/L	685,08	70,41 %	Validación cruzada
P	$4 \rightarrow 1$	SVR/L	685,81	70,72 %	Validación cruzada
$\overline{W}$	$4 \rightarrow 1$	SVR/S	687,61	70,85 %	Validación cruzada
$\overline{T}_a$ $\overline{H}$ P	$9 \rightarrow 1$	SVR/L	710,28	65,82 %	Tr:28-CS / Te:LR-SS
	$7 \rightarrow 1$	SVR/S	711,0	66,08 %	Tr:28-CS / Te:LR-SS
	$9 \rightarrow 1$	SVR/S	724,33	64,15 %	Tr:28-CS / Te:LR-SS

en *La Rita* (debido a las condiciones meteorológicas prevalecientes) fueron generalmente mayores que los observados en *28 Millas*. El coeficiente de determinación R^2 es, de otra manera, menos sensitivo al valor absoluto de la variable bajo estudio. Como referencia, la tabla 4.6 confirma que la variable E_s presenta, para todas la métricas calculadas, valores mayores en La Rita que en 28 Millas. El frente de Pareto para la finca *La Rita* contiene 9 puntos óptimos en el frente de Pareto, mientras que para *28 Millas* hay 5 puntos óptimos en su respectivo frente de Pareto. Para 5 de las 9 las configuraciones en el frente de Pareto de *La Rita*, la variable del promedio de la temperatura del aire ($\overline{T}_a$) está presente . Esta variable está presente también en todas las configuraciones del frente de Pareto para *28 Millas*.

Por otra parte, respecto a los experimentos de Entrenamiento y Pruebas, se tiene que para 28 Millas, si se entrena con los datos de La Rita, se obtienen valores de $RMSE$ y R^2 similares a los de frente de Pareto, pero usando La Rita sin aumento de datos. A diferencia, para La Rita, si se entrena con los datos de 28 Millas, se obtienen valores de R^2 menores y valores de $RMSE$ mayores a los de frente de Pareto, pero en este caso es

Tabla 4.6: Comparación de estadísticas respecto al E_s entre los conjuntos de datos de 28 Millas y La Rita

Métrica	28 Millas	La Rita
Cardinalidad	634	676
Promedio	5142.09	5464.08
Mediana	5176.86	5494.5
Desviación estándar	612.41	1229.76
Coeficiente de variación	0.12	0.23

mejor usar los datos de 28 Millas con aumento de datos.

Como es de esperar, en las tablas 4.4 y 4.5, el frente de Pareto está conformado por pronósticos a una semana adelante (dado que existe mucho menos incertidumbre que dos o tres semanas adelante), pero si interesara pronosticar no una, sino dos o tres semanas adelante, los mejores $RMSE$ y R^2 que se obtienen se presentan en la tabla 4.7.

Tabla 4.7: Mejores $RMSE$ y R^2 al pronosticar 2 y 3 semanas adelante (Sigatoka negra)

Lugar	a	$RMSE$	R^2
La Rita	Pronóstico de E_s en 2 semanas	800,93	59,7 %
	Pronóstico de E_s en 3 semanas	866,37	52,98 %
28 Millas	Pronóstico de E_s en 2 semanas	445,71	51,1 %
	Pronóstico de E_s en 3 semanas	457,63	49,07 %

Con fines de comparación, la tabla 4.8 muestra los mejores $RMSE$ y R^2 obtenidos con los mismos conjuntos de datos pero utilizando otras técnicas de predicción. Como se observa, no superan los valores obtenidos en la presente investigación considerando el criterio definido del frente de Pareto entre $RMSE$ y R^2. Esto se explica por el hecho de que estas técnicas requieren una cantidad mayor de datos. Además de estos resultados, Argüello [9] mostró que los resultados obtenidos no logran superar los presentados en la presente investigación debido a que las redes no logran aprender la estructura por la cantidad de datos disponibles.

La tabla 4.9 muestra los mejores $RMSE$ y R^2 que se obtienen al utilizar las mismas técnicas de la presente investigación pero no como aprendizaje supervisado sino como regresión simple. Además de no superar los resultados de este trabajo, se requerirían estimar las variables climáticas para obtener el valor de predicción de semanas adelante, algo que se desea evitar en la presente investigación.

La figura 4.1 muestra el frente de Pareto entre R^2 y $RMSE$ para las fincas: La Rita y 28 Millas. En esta figura están graficados 6912 resultados de la validación cruzada. En rojo se resalta el frente de Pareto.

La figura 4.2 muestra la aplicación de $tSNE$ entre los conjuntos de datos de La Rita y 28 Millas. Se aprecia que no hay un traslape significativo entre ambos, lo cual se vio reflejado en que al entrenar con una de las fincas y predecir con la otra, se alcanzan

Tabla 4.8: Mejores $RMSE$ y R^2 al pronosticar E_s utilizando otras técnicas (Sigatoka negra)

Lugar	Técnica	$RMSE$	R^2
La Rita	Bayesian Ridge	679,31	58,23 %
	Echo State Networks	1040,05	53,66 %
	Dynamic Time Warping	925,3	53,57 %
	Linear Discriminant Analysis	982,36	50,63 %
	Gradient Descent	870,33	49,88 %
	Gradient Boosting Regressor	791,99	40,81 %
28 Millas	Gradient Descent	445,45	52,67 %
	Echo State Networks	575,13	52,22 %
	Dynamic Time Warping	551,74	51,0 %
	Bayesian Ridge	401,51	50,76 %
	Linear Discriminant Analysis	599,37	46,89 %
	Gradient Boosting Regressor	463,43	36,9 %

Tabla 4.9: Mejores $RMSE$ y R^2 al pronosticar E_s utilizando las mismas técnicas de la investigación, pero no tratadas como aprendizaje supervisado sino como una regresión lineal (Sigatoka negra)

Lugar	Técnica	$RMSE$	R^2
La Rita	SVR/P	1024,43	38,66 %
	OLSR	1256,87	29,8 %
	SVR/G	1165,39	27,31 %
	ENR	1205,17	24,98 %
	SVR/L	1195,81	24,26 %
	SVR/S	1180,56	22,72 %
28 Millas	SVR/P	485,66	33,70 %
	OLSR	571,97	28,89 %
	ENR	563,9	24,88 %
	SVR/L	559,92	24,81 %
	SVR/G	559,23	23,76 %
	SVR/S	559,16	23,56 %

métricas alrededor de 60 % al predecir 28 Millas a partir de La Rita y de 65 % al predecir La Rita a partir de 28 Millas (ver tablas 4.4 y 4.5).

Ahora bien, desde el punto de vista del aporte de la estrategia en sí, en los siguientes párrafos se analizarán aspectos a resaltar.

Propuesta esquemática que favorece la repetibilidad del proceso.

Los trabajos de [12] (figura 4.3), [33] (figura 4.4) y [65] (figura 4.5), presentan un esquema detallado de su propuesta. A pesar de ello, la lectura de sus trabajos carece del mismo

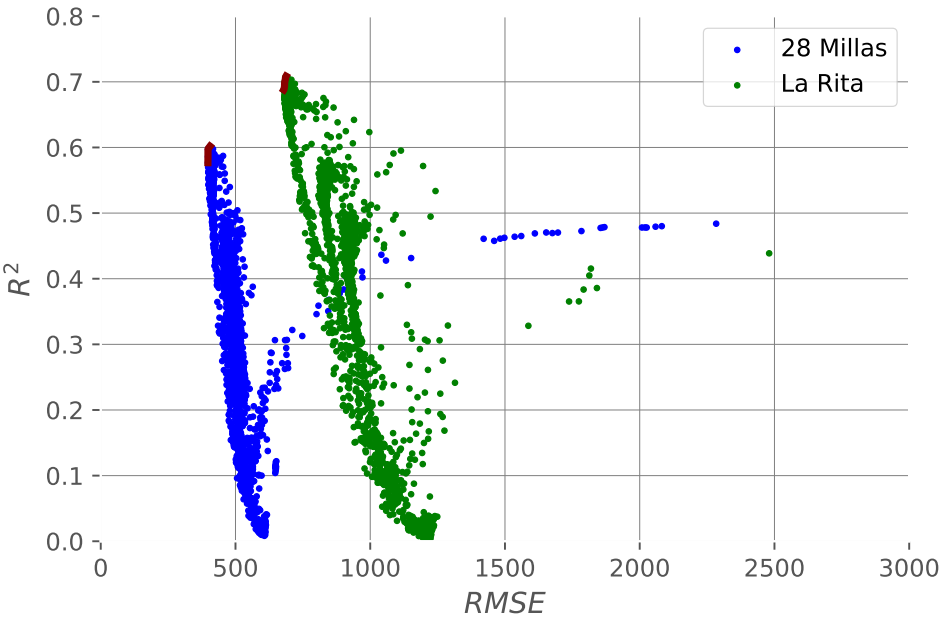


Figura 4.1: Frente de Pareto entre R^2 y $RMSE$ para las fincas: La Rita y 28 Millas, 6912 resultados de la validación cruzada.

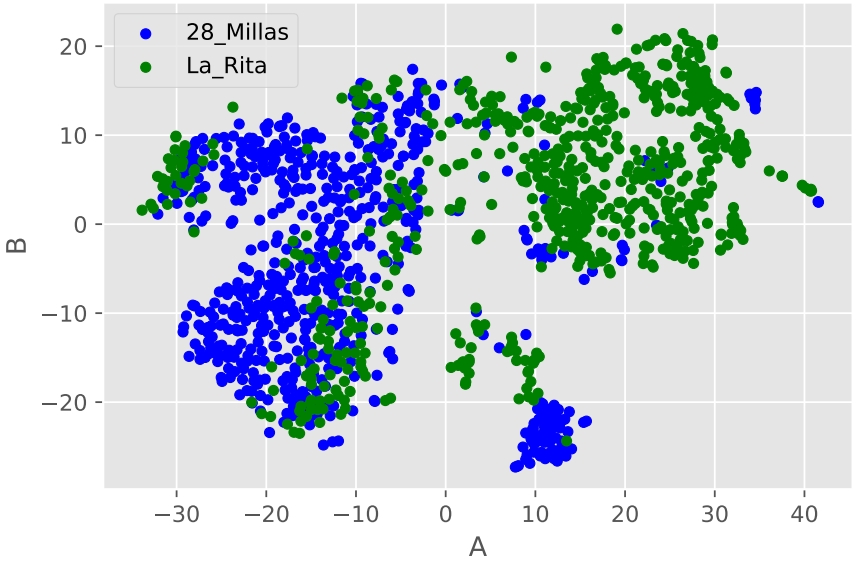


Figura 4.2: Aplicación de tSNE entre los conjuntos de datos de La Rita y 28 Millas, divergencia KL de 0.88.

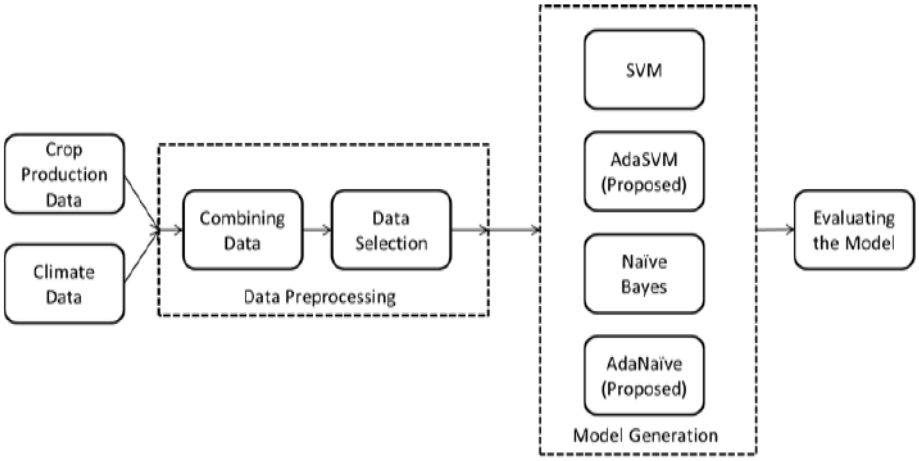


Figura 4.3: Crop Production Ensemble Machine Learning Model for Prediction.
(Tomado de Balakrishnan y Muthukumarasamy [12])

nivel de detalle que se presenta en el Capítulo 3 de este documento. En [12], el esquema propuesto es el de menor detalle de los tres. La propuesta en [33] omite procesos como: aumento de datos, selección de otros conjuntos de datos para entrenamiento y pruebas, y manejo de un histórico de del conocimiento aprendido, como sí lo hace la presente propuesta con el RCA. Por su parte, la propuesta en [65] omite aspectos como optimización multiobjetivo, aumentos de datos, no aclara el tema de la selección paramétrica y se auto-limita al campo de la horticultura.

La estrategia propuesta inicia delimitando su aplicabilidad.

Revisando [2], [12], [24], [33], [65], [94], [97], [107], se puede apreciar que en ninguna de estas propuestas se presenta algo similar al contenido de la sección 3.1 (Etapa preliminar) y aunque sí presentan algunos aspectos similares, no lo hacen de manera esquemática como en las subsecciones 3.1.1 (Delimitación de uso) y 3.1.2 (Comprensión del RCA). En concordancia con el teorema del *no hay almuerzo gratis* [117], la Delimitación de uso acota en qué casos aplica la estrategia propuesta y en cuáles no. Por su parte, la sección Comprensión del RCA, explica el concepto de Repositorio de Conocimiento Aprendido (RCA), no presente en las otras propuestas.

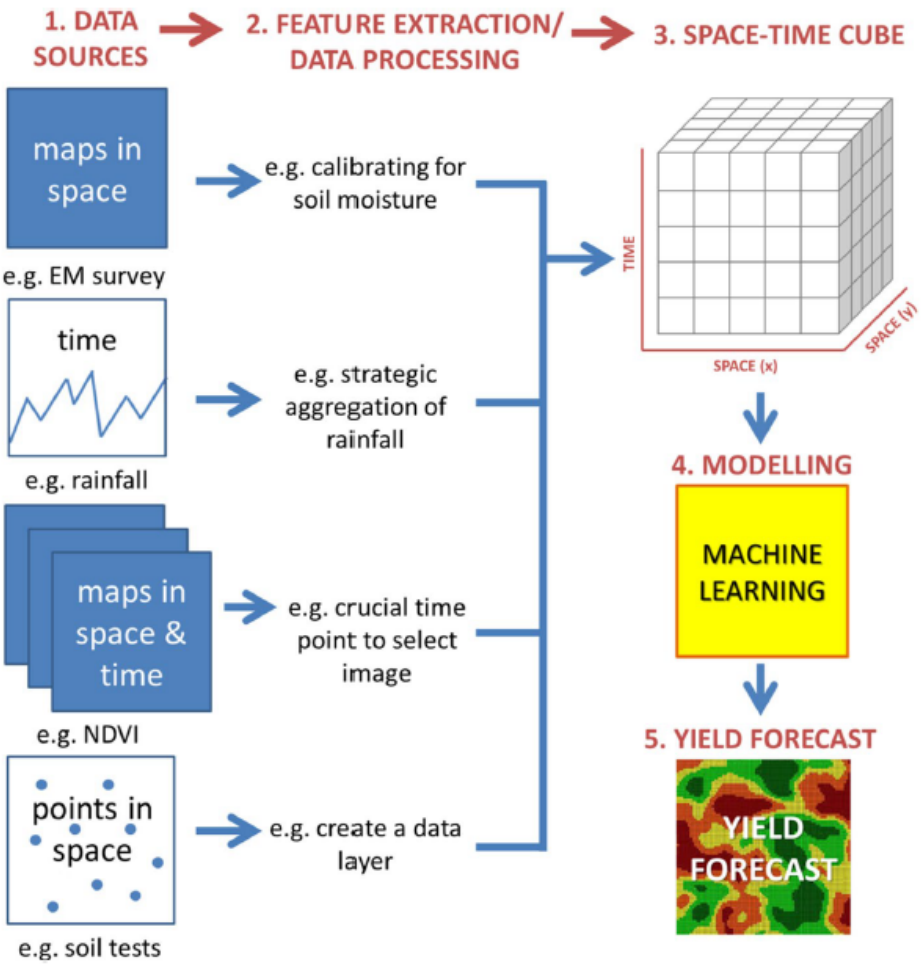


Figura 4.4: An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning.
(Tomado de Filippi, Jones, Wimalathunge y col. [33])

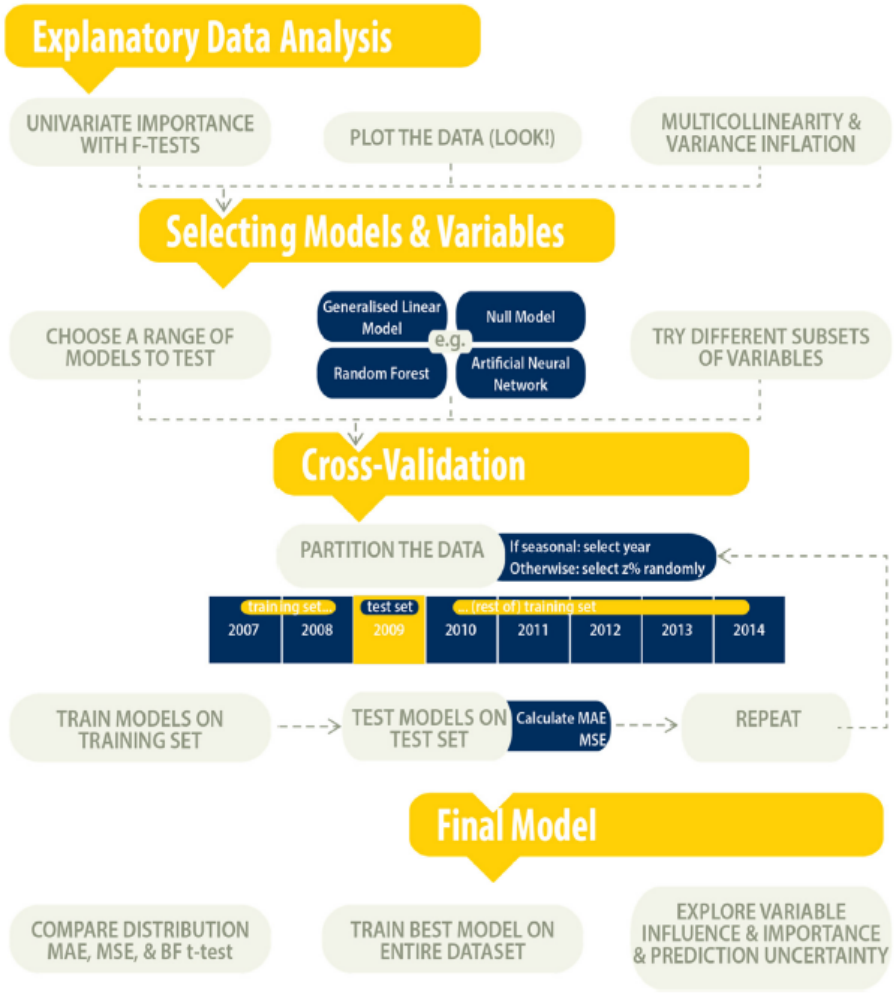


Figura 4.5: Predictive models in horticulture: A case study with Royal Gala apples. (Tomado de Logan, McLeod y Guikema [65])

La estrategia no requiere predecir variables meteorológicas, sino que capta el efecto ya producido en el tiempo.

Como se ha indicado en los Capítulos 1 y 2, se evitan predecir variables climatológicas por cuanto éstas representan fenómenos caóticos y su predicción es toda un área de investigación. El enfoque de aprendizaje supervisado y la manera en que se generan los patrones (ver subsección 3.3.1, Generación de patrones) evitan hacer lo anterior y más bien buscan captar el efecto ya producido por el clima en el proceso biológico. Al respecto, trabajos como [84] profundizan en el concepto de *Agricultura Climáticamente Inteligente* (del inglés *Climate Smart Agriculture*) que es un enfoque para guiar la gestión de la agricultura en la era del cambio climático. La presente estrategia, por medio del aprendizaje automático, busca que se vaya aprendiendo de los datos cómo el cambio climático va afectando el proceso biológico, más que predecir el cambio climático. Si bien se han publicado trabajos que utilizan variables climatológicas como parte de sus variables de entrada, tales como [2], [92], [31], [15], [66], [28], [105], no se encontró una propuesta que incluyera la predicción a más de un periodo adelante y de la manera en que se realiza en esta investigación con la generación de patrones (ver subsección 3.3.1, periodos antes (p) y periodos después (a)).

Propone un método de aumento de datos para tratar de mejorar la predicción dentro de la Estrategia.

Este aspecto no se encontró en los trabajos relacionados del estado del arte. Reconociendo que el aporte es la incorporación de este proceso dentro de la estrategia (ver subsección 3.2.4), más que el modo de generar más muestras a parir del conjunto de datos existente.

No requiere contar con imágenes para iniciar con la experimentación.

Trabajos relacionados utilizan imágenes como parte de sus variables de entrada, [2], [7], [17], [33], [75], [90], [94], [96], [107] entre otros; pero por los motivos indicados en el Capítulo 1, son prescindibles en este trabajo.

Propone una manera de trabajar con el espacio paramétrico de manera heurística.

La propuesta de trabajar con algoritmos genéticos para la selección de parámetros, no se encontró tal cual en otros trabajos relacionados, o al menos, no lo explicitaron.

El uso del frente de Pareto entre el R^2 y el $RMSE$ permite la optimización multiobjetivo.

En los trabajos relacionados se utilizan métricas tales como: Root Mean Square Error ($RMSE$), Root Relative Square Error ($RRSE$), Mean Absolute Percentage Error ($MAPE$), Mean Absolute Error (MAE), Coeficiente de determinación (R^2), pero el uso del frente de Pareto como estrategia de selección de las mejores configuraciones no fue encontrado. Dichos trabajos utilizan dichas métricas de manera independiente a la hora de hacer sus análisis de resultados, no como optimización multiobjetivo, como sí se hace en este

trabajo.

Estrategia de solución vista como aprendizaje supervisado en ventanas de tiempo deslizantes, en lugar del enfoque tradicional de predicción en series temporales.

Si bien hay abundante teoría sobre la predicción en series de tiempo ([11], [63], [52], [118]), su manejo con métodos tradicionales como ARIMA y sus variantes, parte del cumplimiento de ciertos supuestos, en particular la estacionariedad de la serie de tiempo [54]; lo cual pone requisitos adicionales para su análisis y requiere aplicar transformaciones a las series para volverlas estacionarias. El enfoque utilizado en este trabajo, si bien requiere de una etapa de pre-procesamiento, impone menos restricciones a los datos, en particular porque si se desea captar el efecto acumulado de ciertas variables en otra, transformaciones como la diferenciación pueden ocasionar pérdidas de información contenida en la serie de tiempo original. Por ejemplo, si el desarrollo de un fenómeno depende de la acumulación de los valores de una variable, y está es modificada, esta información se modifica con la transformación.

4.2. Propuesta en el proceso de aprendizaje por transferencia

A diferencia de los métodos tradicionales de aprendizaje automático, el aprendizaje por transferencia rompe el supuesto de que los datos de entrenamiento y de prueba deben obedecer a la misma distribución, y aprovecha la experiencia pasada para nuevos dominios, diferentes, pero relacionados [119]. Por otro lado, el escalamiento multidimensional se refiere al conjunto de técnicas para interpretar la similitud o disimilitud entre los datos [37].

El objetivo de la presente sección es resaltar el aporte de la estrategia respecto al aprendizaje por transferencia. Como medio para mostrar la aplicación de la estrategia se utiliza el proceso biológico de la roya del cafeto, si se desean conocer los resultados detallados de los experimentos, se puede consultar el Apéndice C y en cuanto al proceso biológico, este fue descrito en la subsección 2.3.2.

4.2.1. Materiales y métodos

Los datos utilizados fueron proporcionados por el Centro de Investigaciones del ICAFÉ. En la tabla 4.10 se muestra la ubicación de cada finca y en qué periodo fueron recolectados los datos.

Con el fin de evitar los espacios en blanco que separan los nombres con más de una palabra, se utilizan las siguientes equivalencias en la presente sección: Barva: **Barva**, Carrizal: **Carrizal**, San Vito: **SanVito**, San Carlos: **SanCarlos**, Frailes: **Frailes**, Dota: **Dota**, Poás: **Poas**.

Tabla 4.10: Fincas utilizadas en el estudio (Caso: roya)

Finca	Ubicación	Periodo (años)
Barva	Heredia, San Pedro de Barva	2010 a 2017
Carrizal	San José, San Pablo de León Cortés	2013 a 2017
SanVito	Puntarenas, San vito de Coto Brus	2010 a 2015
SanCarlos	San José, San Carlos de Tarazú	2013 a 2017
Frailes	San José, Frailes de Desamparados	2013 a 2017
Dota	San José, Santa María de Dota	2013 a 2017
Poas	Alajuela, San Juan de Poás	2010 a 2015

En cuanto a las variables disponibles, la tabla 4.11 muestra este detalle. Valga mencionar que la variable $\overline{Lw1}$ solo está presente en los *ca* de: Dota, Carrizal y SanVito. La variable a pronosticar es la incidencia de la roya (Ir). Se utiliza periodicidad mensual, por lo que se dispone de la siguiente cantidad de observaciones por *ca*: Barva 82, Frailes 46, Dota 50, Carrizal 53, SanCarlos 51, SanVito 69 y Poas 69.

Tabla 4.11: Variables disponibles (Caso: roya)

Símbolo	Descripción	Unidades
$\overline{T_a}$	Temperatura del aire promedio	[°C]
$\overline{At}$	Temperatura máxima menos mínima	[°C]
$\overline{Hdd}$	Grados día de calor	[°C]
$\overline{Lw1}$	Mojadura foliar - hoja 1	*unidad*
$\overline{H}$	Humedad relativa promedio	[%]
$\overline{R}$	Radiación solar promedio	[W/m^2]
P	Precipitación acumulada	[mm]
$\overline{W}$	Velocidad del viento promedio	[m/s]
Ir	Incidencia de la roya	nivel

Las tablas 4.12, 4.13, 4.14, 4.15, 4.16, 4.17 y 4.18, detallan algunas estadísticas descriptivas de los conjuntos de datos utilizados.

Tabla 4.12: Estadísticas (Conjunto de datos: Barva)

Métrica	$\overline{T_a}$	$\overline{H}$	$\overline{W}$	$\overline{Sr}$	$\overline{Hdd}$	$\overline{At}$	P	Ir
Cardinalidad	82	82	82	82	82	82	82	82
Promedio	20.96	79.81	2.0	198.62	0.0	7.2	214.81	35.0
Mediana	20.92	82.25	1.27	181.82	0.0	9.5	217.8	22.6
Desviación estándar	0.77	7.97	1.75	42.55	0.0	4.82	188.95	33.61
Valor mínimo	19.01	63.56	0.27	108.58	0.0	0.3	0.0	0.0
Valor máximo	22.77	92.44	7.03	321.93	0.01	13.34	663.8	100.0
Rango	3.76	28.88	6.76	213.35	0.01	13.04	663.8	100.0
Coeficiente de variación	0.04	0.1	0.88	0.21	0.46	0.67	0.88	0.96

Tabla 4.13: Estadísticas (Conjunto de datos: Frailes)

Métrica	$\overline{T}_a$	$\overline{H}$	$\overline{W}$	$\overline{Sr}$	$\overline{Hdd}$	$\overline{At}$	P	Ir
Cardinalidad	46	46	46	46	46	46	46	46
Promedio	17.81	84.45	1.08	191.37	0.02	0.31	158.47	9.16
Mediana	17.83	85.66	0.86	186.49	0.01	0.31	99.95	10.3
Desviación estándar	0.71	2.67	0.72	34.86	0.01	0.05	181.79	3.44
Valor mínimo	16.19	77.58	0.15	118.3	0.01	0.19	2.8	1.35
Valor máximo	19.1	88.01	2.92	273.45	0.03	0.41	843.38	13.9
Rango	2.91	10.43	2.78	155.15	0.02	0.22	840.58	12.55
Coeficiente de variación	0.04	0.03	0.67	0.18	0.36	0.15	1.15	0.38

Tabla 4.14: Estadísticas (Conjunto de datos: Dota)

Métrica	$\overline{T}_a$	$\overline{H}$	$\overline{W}$	$\overline{Sr}$	$\overline{Hdd}$	$\overline{Lw1}$	$\overline{At}$	P	Ir
Cardinalidad	50	50	50	50	50	50	50	50	50
Promedio	18.68	80.68	2.62	191.89	0.01	3.84	0.37	163.16	9.23
Mediana	18.76	81.85	2.03	184.37	0.01	4.07	0.34	155.0	9.35
Desviación estándar	0.8	8.32	2.08	36.48	0.0	2.67	0.08	132.29	3.19
Valor mínimo	17.11	64.06	0.11	116.17	0.01	0.07	0.22	0.6	1.45
Valor máximo	20.64	92.33	8.42	279.4	0.02	8.46	0.59	535.2	14.55
Rango	3.53	28.27	8.3	163.23	0.02	8.38	0.37	534.6	13.1
Coeficiente de variación	0.04	0.1	0.79	0.19	0.26	0.69	0.23	0.81	0.35

El conjunto de técnicas utilizadas (T) fue: $\{SVR/L$, SVR/G, SVR/S, SVR/P, ENR, $OLSR\}$. En cuanto a los patrones, se analizaron hasta 12 meses previos y 3 meses adelante, con incrementos de 1 ($p = 12$, $a = 3$, $inc = 1$).

Además, con el fin de acortar los nombres en las tablas a presentar, se utilizan las siguientes abreviaturas: Barva : **B**, Carrizal: **C**, SanVito: **SV**, SanCarlos: **SC**, Frailes: **F**, Dota: **D**, Poas: **P**, Datos sin aumento de datos: **SS**, Datos con aumento de datos: **CS**, conjunto de datos del entrenamiento: **Tr**, y conjunto de datos de Prueba: **Te**:

Tabla 4.15: Estadísticas (Conjunto de datos: Carrizal)

Métrica	$\overline{T}_a$	$\overline{H}$	$\overline{W}$	$\overline{Sr}$	$\overline{Hdd}$	$\overline{Lw1}$	$\overline{At}$	P	Ir
Cardinalidad	53	53	53	53	53	53	53	53	53
Promedio	18.26	84.16	0.65	177.92	0.03	3.55	0.71	165.42	11.25
Mediana	18.26	86.28	0.5	159.92	0.03	2.98	0.7	129.8	12.0
Desviación estándar	0.6	6.79	0.62	38.24	0.01	2.64	0.16	162.15	5.45
Valor mínimo	17.15	62.61	0.01	114.12	0.01	0.0	0.43	0.0	0.2
Valor máximo	20.0	92.92	2.63	270.73	0.04	8.85	1.04	565.4	19.9
Rango	2.85	30.31	2.61	156.61	0.03	8.85	0.61	565.4	19.7
Coeficiente de variación	0.03	0.08	0.96	0.21	0.29	0.74	0.23	0.98	0.48

Tabla 4.16: Estadísticas (Conjunto de datos: San Carlos)

Métrica	$\overline{T_a}$	$\overline{H}$	$\overline{W}$	$\overline{Sr}$	$\overline{Hdd}$	$\overline{At}$	P	Ir
Cardinalidad	51	51	51	51	51	51	51	51
Promedio	19.06	83.6	1.04	172.04	0.01	0.55	166.38	8.97
Mediana	18.87	85.21	0.78	164.94	0.01	0.54	140.4	10.0
Desviación estándar	0.93	5.27	0.74	41.89	0.0	0.17	156.6	3.85
Valor mínimo	17.9	73.1	0.17	100.8	0.0	0.28	0.0	1.0
Valor máximo	21.73	91.04	3.14	255.02	0.02	0.96	647.4	16.15
Rango	3.83	17.93	2.97	154.22	0.02	0.69	647.4	15.15
Coeficiente de variación	0.05	0.06	0.71	0.24	0.53	0.3	0.94	0.43

Tabla 4.17: Estadísticas (Conjunto de datos: San Vito)

Métrica	$\overline{At}$	$\overline{T_a}$	$\overline{H}$	$\overline{W}$	P	$\overline{Sr}$	$\overline{Hdd}$	$\overline{Lw1}$	Ir
Cardinalidad	69	69	69	69	69	69	69	69	69
Promedio	7.94	21.82	81.6	1.24	269.23	185.81	0.0	6.01	45.2
Mediana	8.05	21.8	90.82	1.18	221.4	173.92	0.0	6.87	37.01
Desviación estándar	2.7	0.9	23.19	0.47	223.6	107.29	0.0	2.96	33.89
Valor mínimo	0.0	19.61	3.16	0.0	0.0	0.0	0.0	0.0	2.0
Valor máximo	13.17	25.53	100.0	3.73	963.64	1009.67	0.01	15.0	100.0
Rango	13.17	5.92	96.84	3.73	963.64	1009.67	0.01	15.0	98.0
Coeficiente de variación	0.34	0.04	0.28	0.38	0.83	0.58	1.48	0.49	0.75

4.2.2. Resultados y análisis

Para lograr el objetivo de la presente sección, este apartado se estructura de la siguiente manera:

1. Resultados de la aplicación del paso 3.2.5 (*Determinación de uno o varios ca para el entrenamiento*) al proceso biológico en estudio.

2. Resultado de los experimentos que combinan todos los pares de *ca* del proceso biológico en estudio, esto para poder comparar lo obtenido en el paso anterior con

Tabla 4.18: Estadísticas (Conjunto de datos: Poas)

Métrica	$\overline{At}$	$\overline{T_a}$	$\overline{H}$	$\overline{W}$	P	$\overline{Sr}$	$\overline{Hdd}$	Ir
Cardinalidad	69	69	69	69	69	69	69	69
Promedio	8.74	18.97	84.56	0.95	205.55	163.1	0.02	2.26
Mediana	10.08	19.0	85.95	0.22	144.4	153.25	0.02	0.0
Desviación estándar	4.24	0.66	6.88	1.55	205.55	40.34	0.01	4.47
Valor mínimo	0.0	17.03	69.21	0.0	0.0	83.29	0.0	0.0
Valor máximo	14.94	20.12	95.84	8.02	963.8	246.35	0.05	16.67
Rango	14.94	3.09	26.63	8.02	963.8	163.06	0.04	16.67
Coeficiente de variación	0.49	0.03	0.08	1.62	1.0	0.25	0.41	1.98

el actual.

3. Análisis comparativo de resultados.

4. Conclusiones y aporte de la propuesta en cuanto al proceso de aprendizaje por transferencia.

Determinación de uno o varios *ca* para el entrenamiento

Como caso de estudio, se toman los 7 *ca* relacionados a las fincas que indica la tabla 4.10.

El primer paso es tomar las series temporales que coincidan en la marca temporal (ver columna *Periodo* de la tabla 4.10). De este análisis resulta que se pueden tomar 46 observaciones de cada *ca* que cumplen este requisito, por lo que se tienen $46 \times 7 = 322$ observaciones en total para los 7 *ca*. Valga indicar que si se tratara de solo comparar 2 *ca*, se podrían tomar cantidades diferentes de observaciones en cada *ca* para el estudio, pero como en este caso se desean comparar los 7 *ca* en pares, tomar cantidades diferentes de observaciones y en marcas temporales diferentes, afectaría el valor obtenido en las divergencias KL (Kullback-Leibler) y en la comparación de las normas euclidianas.

En cuanto a las variables a incluir, la tabla 4.11 contabiliza 9 variables (vca). De ellas se excluye para este estudio la mojadura foliar - hoja 1 ($\overline{Lw1}$), por cuanto no todos los *ca* la tienen, y también se excluye la incidencia de la roya (Ir), ya que precisamente la búsqueda de similitud con otros *ca* se debería sobre todo a un nuevo *ca* con pocos datos de esta variable y lo que se buscaría es la similitud considerando otras variables que se podrían obtener con estaciones meteorológicas cercanas a la finca que corresponde a *ca_estudio*, quedando por lo tanto solo 7 variables.

La figura 4.6 muestra la graficación $2D$ luego de aplicar t-SNE con los conjuntos de datos de la roya. En dicha figura se aprecian dos grupos: uno que *pareciera* formado por Poas, Barva y SanVito, y otro por Frailes, Carrizal, SanCarlos y Dota. Además se aprecian puntos de Poas y de Barva en ambos grupos. Se dice *pareciera*, pues en esta figura no se tiene certeza si por el orden de graficación, algún punto está ocultando puntos de otro u otros *ca*.

Para mostrar que calcular la divergencia KL entre pares de *ca* no es criterio suficiente, en la tabla 4.19 se muestran los primeros resultados obtenidos al calcular la Divergencias KL entre pares de *ca*, ordenados de menor a mayor [1]. En dicha tabla se aprecia que las divergencias KL menores corresponden a los pares: SanVito-Frailes, SanVito-Carrizal, SanVito-SanCarlos, SanVito-Dota, Barva-Frailes, Barva-Carrizal y Barva-Dota, que no coinciden con las agrupaciones mostradas en la figura 4.6. Esto sucede porque al calcular la divergencia KL entre pares de *ca*, de manera independiente a los otros *ca* candidatos, los valores obtenidos de Divergencia KL son estimados en el algoritmo tSNE de manera que no guardan relación con los *ca* que estén en ese momento fuera de los datos en estudio. Y si a la técnica t-SNE se le incluyen todos los *ca* a la vez, lo que se obtiene es un único valor de divergencia-KL entre todos los *ca* (en este caso: 0,38392). A diferencia de lo anterior,

[1] En el Apéndice C se pueden consultar los resultados completos.

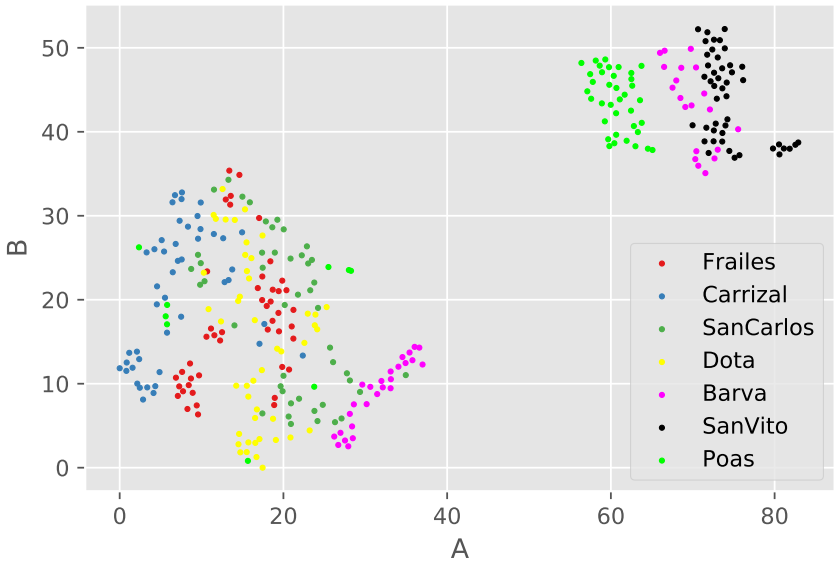


Figura 4.6: Versión gráfica 2D al aplicar tSNE en los conjuntos de datos: roya.

la propuesta de la presente estrategia consiste en calcular la norma euclidiana sobre los datos de los *ca* pero cuando ya están en el espacio embebido, por lo cual las distancias sí guardan relación con lo que se aprecia en la figura 4.6. En la tabla 4.20 se puede apreciar cómo la distancia euclidiana es única para cada *ca*, por ejemplo, 158,5761 para Dota y 161,0641 para Frailes, sin importar con quien se esté comparando.

Tabla 4.19: Primeros 7 resultados al aplicar tSNE en los conjuntos de datos de la roya, ordenados por divergencia KL

Orden	Divergencia KL	ca
1	0,0499	SanVito Frailes
2	0,0716	SanVito Carrizal
3	0,0745	SanVito SanCarlos
4	0,0746	SanVito Dota
5	0,079	Barva Frailes
6	0,092	Barva Carrizal
7	0,1142	Barva Dota

Por otro lado, la estrategia propuesta en esta investigación indica que para que un *ca*

sea considerado un posible buen candidato para ser parte del entrenamiento del modelo para predecir otro *ca*, se deben analizar dos criterios (ver subsección 3.2.5): 1) preferir los valores más bajos de *pnca* y 2) que en el gráfico $3D$, los puntos que representan a los *ca* seleccionados en el punto anterior, estén espacialmente cercanos o incluso alineados.

En la tabla 4.20, se aprecian los primeros 7 resultados respecto a las diferencias relativas entre las normas euclidianas de los diferentes pares de *ca*, ordenadas de menor a mayor según su *pnca*. En este caso, los primeros lugares los ocupan: Dota-Carrizal, Carrizal-Frailes, Dota-Frailes, Poas-Barva, SanCarlos-Frailes, SanCarlos-Carrizal y Dota-SanCarlos; los cuales tienen una mayor coincidencia con las agrupaciones mostradas en la figura 4.6 [2]. Se propone utilizar una distancia relativa, pues el valor de la distancia entre normas puede variar mucho con solo agregar una mayor cantidad de datos, mientras que la distancia relativa es comparable con las otras; por ejemplo, vale notar que las tres primeras posiciones de la tabla 4.20 tienen valores menores que 1 %, y del cuarto lugar en adelante ya pasa a valores mayores o iguales a 6 %. Para las posiciones 20 y 21, el *pnca* llega a 57,41 %.

Tabla 4.20: Primeros 7 resultados al aplicar tSNE en los conjuntos de datos de la roya, ordenados por *pnca* (7 variables)

Orden	**Conjunto de datos** (*ca*)	**Norma euclidiana** (*nca*)	**Distancia** (*dnca*)	**Distancia relativa** (*pnca*)
1	Dota Carrizal	158,5761 158,5768	0,0007	0,0002 %
2	Carrizal Frailes	158,5768 161,0641	2,4873	0,7782 %
3	Dota Frailes	158,5761 161,0641	2,488	0,7784 %
4	Poas Barva	461,3744 401,9853	59,3891	6,8788 %
5	SanCarlos Frailes	189,1821 161,0641	28,118	8,0281 %
6	SanCarlos Carrizal	189,1821 158,5768	30,6053	8,8007 %
7	Dota SanCarlos	158,5761 189,1821	30,606	8,8009 %

El siguiente criterio propuesto corresponde a la graficación $3D$ de los resultados, pero a partir del cálculo con dos componentes. En la figura 4.7 se muestran varias vistas de manera ilustrativa. Se aprecia cómo al graficar los siete planos $2D$ de manera espaciada en la graficación $3D$ se logra dilucidar si hay traslape o no de puntos de diferentes *ca*, lo cual en la representación $2D$ no es factible asegurar; por ejemplo, las subfiguras: (c), (d), (e) y (f), de la figura 4.7, clarifican los traslapes; por ejemplo, Barva tiene puntos cerca de ambos grupos, y Poas, que en el gráfico $2D$ parecía tener casi la totalidad de sus puntos

[2] En el Apéndice C se presentan los resultados completos de la tabla 4.20.

en el grupo de SanVito y Barva, ahora se ve que la distribución no es significativamente diferente, sino que también hay varios puntos en el grupo de Dota, Carrizal, Frailes y SanCarlos. Por su parte, SanVito sí se mantiene en un solo grupo.

Una consideración particular requieren Dota, Carrizal y SanVito, por cuanto sólo estos *ca* cuentan con la variable mojadura foliar - hoja 1 ($\overline{Lw1}$). Si al realizar los experimentos en la validación cruzada, esta variable forma parte de las configuraciones en su frente de Pareto, o no podrán servir como entrenamiento para otros *ca* que no la tengan, o habría que entrenar sin esta variable como parte del conjunto de variables de entrada.

Previo a pasar al siguiente apartado, la estrategia permite que el equipo de investigación tome decisiones sobre las combinaciones de variables considerando otros aspectos, se presentan tres variantes a manera de ejemplo de uso de la estrategia:

- Que se desee incluir la variable que indica la incidencia de la roya, pasando de 7 a 8 variables: En la figura 4.8 (a)-(b), se puede apreciar en la versión 3D (a) y la versión 2D (b), que el resultado no cambia significativamente con lo obtenido sin incluir la incidencia de la roya. En cuanto al *pnca*, de los siete primeros lugares obtenidos en la tabla 4.20 no hay un cambio importante en el orden, aunque sí varía en magnitud. Esa diferencia en magnitud del *pnca* se explica pues al agregar más variables, las normas de las matrices de cada *ca* va a cambiar. De la siguiente lista se subraya la única pareja que aparece ahora entre los 7 primeros lugares que no aparecía en la tabla 4.20:

  1. Poas-SanVito: 1,35 %
  2. SanCarlos-Carrizal: 2,01 %
  3. Dota-Frailes: 3,74 %
  4. Dota-Carrizal: 5,29 %
  5. Dota-SanCarlos: 7,29 %
  6. Carrizal-Frailes: 9,01 %
  7. Poas-Barva: 9,48 %

- Que conociendo la importancia de ciertas variables en el proceso, solo se deseen incluir estas: Se seleccionan por su nominación en estudios previos: temperatura del aire promedio ($\overline{T}_a$), temperatura máxima menos mínima ($\overline{At}$), humedad relativa promedio ($\overline{H}$), precipitación acumulada (P) y velocidad del viento promedio ($\overline{W}$). En la figura 4.8 (c)-(d), nuevamente no se aprecian diferencias significativas con lo obtenido hasta ahora. En cuanto al *pnca*, y de manera similar al caso anterior, de los siete primeros lugares obtenidos en la tabla 4.20, coinciden 6. De la siguiente lista se subraya la única pareja que aparece ahora entre los 7 primeros lugares que no aparecía en la tabla 4.20:

  1. Dota-Carrizal: 2,46 %
  2. SanCarlos-Carrizal: 2,74 %

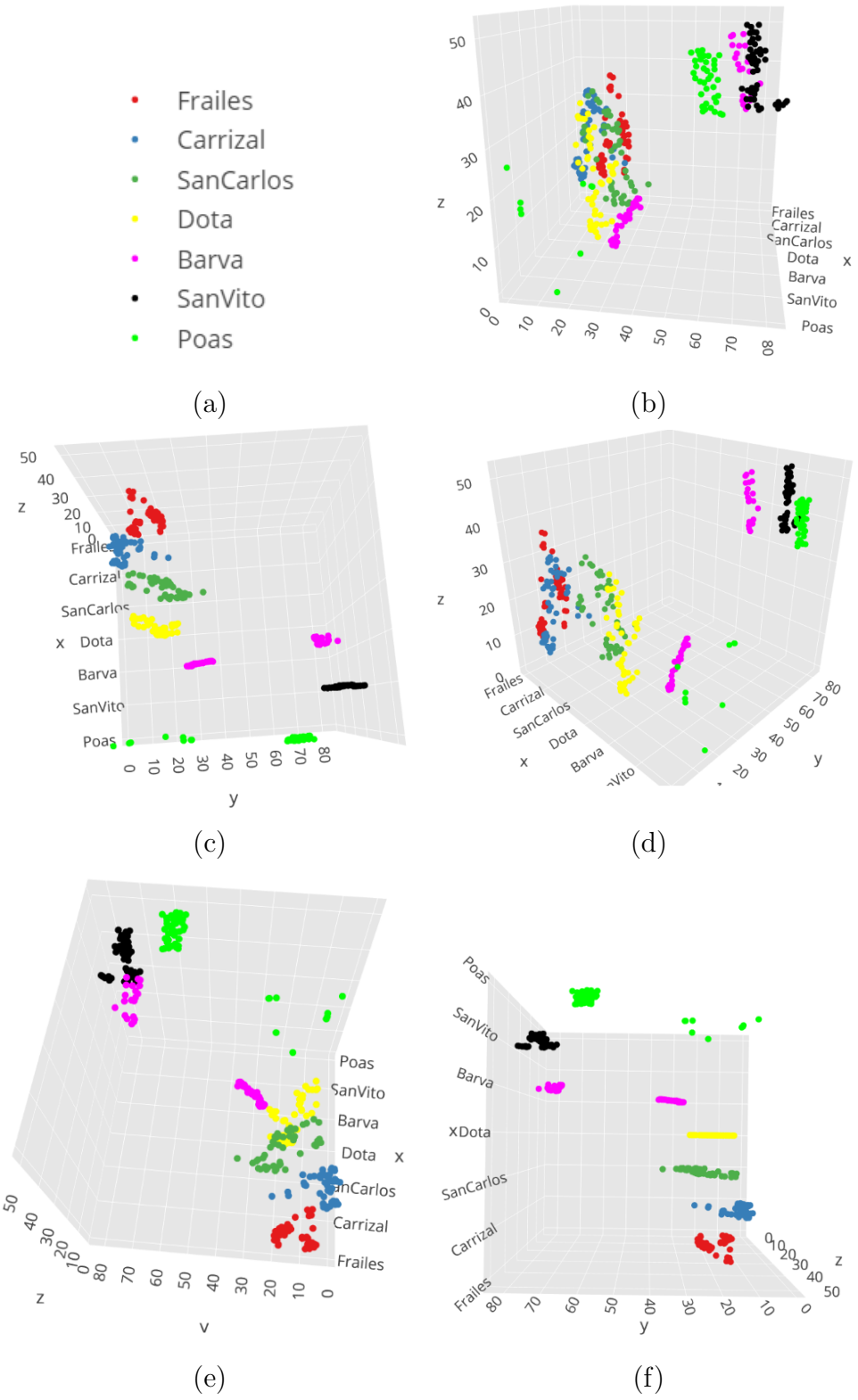


Figura 4.7: Varias vistas de la versión gráfica 3D al aplicar tSNE en los conjuntos de datos de la roya, considerando 7 variables (reducción a 2 componentes): (a) Simbología utilizada. De (b) a (f) diferentes vistas del gráfico 3D.

3. Dota-Frailes: 2,85 %
4. Dota-SanCarlos: 5,19 %
5. Carrizal-Frailes: 5,31 %
6. Poas-SanVito: 7,9 %
7. SanCarlos-Frailes: 8,04 %

- Que se desee analizar el comportamiento, por separado, de los tres *ca* que incluyen la variable mojadura foliar - hoja 1 ($\overline{Lw1}$): Dota, Carrizal y SanVito. La figura 4.8 (e)-(f), reafirman la cercanía de Carrizal con Dota y que SanVito se encuentre más distante, de hecho en cuanto a *pnca*, el de Dota-Carrizal es 0,51 %, mientras que SanVito-Dota llega a 58,83 % y SanVito-Carrizal 59,17 %.

Finalmente, la propuesta de [68] (visualizando datos utilizando t-SNE), permite reducir el resultado no solo a dos componentes, sino a tres o más componentes. Al respecto, podría surgir la pregunta de por qué entonces no graficar el resultado en 3D a partir de la reducción a tres componentes, en lugar de a dos componentes, como se propone en la presente propuesta. En la figura 4.9 se pueden apreciar varias vistas de la graficación en 3D de los resultados utilizando una reducción a tres componentes, en lugar de a dos componentes. Como se observa en las subfiguras (b), (c) y (d), el problema de hacerlo así, es que mantiene la limitación de la graficación en 2D, en que se pueden traslapar puntos de varios grupos. Además, cuando los puntos de varios grupos están muy cercanos, se pueden formar aglomeraciones de puntos que imposibilitaría distinguir puntos de grupos diferentes que se encuentren ocultos dentro de la aglomeración.

Resultados de los experimentos

Se realizaron los siguientes experimentos:

1. Se realiza *ten-fold-cross-validation* con los datos de cada una de las siete fincas por separado (**Validación cruzada**), 7 experimentos.
2. Se hace el entrenamiento con cada una de seis fincas (por separado) y se prueba con la finca restante (**Tr:xx-SS / Te:yy-SS**). Para el Entrenamiento y Pruebas, se utilizan las configuraciones presentes en el frente de Pareto de la finca de entrenamiento, las cuales fueron obtenidas en la validación cruzada de la misma. Solo se omite una configuración si no es posible utilizarla por no contar con las mismas variables en ambos conjuntos de datos, 39 experimentos.
3. Se hace el entrenamiento con cada una de seis fincas (por separado), utilizando el aumento de datos y se prueba con la finca restante sin hacerle aumento de datos (**Tr:xx-CS / Te:yy-SS**). Para el Entrenamiento y Pruebas, se utilizan las configuraciones presentes en el frente de Pareto de la finca de entrenamiento, las cuales fueron obtenidas en la Validación cruzada de la misma. Solo se omite una configuración si no es posible utilizarla por no contar con las mismas variables en ambos conjuntos de datos. 39 experimentos.

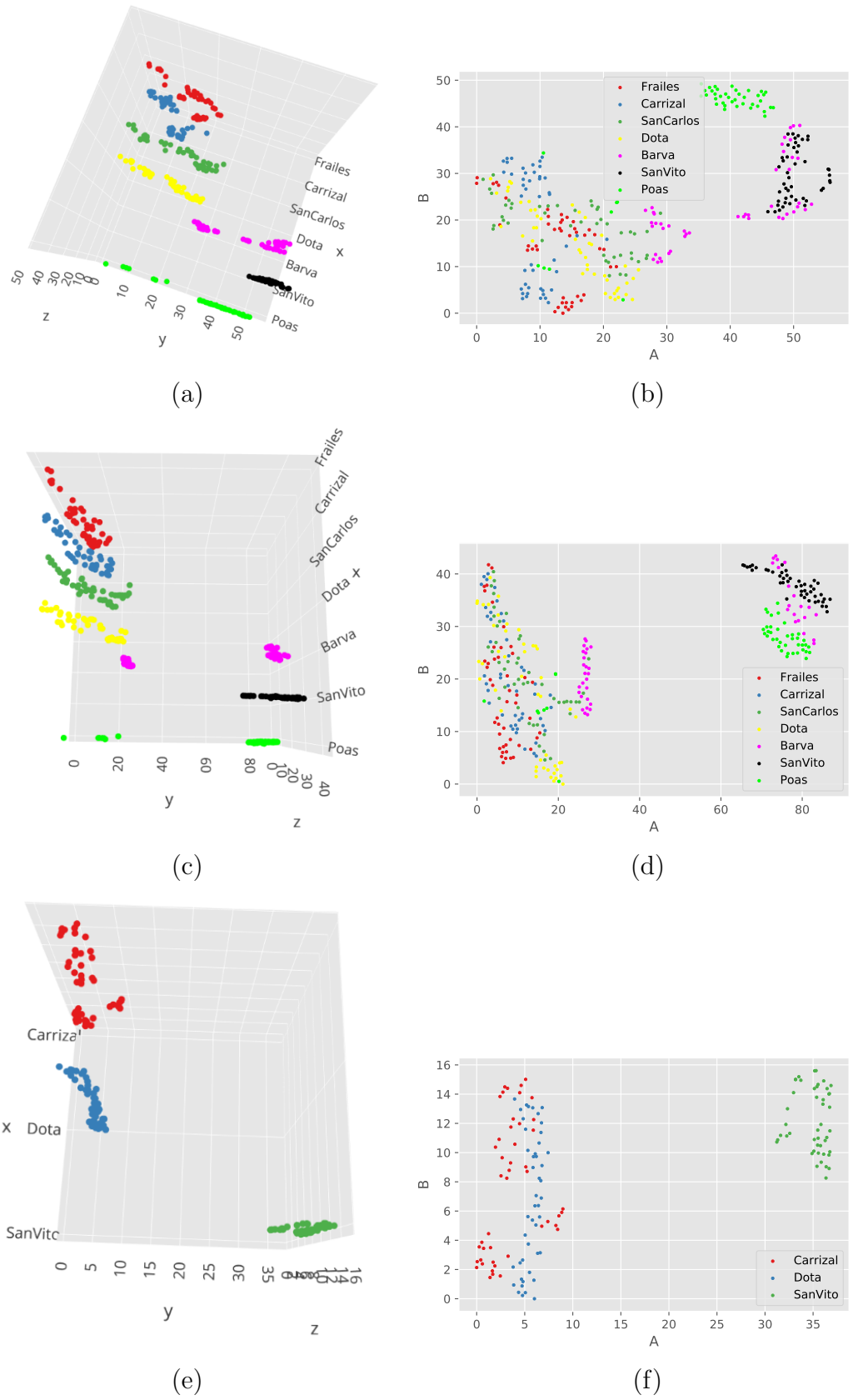
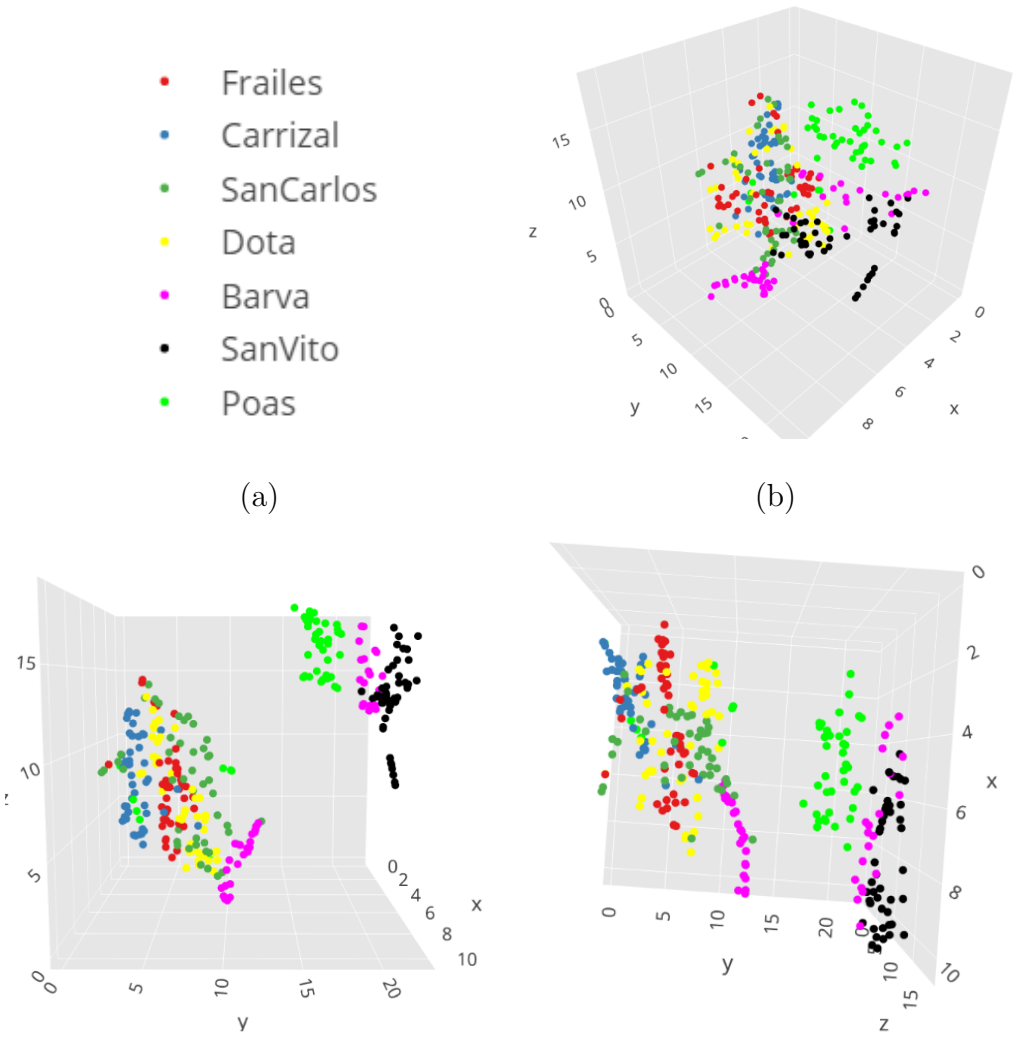


Figura 4.8: Varias vistas de la versión gráfica 3D al aplicar tSNE en los conjuntos de datos de la roya: (a) (b) Incluye incidencia de la roya, (c) (d) Cinco variables seleccionadas, (e) (f) Solo Carrizal-Dota-SanVito.



(a) (b)

(c) (d)

Figura 4.9: Varias vistas de la versión gráfica 3D al aplicar tSNE en los conjuntos de datos de la roya, utilizando reducción a tres componentes en lugar de dos: (a) Simbología, (b) (c) (d) Vistas.

La tabla 4.21 muestra las mejores configuraciones según el frente de Pareto en la etapa de validación cruzada. La tabla está ordenada por Finca (*ca*) y luego por $RMSE$ de menor a mayor [3]. Es de resaltar que salvo en el caso de Poas, no son necesarias todas las variables para obtener los mejores resultados.

Tabla 4.21: Configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para la incidencia de la roya en la etapa de validación cruzada

Lugar	Variables	$p \rightarrow a$	Técnica	$RMSE$	R^2
Barva	$\overline{H}\ \overline{W}\ \overline{Hdd}$	$10 \rightarrow 1$	SVR/P	14,73	76,62 %
	$\overline{T_a}\ \overline{H}\ \overline{W}\ P$	$1 \rightarrow 1$	SVR/G	14,86	77,69 %
		$1 \rightarrow 1$	SVR/P	15,18	80,74 %
Carrizal	$\overline{At}\ \overline{T_a}$	$10 \rightarrow 2$	SVR/P	2,47	76,79 %
	$\overline{At}\ \overline{H}\ \overline{Hdd}\ \overline{Lw1}$	$5 \rightarrow 1$	SVR/P	2,52	80,64 %
Dota	$\overline{W}\ P\ \overline{Hdd}$	$4 \rightarrow 1$	SVR/S	1,64	75,42 %
	$\overline{At}\ \overline{T_a}\ \overline{Hdd}\ \overline{Lw1}$	$2 \rightarrow 1$	SVR/P	1,66	78,2 %
	$P\ \overline{Hdd}$	$12 \rightarrow 2$	SVR/P	1,67	78,3 %
	$\overline{At}\ \overline{T_a}\ P\ \overline{Hdd}$	$3 \rightarrow 1$	SVR/P	1,81	78,9 %
Frailes	$\overline{At}$	$11 \rightarrow 1$	SVR/P	1,27	73,07 %
	$\overline{T_a}\ \overline{W}\ \overline{Hdd}$	$3 \rightarrow 1$	SVR/P	1,56	80,19 %
Poas	All	$1 \rightarrow 1$	SVR/P	2,85	60,59 %
San Carlos	$\overline{At}\ \overline{T_a}\ \overline{W}\ \overline{Hdd}$	$11 \rightarrow 1$	SVR/P	1,72	77,52 %
San Vito	$\overline{At}\ \overline{T_a}\ \overline{H}\ P$	$2 \rightarrow 1$	SVR/P	11,7	84,62 %

Por su parte, la tabla 4.22 muestra las mejores configuraciones pertenecientes al frente de Pareto en la etapa de Entrenamiento/Pruebas. La tabla está ordenada por Finca (*ca*) y luego por $RMSE$ (de menor a mayor) [4]. Consistente con los resultados obtenidos en las secciones previas, se aprecia cómo las fincas que mostraron ser buenas candidatas para hacer transferencia de aprendizaje entre sí, a la hora de ser utilizadas para entrenar, presentan resultados aceptables. Solo como ejemplo, el conjunto de datos de Dota muestra ser un buen predictor de varias de las otras fincas, tal como se analizó en las secciones previas.

La figura 4.10 muestra el frente de Pareto entre R^2 y $RMSE$ para la roya, incluyendo todas las fincas en estudio. Los resultados de esta figura son consistentes con los resultados obtenidos en las tablas 4.21 y 4.22. Lo único es que debe considerarse que por la cantidad de puntos graficados, hay trasposición de los mismos.

Análisis comparativo de resultados

Con el fin de comparar los resultados de la etapa de determinación de uno o varios *ca* para el entrenamiento (tabla 4.20 y figura 4.7) y la etapa de Entrenamiento/Pruebas (tabla 4.22), se presenta la figura 4.11, la cual se comenta a continuación:

[3] En el Apéndice C se pueden observar más resultados.
[4] En el Apéndice C se pueden observar más resultados.

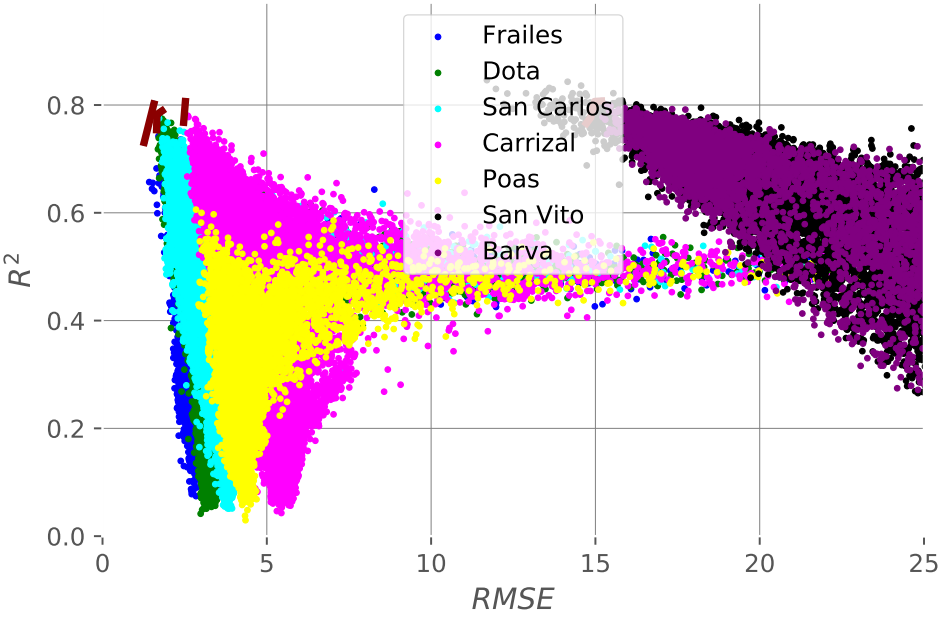


Figura 4.10: Frente de Pareto entre R^2 y $RMSE$ (Proceso biológico: roya).

Entrenamiento	Orden *pnca*	Pruebas
Dota	14	Barva
Dota	1	Carrizal
Frailes SanCarlos Carrizal	3 7 1	Dota
Dota	3	Frailes
Dota	17	Poas
Frailes	5	SanCarlos
Dota	6	SanVito

Figura 4.11: Esquema comparativo de los resultados de la etapa de determinación de uno o varios *ca* para el entrenamiento y la etapa de Entrenamiento/Pruebas, para la roya del cafeto.

Tabla 4.22: Configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para la incidencia de la roya en la etapa de entrenamiento y pruebas

Lugar	Variables	$p \rightarrow a$	Técnica	$RMSE$	R^2	Experimento
Barva	$\overline{W}\ P\ \overline{Hdd}$	$4 \rightarrow 1$	SVR/S	23,86	76,58 %	Tr:D-CS / Te:B-SS
		$4 \rightarrow 1$	SVR/S	23,08	69,25 %	Tr:D-SS / Te:B-SS
Carrizal	$\overline{At}\ \overline{T_a}\ \overline{Hdd}\ \overline{Lw1}$	$2 \rightarrow 1$	SVR/P	3,4	68,26 %	Tr:D-SS / Te:C-SS
Dota	$\overline{At}\ \overline{H}\ \overline{Hdd}\ \overline{Lw1}$	$5 \rightarrow 1$	SVR/P	4,95	56,49 %	Tr:C-SS / Te:D-SS
	$\overline{At}\ \overline{T_a}\ \overline{W}\ \overline{Hdd}$	$11 \rightarrow 1$	SVR/P	4,94	55,41 %	Tr:SC-CS / Te:D-SS
	$\overline{At}$	$11 \rightarrow 1$	SVR/P	3,71	41,35 %	Tr:F-CS / Te:D-SS
		$11 \rightarrow 1$	SVR/P	3,58	40,43 %	Tr:F-SS / Te:D-SS
Frailes	$\overline{W}\ P\ \overline{Hdd}$	$4 \rightarrow 1$	SVR/S	2,66	68,5 %	Tr:D-SS / Te:F-SS
	$\overline{W}\ P\ \overline{Hdd}$	$4 \rightarrow 1$	SVR/S	11,05	54,71 %	Tr:D-CS / Te:P-SS
Poas	$P\ \overline{Hdd}$	$12 \rightarrow 2$	SVR/P	10,08	45,9 %	Tr:D-CS / Te:P-SS
	$\overline{W}\ P\ \overline{Hdd}$	$4 \rightarrow 1$	SVR/S	4,82	44,91 %	Tr:D-SS / Te:P-SS
SanCarlos	$\overline{T_a}\ \overline{W}\ \overline{Hdd}$	$3 \rightarrow 1$	SVR/P	2,16	71,68 %	Tr:F-CS / Te:SC-SS
SanVito	$\overline{W}\ P\ \overline{Hdd}$	$4 \rightarrow 1$	SVR/S	30,51	71,61 %	Tr:D-CS / Te:SV-SS
		$4 \rightarrow 1$	SVR/S	23,02	70,4 %	Tr:D-SS / Te:SV-SS

1. Barva: Aunque en la tabla 4.20, la primer aparición de Barva es en el orden 4, en conjunto con el *ca* de Poas, en el Entrenamiento/Pruebas (tabla 4.22) es con el *ca* de Dota que obtiene sus mejores métricas (Tr:D-CS / Te:B-SS: $RMSE$ 23,86, R^2 76,58 %. Tr:D-SS / Te:B-SS: $RMSE$ 23,08, R^2 69,25 %). Al respecto hay dos explicaciones: 1) aunque no entró en el frente de Pareto, la configuración de Barva con Poas obtuvo las siguientes métricas: Tr:P-SS/Te:B-SS, $RMSE$ 27,82, R^2 53,73 %, cercanas a las del frente de Pareto [5]; 2) con el fin de poder comparar los siete *ca* en pares, se limitó el número de observaciones de cada *ca* a la misma cantidad y con la misma marca temporal, pero en realidad Barva tiene más observaciones que los otros, 82 observaciones Barba (tabla 4.12), 50 Dota (tabla 4.14) y 69 Poas (tabla 4.18); por ello, al realizar el Entrenamiento/Pruebas con la totalidad de observaciones disponibles de Barva, es factible que se genere un resultado en el que Barva se asemeje más a Dota que a Poas.

2. Con los siguientes *ca* se encuentra congruencia entre las configuraciones de su frente de Pareto en Entrenamiento/Pruebas y el orden de su *pnca*. Carrizal: Orden 1, Dota: Orden 1, 3 y 7, SanCarlos: Orden 5, SanVito: Orden 6 y Frailes: Orden 3.

3. Poas: Se presenta una situación muy similar a lo indicado con Barva. Aunque en la tabla 4.20, la primer aparición de Poas es en el orden 4, en conjunto con el *ca* de Barva, en el Entrenamiento/Pruebas (tabla 4.22) es con el *ca* de Dota que obtiene sus mejores métricas (Tr:D-CS / Te:P-SS: $RMSE$ 10,08, R^2 45,9 %. Tr:D-SS / Te:P-SS: $RMSE$ 4,82, R^2 44,91 %). Aun así, el *ca* de Poas con respecto al *ca* de Barva obtiene las siguientes métricas: Tr:B-CS/Te:P-SS, $RMSE$ 15,17, R^2 51,64 %; y,

[5] El detalle completo se puede consultar en la tabla C.3, ubicada en el Apéndice C.

Tr:B-SS/Te:P-SS, $RMSE$ 16,49, R^2 50,6 % [6]; cercanas a las pertenecientes al frente de Pareto de Poas. Además, se debe considerar la mayor cantidad de observaciones consideradas en la etapa de Entrenamiento/Pruebas, 82 observaciones Barba (tabla 4.12), 50 Dota (tabla 4.14) y 69 Poas (tabla 4.18).

Se realizaron experimentos adicionales tomando los cuatro *ca* que se apreciaban más similares en la subsección 4.2.2, a saber: Dota, Carrizal, Frailes y SanCarlos. La tabla 4.23 muestra las configuraciones que conforman el frente de Pareto de cada *ca* al entrenar con los otros dos o tres *ca* restantes. Una diferencia a considerar, es que en este caso la configuraciones utilizadas fueron las del frente de Pareto del *ca* en Pruebas, a diferencia de lo realizado para la tabla 4.22, en que se utilizaron las configuraciones del *ca* en Entrenamiento; esto porque al tener más de un *ca* en el entrenamiento, cada uno tiene su frente de Pareto, por lo que se decide tomar el del *ca* en pruebas. Cuando era solo un *ca* en el entrenamiento, no había duda en utilizar el frente de Pareto del *ca* en entrenamiento.

Tabla 4.23: Configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para la incidencia de la roya en la etapa de entrenamiento y pruebas al utilizar más de una finca en el entrenamiento (D:Dota, C:Carrizal, F:Frailes, SC: SanCarlos)

Lugar	**Variables**	$p \rightarrow a$	**Técnica**	$RMSE$	R^2	**Experimento**
Carrizal	$\overline{At}\ \overline{H}\ \overline{Hdd}$	$5 \rightarrow 1$	SVR/P	5,1	63,68 %	Tr:D-F / Te:C
Dota	$\overline{At}\ \overline{T}_a\ P\ \overline{Hdd}$	$3 \rightarrow 1$	SVR/P	2,15	68,6 %	Tr:F-SC / Te:D
Frailes	$\overline{T}_a\ \overline{W}\ \overline{Hdd}$	$3 \rightarrow 1$	SVR/P	2,31	68,62 %	Tr:D-SC / Te:F
SanCarlos	$\overline{At}\ \overline{T}_a\ \overline{W}\ \overline{Hdd}$	$11 \rightarrow 1$	SVR/P	2,77	59,87 %	Tr:D-F-C / Te:SC
		$11 \rightarrow 1$	SVR/P	2,96	62,46 %	Tr:D-C / Te:SC

Al comparar los resultados para estos cuatro *ca*, los cuales fueron presentados en las tablas: 4.21 (Validación cruzada), 4.22 (Entrenamiento/Pruebas en comparación uno a uno) y 4.23 (Entrenamiento/Pruebas utilizando varios *ca* en el entrenamiento); se observa en la tabla 4.24, que utilizar más de uno de los *ca* en el entrenamiento no necesariamente mejora las métricas obtenidas. Solo para Dota, el usar más de un *ca* en entrenamiento (Frailes y SanCarlos), implicó una mejora sustancial de lo obtenido en el frente de Pareto al entrenar con solo un *ca* (Carrizal), pues en cuanto al $RMSE$ se pasó de 4,95 a 2,15 y en cuanto al R^2 se pasó de 56,49 % a 68,6 %. En Frailes también se presenta una pequeña mejora al incorporar dos *ca* en el entrenamiento, pero la diferencia es tan poca, que sus valores están dentro del margen de error de la predicción.

Conclusiones y aporte

El aporte de la presente estrategia se puede resumir en aprovechar la propuesta de Maaten e Hinton [68], enfocada en el escalamiento multidimensional para la visualización de datos de alta dimensionalidad, agregarle varios criterios de comparación y selección de

[6] El detalle completo se puede consultar en la tabla C.7, ubicada en el Apéndice C.

Tabla 4.24: Comparación de las configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para Dota, Carrizal, Frailes y SanCarlos, en tres etapas: 1) validación cruzada, 2) entrenamiento y pruebas con solo un *ca* en el entrenamiento, y 3) entrenamiento y pruebas con dos o tres *ca* en el entrenamiento

Lugar	Variables	$p \rightarrow a$	Técnica	$RMSE$	R^2	Experimento
Carrizal	$\overline{At}\ \overline{H}\ \overline{Hdd}\ \overline{Lw1}$	$5 \rightarrow 1$	SVR/P	2,52	80,64 %	Validación cruzada
	$\overline{At}\ \overline{T_a}$	$10 \rightarrow 2$	SVR/P	2,47	76,79 %	Validación cruzada
	$\overline{At}\ \overline{T_a}\ \overline{Hdd}\ \overline{Lw1}$	$2 \rightarrow 1$	SVR/P	3,4	68,26 %	Tr:D-SS / Te:C-SS
	$\overline{At}\ \overline{H}\ \overline{Hdd}$	$5 \rightarrow 1$	SVR/P	5,1	63,68 %	Tr:D-F-SS / Te:C-SS
Dota	$\overline{At}\ \overline{T_a}\ P\ \overline{Hdd}$	$3 \rightarrow 1$	SVR/P	1,81	78,9 %	Validación cruzada
	$P\ \overline{Hdd}$	$12 \rightarrow 2$	SVR/P	1,67	78,3 %	Validación cruzada
	$\overline{At}\ \overline{T_a}\ \overline{Hdd}\ \overline{Lw1}$	$2 \rightarrow 1$	SVR/P	1,66	78,2 %	Validación cruzada
	$\overline{W}\ P\ \overline{Hdd}$	$4 \rightarrow 1$	SVR/S	1,64	75,42 %	Validación cruzada
	$\overline{At}\ \overline{T_a}\ P\ \overline{Hdd}$	$3 \rightarrow 1$	SVR/P	2,15	68,6 %	Tr:F-SC-SS / Te:D-SS
	$\overline{At}\ \overline{H}\ \overline{Hdd}\ \overline{Lw1}$	$5 \rightarrow 1$	SVR/P	4,95	56,49 %	Tr:C-SS / Te:D-SS
	$\overline{At}$	$11 \rightarrow 1$	SVR/P	3,58	40,43 %	Tr:F-SS / Te:D-SS
Frailes	$\overline{T_a}\ \overline{W}\ \overline{Hdd}$	$3 \rightarrow 1$	SVR/P	1,56	80,19 %	Validación cruzada
	$\overline{At}$	$11 \rightarrow 1$	SVR/P	1,27	73,07 %	Validación cruzada
	$\overline{T_a}\ \overline{W}\ \overline{Hdd}$	$3 \rightarrow 1$	SVR/P	2,31	68,62 %	Tr:D-SC-SS / Te:F-SS
	$\overline{W}\ P\ \overline{Hdd}$	$4 \rightarrow 1$	SVR/S	2,66	68,5 %	Tr:D-SS / Te:F-SS
SanCarlos	$\overline{At}\ \overline{T_a}\ \overline{W}\ \overline{Hdd}$	$11 \rightarrow 1$	SVR/P	1,72	77,52 %	Validación cruzada
	$\overline{W}\ P\ \overline{Hdd}$	$4 \rightarrow 1$	SVR/S	3,17	65,35 %	Tr:D-SS / Te:SC-SS
	$\overline{At}\ \overline{T_a}\ P\ \overline{Hdd}$	$3 \rightarrow 1$	SVR/P	2,69	64,75 %	Tr:D-SS / Te:SC-SS
	$\overline{At}\ \overline{T_a}\ \overline{W}\ \overline{Hdd}$	$11 \rightarrow 1$	SVR/P	2,96	62,46 %	Tr:D-C-SS / Te:SC-SS
		$11 \rightarrow 1$	SVR/P	2,77	59,87 %	Tr:D-F-C-SS / Te:SC-SS

resultados, para luego aprovechar estos resultados con el objetivo de ser punto de partida para orientar el aprendizaje por transferencia.

En el contexto de los objetivos de esta investigación (ver capítulo 1), este aporte permite que pequeños y mediados productores puedan colaborar mutuamente para mejorar sus predicciones; esto a través de la aplicación de esta propuesta de aprendizaje por transferencia. Esta estrategia permite que conjuntos de datos de dominios diferentes, pero relacionados, puedan servir para iniciar un proceso simbiótico que permita un ganar-ganar entre el grupo de productores.

4.3. Propuesta en el proceso de reducción de atributos

La disminución en el número de atributos (variables) tiene dos objetivos fundamentales [69]: 1) disminuir el número de atributos de condición y 2) maximizar la información contenida en los atributos seleccionados. El logro de estos dos objetivos se ve reflejado en la mejora del tiempo de respuesta de los algoritmos de aprendizaje automático, esto al tener

que realizar menos comparaciones, reducir la cantidad de cálculos y eliminar variables que pueden generar ruido a tal punto que produzcan generalizaciones bajo supuestos incorrectos, todo por considerar variables que aportan poca información para la toma de decisiones o no son representativas del grupo de datos en estudio [69].

En la misma línea indicada en el párrafo anterior, la presente sección tiene como fin resaltar el aporte de la estrategia relativa a la reducción de atributos.

Como medio para mostrar la aplicación de la estrategia, se utiliza el proceso biológico de la floración del banano [7] [8].

4.3.1. Materiales y métodos

Los datos utilizados fueron proporcionados por CORBANA. En la tabla 4.25 se muestra la ubicación de cada finca y en qué periodo fueron recolectados los datos.

Tabla 4.25: Fincas utilizadas en el estudio (Caso: floración del banano)

Finca	Ubicación	Periodo (años)
28 Millas	Siquirres en Limón	2011 a 2014
Las Valquirias	Pococi en Limón	2010 a 2015

En cuanto a las variables disponibles, la tabla 4.26 muestra este detalle. La variable a predecir es el peso del racimo (BW).

En cada Finca, la variable BW es medida para cada uno de los cables de producción que se disponen. En el caso de 28 Millas, se utilizaron 7 cables, cada uno con una cantidad variable de semanas medidas, para un total de 799 observaciones. En el caso de Las Valquirias, se utilizaron 32 cables, para un total de 3839 observaciones. En ambos casos, la periodicidad utilizada es semanal.

Tabla 4.26: Variables disponibles (Caso: Floración del banano)

Símbolo	Descripción	Unidades
$\overline{T}_a$	Temperatura del aire promedio	[°C]
$\overline{R}$	Radiación solar promedio	[W/m^2]
P	Precipitación acumulada	[mm]
BW	Peso del racimo	kg

Las tablas 4.27 y 4.28 muestran estadísticas descriptivas de los conjuntos de datos en estudio.

El conjunto de técnicas utilizadas (T) fue: $\{SVR/L$, SVR/G, SVR/S, SVR/P, ENR, $OLSR\}$.

[7] Si se desean conocer los resultados detallados de los experimentos, los mismos se presentan en los Apéndices A y D.

[8] En cuanto al proceso biológico, éste fue descrito en la subsección 2.3.3.

Tabla 4.27: Estadísticas (Floración - 28 Millas)

Métrica	$\overline{T}_a$	$\overline{Sr}$	P	BW
Cardinalidad	799	799	799	799
Promedio	26.3	26.87	5.01	8.33
Mediana	26.37	27.65	0.5	7.2
Desviación estándar	1.15	9.38	9.82	4.98
Valor mínimo	22.23	0.17	0.0	2.8
Valor máximo	28.82	47.87	114.0	27.29
Rango	6.59	47.7	114.0	24.49
Coeficiente de variación	0.04	0.35	1.96	0.6

Tabla 4.28: Estadísticas (Conjunto de datos: Las Valquirias)

Métrica	$\overline{T}_a$	$\overline{Sr}$	P	BW
Cardinalidad	3839	3839	3839	3839
Promedio	26.3	25.19	10.0	6.2
Mediana	26.39	26.18	2.0	4.85
Desviación estándar	1.37	8.99	23.71	4.76
Valor mínimo	21.11	1.9	0.0	0.0
Valor máximo	29.14	44.06	183.0	37.85
Rango	8.03	42.16	183.0	37.85
Coeficiente de variación	0.05	0.36	2.37	0.77

Los métodos utilizados corresponden a los descritos en el Capítulo 3, subsecciones 3.2.8 (Determinación de patrones) y 3.3.1 (Generación de patrones).

4.3.2. Resultados y análisis

En la predicción de procesos biológicos en el campo agrícola, en que se desea conocer el efecto acumulado de ciertas variables sobre la variable a predecir, la presente estrategia colabora en el proceso de determinar la cantidad de observaciones previas que se requieren para lograr un nivel de predicción particular. Además, hay ocasiones en que se desea conocer el nivel de predicción no solo de uno, sino de varios periodos adelante. Por ejemplo, cuando por planificación de la producción y la logística relacionada con la siguiente etapa de comercialización, conocer solo la predicción de un periodo adelante, no es suficiente. Y junto a la determinación de las semanas previas y adelante, es relevante la determinación de los atributos que deben ser considerados para ese nivel de predicción.

En el caso de la floración del banano, interesa a los investigadores conocer los niveles de predicción con patrones desde 30 semanas previas y hasta 20 semanas adelante. Si bien la estrategia permite incrementos de un periodo en un periodo, solo para efectos del ejemplo se utilizaron incrementos de tres periodos en tres periodos ($p = 30$, $a = 20$, $inc = 3$).

Siguiendo la estrategia propuesta, se tiene:

- Determinación del método de entrenamiento y validación
  - MEV = ValidacionCruzada
- Determinación de la combinación de variables en A
  - $A = [\overline{T}_a, \overline{Sr}, P, BW]$
  - $N = [\overline{T}_a, \overline{Sr}, P]$
- Determinación de patrones
  - $p = 30$, $a = 20$, $inc = 3$
  - $Prev$ contiene 10 elementos [1,4,7,10,13,16,19,22,25,28]
  - $Adel$ contiene 7 elementos [1,4,7,10,13,16,19]
  - Al realizar el producto cartesiano vectorial entre $Prev$ y $Adel$, Pat contiene 70 elementos
- Generación de patrones
  - F contiene 7 matrices de patrones, a continuación se indican los vectores columna de los atributos incluidos en cada matriz: [$[\overline{T}_a, \overline{Sr}, P, BW]$, $[\overline{T}_a, \overline{Sr}, BW]$, $[\overline{T}_a, P, BW]$, $[\overline{Sr}, P, BW]$, $[\overline{T}_a, BW]$, $[\overline{Sr}, BW]$, $[P, BW]$]
  - Dado que la cantidad de elementos en Pat es 70 y la cantidad de matrices de patrones en F es 7, la cantidad de elementos en S es: $70 \cdot 7 = 490$

Se continúa la estrategia propuesta tal como se explica en el Capítulo 3 y se obtienen los resultados en la etapa de validación cruzada que se resumen en la tabla 4.29 para 28 Millas y la tabla 4.30 para Las Valquirias.

Los resultados obtenidos permiten al equipo investigador, particularmente a las personas expertas en el dominio, tomar decisiones sabiendo qué R^2 y qué $RMSE$ es el estimado; esto dependiendo de cuántas semanas previas y cuántas semanas adelante se requieren predecir.

Junto al promedio obtenido en cada una de estas métricas, se presenta la desviación estándar de las 10 corridas en la validación cruzada. Como es de esperar, a medida que se desea predecir más semanas adelante, el promedio del R^2 va disminuyendo y aumenta su desviación estándar y de manera congruente, el promedio del $RMSE$ va aumentando y la desviación estándar aumenta también. Lo anterior se explica por el aumento en la incertidumbre del resultado, al querer predecir más semanas adelante.

La figura 4.12 grafica el frente de Pareto de Las Valquirias y 28 Millas. Como se puede apreciar, Las Valquirias (en color azul) obtiene mejores valores en el frente de Pareto, mayores valores de R^2 con menores valores de $RMSE$. El frente de Pareto se enmarca en color rojo.

Las tablas 4.31 y 4.32 muestran un resumen de los resultados obtenidos en cuanto R^2 y $RMSE$ para la floración del banano en cada una de las fincas en estudio. En estas

Tabla 4.29: Configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para diferentes periodos adelante (a), en la etapa de validación cruzada para la floración del banano (28 Millas)

Variables	$p \rightarrow a$	**Técnica**	$RMSE$		R^2	
			mean	*stdev*	*mean*	*stdev*
$\overline{T}_a$	$16 \rightarrow 1$	SVR/G	0,98	0,44	95,28 %	4,52 %
Ta-P-Sr	$1 \rightarrow 1$	SVR/G	0,98	0,4	95,43 %	3,4 %
$\overline{Sr}$ P	$16 \rightarrow 1$	SVR/G	0,98	0,4	95,74 %	2,81 %
P	$13 \rightarrow 1$	SVR/S	0,99	0,44	95,85 %	2,55 %
$\overline{T}_a$ $\overline{Sr}$	$22 \rightarrow 4$	SVR/P	1,72	0,77	87,82 %	8,63 %
	$16 \rightarrow 7$	SVR/P	2,28	0,91	79,71 %	10,76 %
$\overline{T}_a$ P	$19 \rightarrow 10$	SVR/G	2,72	0,45	72,06 %	9,83 %
	$28 \rightarrow 13$	SVR/G	3,0	0,78	68,68 %	8,4 %
$\overline{Sr}$	$10 \rightarrow 16$	SVR/G	3,31	0,76	57,1 %	9,26 %
	$10 \rightarrow 16$	SVR/P	3,41	0,93	61,33 %	10,93 %
	$13 \rightarrow 16$	SVR/P	3,51	0,87	61,94 %	9,65 %
	$28 \rightarrow 16$	SVR/P	3,56	0,47	63,24 %	6,28 %
Ta-P-Sr	$16 \rightarrow 19$	SVR/G	3,22	0,7	64,15 %	10,32 %
$\overline{T}_a$ $\overline{Sr}$	$10 \rightarrow 19$	SVR/G	3,35	0,63	64,19 %	10,56 %

Tabla 4.30: Configuraciones del frente de Pareto en cuanto R^2 y $RMSE$, para diferentes periodos adelante (a), en la etapa de validación cruzada para la floración del banano (Las Valquirias)

Variables	$p \rightarrow a$	**Técnica**	$RMSE$		R^2	
			mean	*stdev*	*mean*	*stdev*
$\overline{Sr}$ P	$4 \rightarrow 1$	SVR/P	0,86	0,27	96,6 %	1,56 %
$\overline{Sr}$	$10 \rightarrow 4$	SVR/G	1,31	0,28	91,53 %	4,02 %
$\overline{Sr}$ P	$16 \rightarrow 4$	SVR/P	1,31	0,28	92,26 %	2,58 %
	$13 \rightarrow 7$	SVR/G	1,51	0,43	88,12 %	8,91 %
$\overline{T}_a$ $\overline{Sr}$	$13 \rightarrow 7$	SVR/G	1,53	0,25	89,62 %	3,45 %
$\overline{Sr}$	$13 \rightarrow 10$	SVR/G	1,68	0,24	86,81 %	2,36 %
$\overline{Sr}$ P	$10 \rightarrow 13$	SVR/G	1,93	0,21	82,84 %	4,25 %
	$7 \rightarrow 16$	SVR/G	2,06	0,27	80,31 %	3,91 %
Ta-P-Sr	$4 \rightarrow 19$	SVR/G	2,26	0,32	76,69 %	3,66 %

tablas se resaltan los resultados obtenidos al tomar como conjunto de entrenamiento una de las fincas y predecir con la otra. Se muestra que hacer validación cruzada con sus propios datos es mejor en ambos casos, lo cual era esperable, pero aún así, tomar como entrenamiento la otra finca da valores de R^2 superiores a 90 % en varias configuraciones.

Finalmente, respecto al aporte de la estrategia propuesta, se puede resumir en los siguientes aspectos:

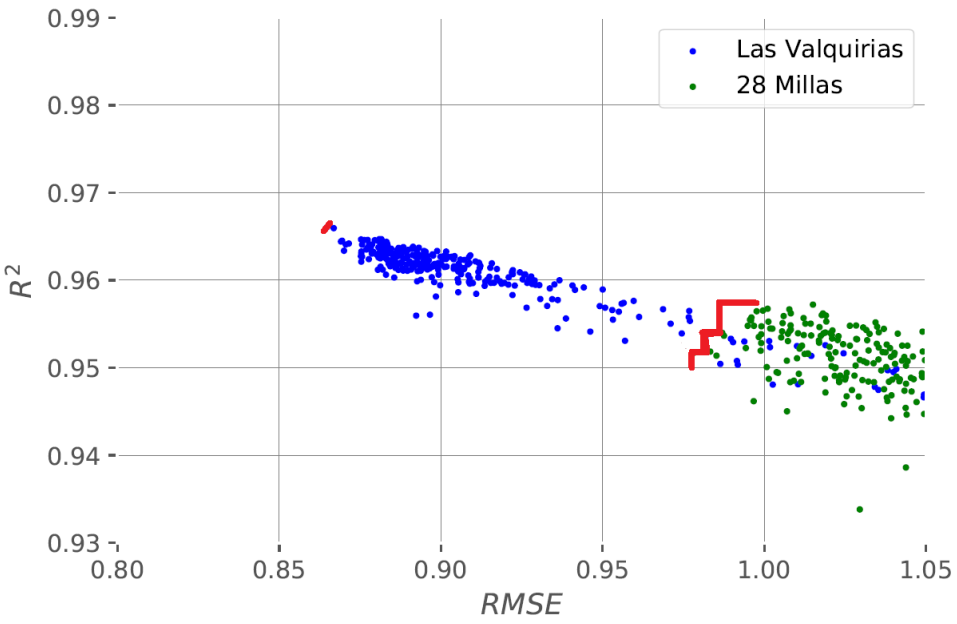


Figura 4.12: Frente de Pareto entre R^2 y $RMSE$ (Proceso biológico: Floración).

Tabla 4.31: Resumen de los resultados obtenidos en cuanto R^2 y $RMSE$ para la floración (Las Valquirias)

Variables	$p \rightarrow a$	**Técnica**	**RMSE**	R^2	**Experimento**
$\overline{Sr}$ P	$4 \rightarrow 1$	SVR/P	0,86	96,6 %	Validación cruzada
P	$13 \rightarrow 1$	SVR/S	0,93	96,22 %	Tr:28-SS / Te:LV-SS
Ta-P-Sr	$1 \rightarrow 1$	SVR/G	0,93	96,11 %	Tr:28-SS / Te:LV-SS
$\overline{Sr}$ P	$16 \rightarrow 1$	SVR/G	0,99	95,64 %	Tr:28-CS / Te:LV-SS
	$16 \rightarrow 1$	SVR/G	1,0	95,51 %	Tr:28-SS / Te:LV-SS
$\overline{T}_a$	$16 \rightarrow 1$	SVR/G	1,07	94,33 %	Tr:28-CS / Te:LV-SS
	$16 \rightarrow 1$	SVR/G	1,08	94,15 %	Tr:28-SS / Te:LV-SS
Ta-P-Sr	$1 \rightarrow 1$	SVR/G	1,75	88,4 %	Tr:28-CS / Te:LV-SS
P	$13 \rightarrow 1$	SVR/S	16,2	45,92 %	Tr:28-CS / Te:LV-SS

Reducción de atributos con miras a reducir el número de sensores y tomas de datos manuales necesarios para la predicción.

En el caso concreto del dominio al que va dirigida la presente propuesta, disminuir el número de atributos pretende: 1) dejar solo los atributos que aportan información y no son más bien generadores de ruido en el proceso, y 2) lograr reducir atributos significa en el fondo disminuir la cantidad de sensores y/o toma de datos manuales, lo cual normalmente va acompañado de una reducción de costos para el productor. Por lo anterior, las técnicas de reducción de atributos que proyectan los atributos a una dimensionalidad menor (tipo Análisis de Componentes Principales) no son preferidas, pues requerirían mantener la misma cantidad de sensores y/o toma de datos.

Ahora bien, como indica [116], la reducción de atributos se puede realizar con dos métodos: uno es llamado método filtro, en el cual el conjunto de atributos se filtra para producir

Tabla 4.32: Resumen de los resultados obtenidos en cuanto R^2 y $RMSE$ para la floración (28 Millas)

Variables	$p \rightarrow a$	Técnica	RMSE	R^2	Experimento
$\overline{T}_a$	$16 \rightarrow 1$	SVR/G	0,98	95,28 %	Validación cruzada
Ta-P-Sr	$1 \rightarrow 1$	SVR/G	0,98	95,43 %	Validación cruzada
$\overline{Sr}$ P	$16 \rightarrow 1$	SVR/G	0,98	95,74 %	Validación cruzada
P	$13 \rightarrow 1$	SVR/S	0,99	95,85 %	Validación cruzada
$\overline{Sr}$ P	$4 \rightarrow 1$	SVR/P	1,11	95,36 %	Tr:LV-CS / Te:28-SS
	$4 \rightarrow 1$	SVR/P	1,14	94,85 %	Tr:LV-SS / Te:28-SS

el subconjunto más prometedor antes de que comience el aprendizaje; y el el segundo se denomina método de envoltura, porque el algoritmo de aprendizaje está envuelto en el procedimiento de selección.

La estrategia propuesta considera ambos momentos del proceso de reducción de atributos. Con respecto al método filtro, en la subsección 3.2.7 (Determinación de la combinación de variables en A) se propone utilizar el criterio experto o una propuesta del mismo autor de esta investigación ([20]), donde se utilizan los conceptos de ganancia de información y conjuntos aproximados para la reducción de atributos. Y respecto al método envoltura, en la sección 3.4 (Etapa de entrenamiento y validación), aplicadas las técnicas y realizado el análisis estadístico, las configuraciones que conforman el frente de Pareto indican el conjunto de variables a seleccionar para cada configuración, lo que en el fondo es una recomendación concreta sobre la reducción de atributos.

Recomendación de la configuración de periodos previos (p) y periodos adelante (a) a utilizar, incluyendo la combinación de variables asociada, esto a partir de los elementos que conforman el frente de Pareto.

En la sección 3.2.8 (Determinación de patrones), el equipo investigador define los valores de las semanas previas (p), las semanas adelante (a) y los incrementos (inc) a experimentar; en la sección 3.3.1 (Generación de patrones) se producen los patrones a utilizar; y en la sección 3.4 (Etapa de entrenamiento y validación) se aplican las técnicas, se realiza el análisis estadístico y se determinan las configuraciones que conforman el frente de Pareto.

Las configuraciones que forman parte del frente de Pareto se convierten en una recomendación al equipo investigador de qué valores de p y a utilizar para alcanzar determinados valores de $RMSE$ y R^2.

Indicios de la relación entre los atributos de entrada y el de predicción.

Aunque no es el foco de la presente investigación, las configuraciones que conformen el frente de Pareto también dan indicios de la relación entre las variables independientes y la dependiente, por ejemplo, si la mayoría de configuraciones que conforman el frente de Pareto corresponden a técnicas como: $OLSR$ o ENR, esto sería un indicio que es probable que exista una relación de linealidad en el modelo.

Capítulo 5

Conclusiones

El creciente aumento en el número de investigaciones en inteligencia artificial, particularmente en el área del aprendizaje automático, ha permitido impactar en la solución de problemas de la vida cotidiana. Al respecto, un peligro latente es pretender que las propuestas de predicción sean aplicables para todos los casos y contextos [1], incluso a lo que Nassim Nicholas Taleb llama: cisnes negros [103]. Este autor propone que un cisne negro es una realidad que: 1) era muy difícil predecir antes de que sucediera, 2) tiene un gran impacto con su aparición, y 3) una vez que aparece, la naturaleza humana hace que aparezcan muchas explicaciones de por qué sucedió y de cómo predecirlo.

Por lo anterior, la presente investigación se centró en proponer una estrategia para la predicción en procesos biológicos del campo agrícola con datos limitados, indicando con claridad en qué casos se recomienda su aplicación.

Además, se privilegió proponer una solución para pequeños y medianos productores (sin excluir que grandes productores la puedan utilizar) antes que privilegiar obtener los mejores valores en métricas como $RMSE$ y R^2, pero a precio de requerir datos en volúmenes poco realistas de conseguir para el tipo de productores en cuestión, o cuyo costo de adquisición sería prohibitivo. Por ejemplo, variables de entrada como CO_2 pueden aportar información a procesos biológicos, pero el costo de este tipo de sensores es elevado. Otros ejemplos serían: el uso de imágenes multi o hiper espectrales, imágenes tomadas con cámaras de alta resolución, o imágenes satelitales en grandes cantidades, cuya inclusión podría aportar información relevante, pero cuyo costo de adquisición también suele ser prohibitivo.

Acorde al objetivo general de esta investigación, se ha propuesto una nueva estrategia de predicción en procesos biológicos del campo agrícola con datos limitados. Para ello se caracterizaron los casos de aplicación de la estrategia propuesta, en las cuales se precisó un conjunto de requisitos verificables que guíen al equipo investigador a determinar si la estrategia propuesta es una buena candidata, o no, para el pronóstico del proceso biológico que desean trabajar.

La arquitectura de aprendizaje automático propuesta es capaz de modelar los datos del

[1] Teorema del no hay almuerzo gratis [117].

proceso biológico en estudio, de guiar la creación del modelo particular y de preparar los datos para su uso.

Para validar el modelo propuesto, la estrategia proporciona una etapa que permite calcular las métricas determinadas por el equipo investigador con el fin de determinar si los resultados son de utilidad para su investigación.

Y finalmente, para aprender de las iteraciones de la aplicación de la estrategia, aunque toda la estrategia está basada en el concepto de aprendizaje automático, la gestión del Repositorio de Conocimiento Aprendido (RCA) es la principal respuesta a este objetivo.

5.1. Principales aportes

Las contribuciones más importantes de este trabajo pueden ser resumidas de la siguiente manera:

- Propone una estrategia esquemática que favorece la repetibilidad del proceso (ver capítulo 3).
- Delimita la aplicabilidad de la propuesta (ver secciones 3.1.1 y 3.1.2).
- No necesita predecir variables meteorológicas, sino que capta el efecto ya producido en el tiempo (ver capítulos 1 y 3).
- Propone un método de aumento de datos para mejorar la predicción de la estrategia (ver subsección 3.2.4).
- No requiere contar con imágenes para iniciar con la experimentación (ver capítulo 3).
- Propone una manera de trabajar con el espacio paramétrico de manera heurística (ver subsección 3.4.1).
- Usa el frente de Pareto entre el R^2 y el $RMSE$ para permitir una optimización multiobjetivo (ver subsección 3.4.3).
- Propone una estrategia de solución vista como aprendizaje supervisado en ventanas de tiempo deslizantes, en lugar del enfoque tradicional de predicción en series temporales (ver subsección 3.2.8).
- Aprovecha la propuesta de [68], enfocada en el escalamiento multidimensional para la visualización de datos de alta dimensionalidad. Agrega varios criterios de comparación y selección de resultados, para luego aprovechar estos resultados con el objetivo de ser punto de partida para orientar el aprendizaje por transferencia (ver subsección 3.2.5).

- Aporta en cuanto a la reducción de atributos: 1) en la selección previa de atributos que integren el conjunto de variables a estudiar (ver subsección 3.2.7 y [20]), 2) en el proceso de aprendizaje mismo, en el que con las configuraciones del frente de Pareto se hacen recomendaciones de los atributos a incluir en el modelo (ver sección 3.4), y 3) recomienda la configuración de periodos previos (p), periodos adelante (a) y variables a considerar, esto a partir de los elementos que conforman el frente de Pareto (ver secciones 3.2.8, 3.3.1 y 3.4). La reducción de atributos también pretende reducir costos, esto al lograr disminuir el número de sensores y tomas de datos necesarios para la predicción.

- Dar indicios de la relación entre los atributos de entrada y el de predicción en cuanto al tipo de relación entre dichos atributos (ver subsecciones 3.4.3 y 3.5.1).

5.2. Trabajo futuro

Al finalizar la presente investigación, quedan varias líneas de profundización abiertas para dar continuidad al trabajo iniciado, particularmente se destacan:

- Una dificultad que se presenta en la toma de datos manuales en el campo agrícola, es la subjetividad de quien registra los datos, de manera que de una toma de datos a otra, de una persona a otra, los criterios para definir un nivel de la variable pueden variar, por lo que es relevante generar procesos automatizados para minimizar la subjetividad en esta toma de datos. En esta línea de investigación van los trabajos de León Sarkis [62] (Un análisis comparativo de los algoritmos Fast Radial Symmetry Transform y Hough Transform para la detección automática de granos de café en imágenes) y de García Sanabria [40] (Desarrollo de un método de detección de Sigatoka negra utilizando atributos intrínsecos del huésped y el anfitrión por medio de técnicas de visión por computadora), las cuales, utilizando imágenes tomadas por medio de las cámaras de teléfonos celulares, determinan el nivel de un proceso biológico.

- Si se logra automatizar a un precio razonable la toma de datos descrita en el punto anterior, esto permitiría aumentar la periodicidad a utilizar en la estrategia. Dependiendo del aumento, se podrían incluir en el modelo otras técnicas que requieren conjuntos de datos mucho mayores que los disponibles actualmente, por ejemplo el aprendizaje profundo y técnicas de redes bayesianas. En este sentido se encuentran los trabajos de Mena Arias [74] (Evaluación del uso de distintas métricas de distancia de texto en un algoritmo agregado para la imputación de valores faltantes mediante clasificación), Alfaro Barboza [4] (Cubic Spline Interpolation como medida de distancia utilizada en el descubrimiento de reglas significativas en series temporales complejas y en presencia de ruido, y de Vallejos Peña [111] (Propuesta de algoritmo que combina el agrupamiento en subespacios basado en densidad y el agrupamiento basado en restricciones para la detección de grupos que incluyan

atributos de interés en conjuntos de datos de alta dimensionalidad), publicado en [19].

- La propuesta de aumento de datos requiere más investigación con el fin de contar con más criterios para su valoración. En particular, cuando la aplicación de la propuesta produce conjuntos de datos de 20 o más veces el tamaño del conjunto de datos inicial, se requiere más investigación sobre qué técnicas de aprendizaje automático podrían ya ser utilizadas. En esta línea de investigación está el trabajo de Argüello [9] (Evaluación del uso de Redes Bayesianas Dinámicas para la predicción del avance de la Sigatoka negra y la productividad en cultivos agrícolas), publicado en [18].

- En las predicciones de procesos biológicos en el campo agrícola, es muy común utilizar variables de tipo meteorológico. A la vez, es frecuente que este tipo de variables presente valores faltantes e incluso valores atípicos, los primeros por mal funcionamiento de los sensores o del medio de comunicación de los datos, y los segundos por mal funcionamiento del equipo o afectaciones externas al sensor, como cuando un objeto o ser vivo cae sobre algún sensor y obstaculiza su funcionamiento. Es por esto que promover investigaciones en los siguientes campos sería importante para unir con la presente investigación: 1) la imputación de datos, en todas sus modalidades, univariable y multivariable, simple y múltiple, 2) la interpolación espacial de datos y 3) la detección de valores atípicos.

- Los campos de la reducción de atributos y la selección de patrones de periodos antes y periodos después en la predicción, como se hace en la presente investigación, presentan oportunidades para incorporar otras técnicas de aprendizaje automático (como por ejemplo, redes bayesianas y aprendizaje profundo). Además, como se indicó en la sección 4.3 (Propuesta en el proceso de reducción de atributos), el tema de buscar una relación entre las técnicas que prevalezcan en las configuraciones del frente de Pareto y la estructura de los datos del proceso biológico, es un tema abierto a la profundización.

- Los actuales resultados podrán ser complementados con otras investigaciones —en particular con las que consideran imágenes— con el fin de plasmar estos resultados en una aplicación que quede al servicio de los productores. Aplicación que podrá colaborar con sus usuarios para aumentar la productividad de sus cultivos, manteniendo un nivel de costos razonable al utilizar las predicciones obtenidas de los procesos biológicos en estudio.

Bibliografía

[1] A. T. M. S. Ahamed, N. T. Mahmood, N. Hossain, M. T. Kabir, K. Das, F. Rahman y R. M. Rahman, «Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh», en *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, jun. de 2015, págs. 1-6. DOI: 10.1109/SNPD.2015.7176185.

[2] D. Al Bashish, M. Braik y S. Bani-Ahmad, «A framework for detection and classification of plant leaf and stem diseases», *Proceedings of the 2010 International Conference on Signal and Image Processing, ICSIP 2010*, págs. 113-118, dic. de 2010. DOI: 10.1109/ICSIP.2010.5697452.

[3] A. Albert y L. Zhang, «A novel definition of the multivariate coefficient of variation», *Biometrical Journal*, vol. 52, n.º 5, págs. 667-675, 2010. DOI: 10.1002/bimj.201000030.

[4] D. E. Alfaro Barboza, «Cubic Spline Interpolation como medida de distancia utilizada en el descubrimiento de reglas significativas en series temporales complejas y en presencia de ruido», Tesis de Maestría, Instituto Tecnológico de Costa Rica, Escuela de Ingeniería en Computación, Maestría Académica en Ciencias de la Computación, 2017.

[5] J. Alonso, Á. R. Castañón y A. Bahamonde, «Support Vector Regression to predict carcass weight in beef cattle in advance of the slaughter», *Computers and Electronics in Agriculture*, vol. 91, págs. 116-120, feb. de 2013. DOI: 10.1016/j.compag.2012.08.009.

[6] M. A. Alvarado, R. Foroughbakhch, E. Jurado y A. Rocha, «El cambio climático y la fenología de las plantas», *Ciencia UANL*, vol. V, n.º 5, págs. 493-500, 2002.

[7] J. Amara, B. Bouaziz y A. Algergawy, «A Deep Learning-based Approach for Banana Leaf Diseases Classification», *BTW Workshop*, págs. 79-88, mar. de 2017.

[8] J. A. Aranda Pinilla y J. A. Orjuela Castro, «Optimización multiobjetivo en la gestión de cadenas de suministro de biocombustibles. Una revisión de la literatura», *Ingeniería*, vol. 20, n.º 1, págs. 37-63, 2015.

[9] S. Argüello, «Diseño de una herramienta basada en Redes Bayesianas Dinámicas para la predicción del avance de la Sigatoka Negra y la productividad en cultivos agrícolas», Tesis de Maestría, Instituto Tecnológico de Costa Rica, Escuela de Ingeniería en Computación, Maestría Académica en Ciencias de la Computación, 2017.

[10] T. O. Ayodele, «Types of machine learning algorithms», *New Advances in Machine Learning*, n.º 17, pág. 446, 2010. DOI: 10.5772/9385.

[11] R. T. Baillie, «Robust Inference in Time Series Regressions: Limitations and Feasible GLS Alternatives», *Rimini Center for Economic Analysis*, págs. 1-39, jun. de 2017.

[12] N. Balakrishnan y G. Muthukumarasamy, «Crop Production - Ensemble Machine Learning Model for Prediction», *International Journal of Computer Science and Software Engineering*, vol. 5, n.º 7, págs. 148-153, 2016.

[13] M. Barquero Miranda, *Recomendaciones para el combate de la roya del cafeto*, 3ra. San José, Costa Rica: ICAFE, 2013.

[14] H. Bendini, W. Moraes, da S. Silva, E. Tezuka y P. Cruvinel, «Análise de risco da ocorrência de Sigatoka-negra baseada em modelos polinomiais: um estudo de caso», *Tropical Plant Pathology*, vol. 38, n.º 1, págs. 35-43, 2013.

[15] H. N. Bendini, W. S. Moraes, S. S. Da Costa, E. S. Lopes, T. S. Körting y L. M. Fonseca, «Proposta de Sistema de monitoramento da sigatoka-negra baseado em variáveis ambientais utilizando o TerraMA2», en *Proceedings of the Brazilian Symposium on GeoInformatics*, vol. 15, dic. de 2014, págs. 168-173.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[17] G. C. Bora, D. Lin, P. Bhattacharya, S. K. Bali y R. Pathak, «Application of Bio-Image Analysis for Classification of Different Ripening Stages of Banana», *Journal of Agricultural Science*, vol. 7, n.º 2, págs. 152-160, 2015. DOI: 10.5539/jas.v7n2p152.

[18] L. A. Calvo-Valverde, S. Argüello y M. Guzmán-Quesada, «Evaluación del uso de Redes Bayesianas Dinámicas para la predicción del avance de la Sigatoka negra y la productividad en cultivos agrícolas Evaluation of Dynamic Bayesians Networks for predicting the progress of the Black Sigatoka and the productivity in crops», *Revista Tecnología en Marcha*, vol. 32, págs. 158-170, 2019.

[19] L. A. Calvo-Valverde y A. Vallejos-Peña, «Algoritmo semisupervisado de agrupamiento que combina SUBCLU y el agrupamiento basado en restricciones, para la detección de grupos en conjuntos de alta dimensionalidad», *Revista Tecnología en Marcha*, vol. 31, n.º 3, págs. 74-85, 2018. DOI: 10.18845/tm.v31i3.3904.

[20] L.-A. Calvo-Valverde, «Estrategia basada en el aprendizaje de máquina para tratar con conjuntos de datos no etiquetados usando conjuntos aproximados y/o ganancia de información», *Tecnología en marcha - Edición especial - Matemática aplicada*, págs. 4-15, 2016.

[21] L.-A. Calvo-Valverde, M. Guzman Quesada, J.-A. Guzmán Alvares y P. Alvarado-Moya, en *Abstracts of Presentations at the 56th Annual Meeting of the APS Caribbean Division*, PMID: 28665239, vol. 107, 2017, S4.7-S4.21. DOI: 10.1094/PHYTO-107-7-S4.7.

[22] A. Camargo, J. Molina, J. Cadena-Torres, N. Jiménez y J. Kim, «Intelligent systems for the assessment of crop disorders», *Computers and Electronics in Agriculture*, vol. 85, págs. 1-7, 2012. DOI: 10.1016/j.compag.2012.02.017.

[23] J. Champion, *El banano: Técnicas agrícolas de producciones tropicales*. Barcelona, España: Editorial Blume, 1976.

[24] A. Chlingaryan, S. Sukkarieh y B. Whelan, «Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review», *Computers and Electronics in Agriculture*, vol. 151, n.º May, págs. 61-69, 2018. DOI: 10.1016/j.compag.2018.05.012.

[25] T. Chuang y M. Jeger, «Predicting the Rate of Development of Black Sigatoka (*Mycosphaerella fijiensis var. difformis*) Disease in Southern Taiwan», *Phytopathology*, vol. 77, págs. 1542-1547, 1987.

[26] P. de Comercio Exterior de Costa Rica (PROCOMER). (2017). Portal oficial estadístico de comercio exterior, dirección: http://servicios.procomer.go.cr/estadisticas/inicio.aspx.

[27] E. J. Coopersmith, B. S. Minsker, C. E. Wenzel y B. J. Gilmore, «Machine learning assessments of soil drying for agricultural planning», *Computers and Electronics in Agriculture*, vol. 104, págs. 93-104, 2014. DOI: 10.1016/j.compag.2014.04.004.

[28] D. C. Corrales, «Toward detecting crop diseases and pest by supervised learning», *Ingenieria y Universidad*, vol. 19, n.º 1, pág. 207, 2015. DOI: 10.11144/Javeriana.iyu19-1.tdcd.

[29] D. Corrales, A. Figueroa A.and Ledezma y J. Corrales, «An Empirical Multi-classifier for Coffee Rust Detection in Colombian Crops», *Computational Science and Its Applications*, vol. 9155, págs. 60-74, 2015.

[30] R. D'Agostino y E. S. Pearson, «Tests for departure from normality», *Biometrika*, vol. 60, págs. 613-622, 1973.

[31] M. C. De Alves, L. G. De Carvalho, E. a. Pozza, L. Sanches y J. C. S. De Maia, «Ecological zoning of soybean rust, coffee rust and banana black Sigatoka based on Brazilian climate changes», *Procedia Environmental Sciences*, vol. 6, págs. 35-49, 2011. DOI: 10.1016/j.proenv.2011.05.005.

[32] B. Demir y L. Bruzzone, «A multiple criteria active learning method for support vector regression», *Pattern Recognition*, págs. 2558-2567, 2014. DOI: 10.1016/j.patcog.2014.02.001.

[33] P. Filippi, E. J. Jones, N. S. Wimalathunge, P. D. Somarathna, L. E. Pozza, S. U. Ugbaje, T. G. Jephcott, S. E. Paterson, B. M. Whelan y T. F. Bishop, «An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning», *Precision Agriculture*, n.º 20, 2019. DOI: 10.1007/s11119-018-09628-4.

[34] F.-A. Fortin, François-Michel, M.-A. Gardner, M. Parizeau y C. Gagné, «DEAP: Evolutionary Algorithms Made Easy», *Journal of Machine Learning Research*, vol. 13, págs. 2171-2175, jul. de 2012.

[35] E. Fouré, «Black leaf Streak disease of bananas and plantains (*Mycosphaerella fijiensis Morelet*). Study of the symptoms and stages of the disease in Gabon.», *IRFA*, 1985.

[36] ——, «Stratégies de lutte contre la cercosporioses noire des bananiers et plantains provoquée par *Mycosphaerella fijiensis Morelet*. Lávertissement biologique au Cameroun. Evaluation des possibilités d'amélioration.», *Fruits*, vol. 43, n.º 5, págs. 269-274, 1988.

[37] S. L. France y J. D. Carroll, «Two-way multidimensional scaling: A review», *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 41, n.º 5, págs. 644-661, 2011. DOI: 10.1109/TSMCC.2010.2078502.

[38] J. Ganry y E. Laville, «La lutte controlée contre le Cercospora aux Antilles. Bases climatiques de lávertissement», *Fruits*, vol. 27, págs. 665-676, 1972.

[39] J. Ganry y J. Meyer, «Les cerco-sporioses du bananier et leurs traitements. Evolution des méthods de traitement», *Fruits*, vol. 38, págs. 3-20, 1983.

[40] A. García Sanabria, «Desarrollo de un método de detección de Sigatoka negra utilizando atributos intrínsecos del huésped y el anfitrión por medio de técnicas de visión por computadora», Tesis de Maestría, Instituto Tecnológico de Costa Rica, Escuela de Ingeniería en Computación, Maestría Académica en Ciencias de la Computación, 2019.

[41] T. Glezakos, G. Moschopoulou, T. Tsiligiridis, S. Kintzios y C. Yialouris, «Plant virus identification based on neural networks with evolutionary preprocessing», *Computers and Electronics in Agriculture*, vol. 70, págs. 263-275, 2010.

[42] M. González, J. A. Guzmán y F. Blanco, «Efecto del clima sobre la fenología del cultivo del banano (Musa AAA)», en *V Congreso Internacional sobre banano*, C. S.A., ed., vol. V, 2014.

[43] D. Grenón, *Aplicaciones informáticas en la empresa agropecuaria. PNATTI. Subsecretaría de Informática y Desarrollo*. Buenos Aires, 1984.

[44] X. Gu, R. T. Kwok, J. W. Lam y B. Z. Tang, «AIEgens for biological process monitoring and disease theranostics», *Biomaterials*, vol. 146, págs. 115-135, 2017. DOI: 10.1016/j.biomaterials.2017.09.004.

[45] M. Guzmán, «Control biológico y cultural de la sigatoka- negra», *Tropical Plant Pathology*, pág. 4, sep. de 2012. DOI: 10.13140/2.1.2927.7442.

[46] G. Heiman, *Understanding Research Methods and Statistics*. Cengage Learning, Inc, oct. de 2002.

[47] T. Hengl, *A Practical Guide to Geostatistical Mapping*, 2.ª ed. Office for Official Publications of the European Communities, sep. de 2009.

[48] S. Hernández, «Del diseño convencional al diseño óptimo. Posibilidades y variantes», *Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería*, vol. 9, págs. 259-270, 1993.

[49] J. Holloway y K. Mengersen, «Statistical Machine Learning Methods and Remote Sensing for Sustainable Development Goals: A Review», *Remote Sensing*, vol. 10, n.º 9, pág. 1365, 2018. DOI: 10.3390/rs10091365.

[50] Y. Huang, Y. Lan, S. Thomson, A. Fang, W. Hoffmann y R. Lacey, «Development of soft computing and applications in agricultural and biological engineering», *Computers and Electronics in Agriculture*, vol. 71, n.º 2, págs. 107-127, 2010. DOI: 10.1016/j.compag.

[51] G. Humphries, D. R. Magness y F. Huettmann, *Machine Learning for Ecology and Sustainable Natural Resource Management*. Switzerland: Springer, 2018.

[52] W. Hussain, F. K. Hussain, M. Saberi, O. K. Hussain y E. Chang, «Comparing time series with machine learning-based prediction approaches for violation management in cloud SLAs», *Future Generation Computer Systems*, vol. 89, págs. 464-477, 2018. DOI: 10.1016/j.future.2018.06.041.

[53] N. Ibrahim y A. Wibowo, «Time Series Support Vector Regression with Missing Data Treatment Based Variables Selection for Water Level Prediction of Galas River in Kelantan Malaysia.», *International Journal of Applied Research in Engineering and Science*, vol. 3, págs. 25-36, 2014.

[54] R. Ihaka, *Time Series Analysis*. Statistics Department University of Auckland, abr. de 2005.

[55] Instituto del Café de Costa Rica, *Modelo de costos de producción agrícola de café fruta cosecha 2018-2019*, 2019.

[56] ——, *Modelo de costos de renovación de cafetales cosecha 2018-2019*, 2019.

[57] M. K P y V. P. CH, «Role of image processing and machine learning techniques in disease recognition, diagnosis and yield prediction of crops: A review», *International Journal of Advanced Research in Computer Science*, vol. 9, n.º 2, págs. 788-795, 2018.

[58] Y. Kim, S. Yoo, Y. Gu, J. Lim, D. Han y S. Baik, «Crop Pests Prediction Method Using Regression and Machine Learning Technology: Survey», *IERI Procedia*, vol. 6, págs. 52-56, 2014. DOI: 10.1016/j.ieri.2014.03.009.

[59] L. Konstantinos, P. Busato, D. Moshou, S. Pearson y D. Bochtis, «Machine Learning in Agriculture: A Review», *Sensors*, vol. 18, n.º 8, pág. 2674, 2018. DOI: 10.3390/s18082674.

[60] W. H. Kruskal y W. A. Wallis, «Use of Ranks in One-Criterion Variance Analysis», *Journal of the American Statistical Association*, vol. 47, n.º 260, págs. 583-621, 1952.

[61] E. Lasso, T. T. Thamada, C. A. Alves y J. C. Corrales, «Graph Patterns as Representation of Rules Extracted from Decision Trees for Coffee Rust Detection», *Communications in Computer and Information Science*, vol. 544, págs. 405-414, 2015. DOI: 10.1007/978-3-319-24129-6_35.

[62] M. León Sarkis, «Un análisis comparativo de los algoritmos Fast Radial Symmetry Transform y Hough Transform para la detección automática de granos de café en imágenes», Tesis de Maestría, Instituto Tecnológico de Costa Rica, Escuela de Ingeniería en Computación, Maestría Académica en Ciencias de la Computación, 2017.

[63] C. Li, Y. Yang y S. Liu, «A new method to mitigate data fluctuations for time series prediction», *Applied Mathematical Modelling*, vol. 65, págs. 390-407, 2019. DOI: 10.1016/j.apm.2018.08.017.

[64] G. Lobet, «Image Analysis in Plant Sciences: Publish Then Perish», *Trends in Plant Science*, vol. 22, n.º 7, págs. 559-566, 2017.

[65] T. M. Logan, S. McLeod y S. Guikema, «Predictive models in horticulture: A case study with Royal Gala apples», *Scientia Horticulturae*, vol. 209, págs. 201-213, 2016. DOI: 10.1016/j.scienta.2016.06.033.

[66] O. Luaces, L. H. a. Rodrigues, C. A. Alves Meira y A. Bahamonde, «Using nondeterministic learners to alert on coffee rust disease», *Expert Systems with Applications*, vol. 38, n.º 11, págs. 14 276-14 283, 2011. DOI: 10.1016/j.eswa.2011.05.003.

[67] L. van der Maaten, «Accelerating t-SNE using Tree-Based Algorithms», *Journal of Machine Learning Research*, vol. 15, págs. 1-21, 2014.

[68] L. van der Maaten y G. Hinton, «Visualizing Data using t-SNE», *Journal of Machine Learning Research*, vol. 9, págs. 2579-2605, 2008.

[69] P. Mahajan, «Rough Set Approach in Machine Learning : A Review», *International Journal of Computer Applications*, vol. 56, n.º 10, págs. 1-13, 2012.

[70] D. Marín Vargas y R. Romero Calderón, «El combate de la Sigatoka Negra», en *Boletín Departamento de Investigaciones*, Costa Rica: Corbana, 1995, págs. 1-23.

[71] D. Marín, R. Romero, M. Guzmán y T. Sutton, «Black Sigatoka: An increasing threat to banana cultivation», *Plant Disease*, vol. 87, n.º 3, págs. 208-222, 2003.

[72] R. T. Marler y J. S. Arora, «Survey of multi-objective optimization methods for engineering», *Structural and multidisciplinary optimization*, vol. 26, n.º 6, págs. 369-395, 2004.

[73] K. P. Mayuri y C. Vani Priya, «Role of image processing and machine learning techniques in disease recognition, diagnosis and yield prediction of crops: A review», *International Journal of Advanced Research in Computer Science*, vol. 9, n.º 2, págs. 788-795, 2018.

[74] J. A. Mena Arias, «Evaluación del uso de distintas métricas de distancia de texto en un algoritmo agregado para la imputación de valores faltantes mediante clasificación», Tesis de Maestría, Instituto Tecnológico de Costa Rica, Escuela de Ingeniería en Computación, Maestría Académica en Ciencias de la Computación, 2017.

[75] A. Mizushima y R. Lu, «An image segmentation method for apple sorting and grading using support vector machine and Otsu's method», *Computers and Electronics in Agriculture*, vol. 94, págs. 29-37, jun. de 2013. DOI: 10.1016/j.compag.2013.02.009.

[76] H. Y. Montero González, «Simulación de la floración del cultivo del banano (Musa AAA cv. Grande Naine) mediante el modelo SIMBA-CR adaptado a la vertiente Caribe de Costa Rica», Tesis de Licenciatura, Facultad de Ciencias Agroalimentarias, Turrialba, Costa Rica, 2016.

[77] D. C. Montgomery, *Design and analysis of experiments*, 8.ª ed. New York, USA: John Wiley & Sons, Inc., 2001.

[78] L. Mozaffari, A. Mozaffari y N. L. Azad, «Vehicle speed prediction via a sliding-window time series analysis and an evolutionary least learning machine: A case study on San Francisco urban roads», *Engineering Science and Technology, an International Journal*, vol. 18, n.º 2, págs. 150-162, 2015. DOI: 10.1016/j.jestch.2014.11.002.

[79] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Massachusetts, USA: The MIT Press, 2012.

[80] F. Organización de las Naciones Unidas para la Alimentación y la Agricultura, *El estado mundial de la agricultura y la alimentación*. 2012.

[81] ——, *E-Agriculture in action*, Bangkok, 2017.

[82] ——, *Information and Communication Technology (ICT) in Agriculture - A Report to the G20 Agricultural Deputies*, Rome, 2017.

[83] ——, *The state of food and agriculture - Leverangin food systems for inclusive rural transformation*, Rome, 2017.

[84] ——, *Climate smart agriculture - Building resilience to climate change*, Switzerland, 2018.

[85] ——, (2018). FAOSTAT - Sitio oficial de estadísticas de la FAO, dirección: http://www.fao.org/faostat/es/#data.

[86] 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 y E. Duchesnay, «Scikit-learn: Machine Learning in Python», *Journal of Machine Learning Research*, vol. 12, págs. 2825-2830, 2011.

[87] C. Perez-Ariza, A. Nicholson y M. Flores, «Prediction of coffee rust disease using Bayesian networks», English, en *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models*, Üniversity of Granada", 2012, págs. 259-266.

[88] W. H. Press, S. A. Teukolsky, W. T. Vetterling y B. P. Flannery, *Numerical Recipes. The Art of Scientific Computing*, Third. Cambridge University Press, 2007.

[89] S. M. Ramírez e Y. Q. Ballestero, «El combate de la Sigatoka Negra», en *Boletín estadístico agropecuario*, Costa Rica: Secretaría Ejecutiva de Planificación Sectorial Agropecuaria (SEPSA), 2017, págs. 1-230.

[90] P. Ramos, F. Prieto, E. Montoya y C. Oliveros, «Automatic fruit count on coffee branches using computer vision», *Computers and Electronics in Agriculture*, vol. 137, págs. 9-22, 2017. DOI: 10.1016/j.compag.2017.03.010.

[91] I. Ravi y M. M. Mustaffa, «Adaptation and mitigation strategies for climate-resilient horticulture», en *Climate-Resilient Horticulture: Adaptation and Mitigation Strategies*, H. S. et al., ed., 2013, págs. 1-12. DOI: 10.1007/978-81-322-0974-4_1.

[92] R. Romero Calderón, «Dynamics of fungicide resistant populations of *Mycosphaerella fijiensis* and Epidemiology of black Sigatoka of banana», Tesis doct., Department of Plant Pathology, North Carolina State University, 1995.

[93] T. Rumpf, A.-K. Mahlein, U. Steiner, E.-C. Oerke, H.-W. Dehne y L. Plümer, «Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance», *Computers and Electronics in Agriculture*, vol. 74, n.º 1, págs. 91-99, 2010. DOI: 10.1016/j.compag.2010.06.009.

[94] K. Saldana Ochoa y Z. Guo, «A framework for the management of agricultural resources with automated aerial imagery detection», *Computers and Electronics in Agriculture*, vol. 162, págs. 53-69, feb. de 2019. DOI: 10.1016/j.compag.2019.03.028.

[95] A. L. Samuels, «Some studies in machine learning using game of checkers», *IBM Journal of Research and Development*, vol. 3, n.º 3, págs. 210-229, 1959.

[96] A. Sanaeifar, A. Bakhshipour y M. de la Guardia, «Prediction of banana quality indices from color features using support vector regression», *Talanta*, vol. 148, págs. 54-61, 2016. DOI: 10.1016/j.talanta.2015.10.073.

[97] F. B. de Santana, A. M. de Souza y R. J. Poppi, «Green methodology for soil organic matter analysis using a national near infrared spectral library in tandem with learning machine», *Science of the Total Environment*, vol. 658, págs. 895-900, 2019. DOI: 10.1016/j.scitotenv.2018.12.263.

[98] V. Singha y A. Misrab, «Detection of plant leaf diseases using image segmentation and soft computing techniques», *Information Processing in Agriculture*, vol. 4, n.º 1, págs. 41-49, 2017.

[99] A. J. Smola y B. Schölkopf, «A tutorial on support vector regression», *Statistics and Computing*, vol. 14, págs. 199-222, 2004.

[100] G. W. Snedecor y W. G. Cochran, *Statistical Methods, Eighth Edition*. USA: Iowa State University Press., 1989.

[101] J. Soares, M. Pasqual y W. Lacerda, «Utilization of artificial neural networks in the prediction of the bunches' weight in banana plants», *Scientia Horticulturae*, vol. 155, págs. 24-29, 2013.

[102] J. Soares, M. Pasqual, W. Lacerda, S. Silva y S. Donato, «Comparison of techniques used in the prediction of yield in banana plants», *Scientia Horticulturae*, vol. 167, págs. 84-90, 2014.

[103] N. N. Taleb, *The Black Swan: The Impact of the Highly Improbable*. New York: Random House Publishing Group, 2010.

[104] E. Ternesien, *La cercorporioses des bananiers et plantains. Methodes de lutte-Avertissements. Perspectives au Cameroun. Memoire de fin d'etudes*, Paris, 1985.

[105] T. T. Thamada, L. H. A. Rodrigues y C. A. A. Meira, «Predição da taxa de progresso da ferrugem do cafeeiro por meio de ensembles», *IX Simpósio de Pesquisa dos Cafés do Brasil*, págs. 1-4, 2015.

[106] R. Tibshirani, «Regression shrinkage and selection via the lasso», *Journal of the Royal Statistical Society: Series B*, vol. 58, n.º 1, págs. 267-288, 1996.

[107] J. Treboux y D. Genoud, «Improved machine learning methodology for high precision agriculture», *2018 Global Internet of Things Summit, GIoTS 2018*, págs. 1-6, 2018. DOI: 10.1109/GIOTS.2018.8534558.

[108] D. Triantakonstantis y S. Barr, «An Empirical Multi-classifier for Coffee Rust Detection in Colombian Crops David», vol. 3, págs. 221-236, 2009. DOI: 10.1007/978-3-319-21413-9.

[109] N. Unidas, *Report of the World Commission on Environment and Development - Our Common Future*. 1987.

[110] R. Vaitheeshwari y V. Sathieshkumar, «Performance analysis of epileptic seizure detection system using neural network approach», *ICCIDS 2019 - 2nd International Conference on Computational Intelligence in Data Science, Proceedings*, págs. 1-5, oct. de 2019. DOI: 10.1109/ICCIDS.2019.8862158.

[111] A. Vallejos Peña, «Propuesta de algoritmo que combina el agrupamiento en subespacios basado en densidad y el agrupamiento basado en restricciones para la detección de grupos que incluyan atributos de interés en conjuntos de datos de alta dimensionalidad», Tesis de Maestría, Instituto Tecnológico de Costa Rica, Escuela de Ingeniería en Computación, Maestría Académica en Ciencias de la Computación, 2017.

[112] M. Wattenberg, F. Viégas e I. Johnson, «How to Use t-SNE Effectively», *Distill*, 2016. DOI: 10.23915/distill.00002.

[113] Z. Wei, T. Tao, D. ZhuoShu y E. Zio, «A dynamic particle filter-support vector regression method for reliability prediction.», *Reliability Engineering & System Safety*, págs. 109-116, nov. de 2013. DOI: doi:10.1016/j.ress.2013.05.021.

[114] C. White, *Strategic Managemet*. New York, USA: Palgrave MacMillan, 2004.

[115] F. Wilcoxon, «Individual Comparisons by Ranking Methods», *International Biometric Society*, vol. 1, págs. 80-83, 1945. DOI: 10.2307/3001968.

[116] I. H. Witten, F. Eibe y M. A. Hall., *Data Mining*, Third. United States: Morgan Kaufmann Publishers, 2011.

[117] D. H. Wolpert y W. G. Macready, «No free lunch theorems for optimization», *IEEE Transactions on Evolutionary Computation*, vol. 1, n.º 1, págs. 67-82, 1997. DOI: 10.1109/4235.585893.

[118] B. Ye, J. Chen, C. Ju, H. Li y X. Wang, «Distinguishing chaotic time series from noise: A random matrix approach», *Communications in Nonlinear Science and Numerical Simulation*, 2016. DOI: 10.1016/j.cnsns.2016.08.018.

[119] R. Ye y Q. Dai, «A novel transfer learning framework for time series forecasting», *Knowledge-Based Systems*, vol. 156, págs. 74-99, dic. de 2018. DOI: 10.1016/j.knosys.2018.05.021.

[120] H. Zou y T. Hastie, «Regularization and variable selection via the elastic net», *Journal of the Royal Statistical Society: Series B*, vol. 67, págs. 301-320, 2005.

Apéndice A

Ejemplo detallado: floración del banano

Este caso de estudio mostrará la aplicación de la estrategia con la RCA vacía y la inclusión del proceso biológico de la floración del banano, medida por el peso de los racimos, lugar 28 Millas (Siquirres, Limón, Costa Rica).

1. **Etapa preliminar**

   - Proceso biológico: floración del banano
   - Tipo de aprendizaje: supervisado, regresión
   - Variable a pronosticar: peso del racimo
   - Lugar: 28 Millas, (Siquirres, Limón, Costa Rica)
   - RCA:
     - VCA: sin elementos
     - EPB: sin elementos
     - CA: sin elementos

2. **Etapa de creación del experimento**

   2.1. Creación del *epb*

   - *epb*:
     - *id_epb*: epb00001
     - *descripcion*: Experimentación con datos de Corbana sobre la floración de banano por peso neto kilo del racimo, finca 28 Millas
     - *ca_estudio*: ND
     - *variables*: ND
     - C: ND
     - Pat: ND

- T: ND
- T': ND
- E: ND
- O: ND
- M: ND
- MEV: ND
- S: ND
- R: ND
- AE: ND
- U: ND
- *observaciones*: ND

2.2. Creación de un nuevo *ca*

- *vca*:
  - *id_vca*: vca00001Ta
  - *descripcion*: Promedio de la temperatura del aire
  - *mostrar_como*: $\overline{T}_a$
  - *unidad*: [°C]
  - *origenes*:
    - *id_vca_dat*: vcadat00001
    - *id_vca*: vca00001Ta
    - *origen*: 28 Millas - Siquirres, Limón, Costa Rica
    - *datos*: Acá vendrían los datos, no se incluyen en detalle para guardar la privacidad de los mismos, pero son series semanales entre los años 2011 y 2015, 799 registros en total
  - $filtro$: *promedio*
  - *metodo_imputacion*: spline
- *vca*:
  - *id_vca*: vca00002Sr
  - *descripcion*: Promedio de la radiación solar
  - *mostrar_como*: $\overline{Sr}$
  - *unidad*: [W/m^2]
  - *origenes*:
    - *id_vca_dat*: vcadat00002
    - *id_vca*: vca00002Sr
    - *origen*: 28 Millas - Siquirres, Limón, Costa Rica
    - *datos*: Acá vendrían los datos, no se incluyen en detalle para guardar la privacidad de los mismos, pero son series semanales entre los años 2011 y 2015, 799 registros en total

- *filtro*: *promedio*
- *metodo_imputacion*: spline

- *vca*:
  - *id_vca*: vca00003P
  - *descripcion*: Precipitación acumulada
  - *mostrar_como*: P
  - *unidad*: [mm]
  - *origenes*:
    - *id_vca_dat*: vcadat00003
    - *id_vca*: vca00003P
    - *origen*: 28 Millas - Siquirres, Limón, Costa Rica
    - *datos*: Acá vendrían los datos, no se incluyen en detalle para guardar la privacidad de los mismos, pero son series semanales entre los años 2011 y 2015, 799 registros en total
  - *filtro*: *suma*
  - *metodo_imputacion*: spline
  - *vca*:
    - *id_vca*: vca00004BW
    - *descripcion*: Peso del racimo
    - *mostrar_como*: BW
    - *unidad*: kg
    - *origenes*:
      - *id_vca_dat*: vcadat00004
      - *id_vca*: vca00004BW
      - *origen*: 28 Millas - Siquirres, Limón, Costa Rica
      - *datos*: Acá vendrían los datos, no se incluyen en detalle para guardar la privacidad de los mismos, pero son series semanales entre los años 2011 y 2015, 830 registros en total
    - *filtro*: *suma*
    - *metodo_imputacion*: spline

- *ca*
  - *id_ca*: ca00001
  - *id_epb*: epb00001
  - *tipo_aumento_datos*: nulo
  - A: $[\overline{T}_a, \overline{Sr}, P, BW]$
  - N: $[\overline{T}_a, \overline{Sr}, P]$
  - X: Provienen de los *vca_dat*: vcadat00001, vcadat00002 y vcadat00003
  - y: Provienen del *vca_dat*: vcadat00004

- *detalles*:
  - $\overline{T}_a$:
    - ◇ *ca_id_vca*: caidvca00001Ta
    - ◇ *periodicidad*: semanal
    - ◇ *marca_temporal_inicio*: 2011
    - ◇ *marca_temporal_fin*: 2015
    - ◇ *maximo*: 50
    - ◇ *minimo*: 0
  - $\overline{Sr}$:
    - ◇ *ca_id_vca*: caidvca00002Sr
    - ◇ *periodicidad*: semanal
    - ◇ *marca_temporal_inicio*: 2011
    - ◇ *marca_temporal_fin*: 2015
    - ◇ *maximo*: 60
    - ◇ *minimo*: 0
  - P:
    - ◇ *ca_id_vca*: caidvca00003P
    - ◇ *periodicidad*: semanal
    - ◇ *marca_temporal_inicio*: 2011
    - ◇ *marca_temporal_fin*: 2015
    - ◇ *maximo*: 200
    - ◇ *minimo*: 0
  - BW:
    - ◇ *ca_id_vca*: caidvca00004BW
    - ◇ *periodicidad*: semanal
    - ◇ *marca_temporal_inicio*: 2011
    - ◇ *marca_temporal_fin*: 2015
    - ◇ *maximo*: 100
    - ◇ *minimo*: 0

2.3. Estructuración del *ca* para el pronóstico

- Actualización del *epb*
  - *epb*:
    - *id_epb*: epb00001
    - *descripcion*: Experimentación con datos de Corbana sobre la floración de banano por peso neto kilo del racimo, finca 28 Millas
    - *ca_estudio*: ca00001
    - *variables*: [vca00001Ta, vca00002Sr, vca00003P y vca00004BW]
    - C: ND

- Pat: ND
- T: ND
- T': ND
- E: ND
- O: ND
- M: ND
- MEV: ND
- S: ND
- R: ND
- AE: ND
- U: ND
- *observaciones*: ND

- Se calculan métricas estadísticas básicas del *ca*, las cuales se muestran en la tabla A.1

Tabla A.1: Estadísticas *ca* - 28 Millas - BW

Métrica	$\overline{T}_a$	$\overline{Sr}$	P	BW
Cardinalidad	799	799	799	799
Promedio	26.3	26.87	5.01	8.33
Mediana	26.37	27.65	0.5	7.2
Desviación estándar	1.15	9.38	9.82	4.98
Valor mínimo	22.23	0.17	0.0	2.8
Valor máximo	28.82	47.87	114.0	27.29
Rango	6.59	47.7	114.0	24.49
Coeficiente de variación	0.04	0.35	1.96	0.6

- D: Concatenación vertical de X e y

2.4. Generación de un nuevo *ca* con aumento de datos

- Se genera un nuevo *ca* a partir del *ca_estudio*
- Coeficiente de variación multidimensional (*cvm*): 0,16480
- q: 16
- Se utiliza el proceso propuesto en la Sección 3.2.4
- Nuevo *ca*:
  - *id_ca*: ca00002
  - *id_epb*: epb00001
  - *tipo_aumento_datos*: -CS
  - A: $[\overline{T}_a, \overline{Sr}, P, BW]$
  - N: ND
  - X: Proviene de la matriz DFcs generada (primeras $m-1$ columnas). 13567 filas

- y: Proviene de la matriz DFcs generada (última columna). 13567 elementos
- *detalles*:
  - $\overline{T}_a$:
    - ca_id_vca: caidvca00001Ta
    - *periodicidad*: semanal
    - $marca\_temporal\_inicio$: 2011
    - $marca\_temporal\_fin$: 2015
    - $maximo$: 50
    - $minimo$: 0
  - $\overline{Sr}$:
    - ca_id_vca: caidvca00002Sr
    - *periodicidad*: semanal
    - $marca\_temporal\_inicio$: 2011
    - $marca\_temporal\_fin$: 2015
    - $maximo$: 60
    - $minimo$: 0
  - P:
    - ca_id_vca: caidvca00003P
    - *periodicidad*: semanal
    - $marca\_temporal\_inicio$: 2011
    - $marca\_temporal\_fin$: 2015
    - $maximo$: 200
    - $minimo$: 0
  - BW:
    - ca_id_vca: caidvca00004BW
    - *periodicidad*: semanal
    - $marca\_temporal\_inicio$: 2011
    - $marca\_temporal\_fin$: 2015
    - $maximo$: 100
    - $minimo$: 0

2.5. Determinación de uno o varios *ca* para el entrenamiento

- Se desea validación cruzada en el mismo *ca*, por tanto el atributo C del *epb* en proceso es igual a []

2.6. Determinación del método de entrenamiento y validación

- MEV del *epb* en proceso: ValidacionCruzada

2.7. Determinación de la combinación de variables en A

- N de *ca_estudio*: $[\overline{T}_a, \overline{Sr}, P]$

2.8. Determinación de patrones

- Se desea experimentar con un grupo amplio de patrones
- Configuración para Pat del *epb* en proceso:
  - $p = 30$, $a = 20$, $inc = 3$
  - $Prev$ contiene 10 elementos [1,4,7,10,13,16,19,22,25,28]
  - $Adel$ contiene 7 elementos [1,4,7,10,13,16,19]
  - Al realizar el producto cartesiano vectorial entre $Prev$ y $Adel$, Pat contiene 70 elementos

2.9. Determinación de técnicas a utilizar

- T del *epb* en proceso: $\{SVR/L, SVR/G, SVR/S, SVR/P, ENR, OLSR\}$

2.10. Determinación del espacio paramétrico

- T' del *epb* en proceso: $\{SVR/L, SVR/G, SVR/S, SVR/P, ENR\}$
- E del *epb* en proceso:
  - SVR/L: $1 \leq C \leq 10000000$, $0 \leq epsilon \leq 5000$
  - SVR/G: $1 \leq C \leq 10000000$, $0 \leq epsilon \leq 5000$, $0{,}001 \leq \gamma \leq 0{,}999$.
  - SVR/S: $1 \leq C \leq 10000000$, $0 \leq epsilon \leq 5000$, $0{,}001 \leq \gamma \leq 0{,}999$, $0 \leq c \leq 10$.
  - SVR/P: $1 \leq C \leq 10000000$, $0 \leq epsilon \leq 5000$, $0{,}001 \leq \gamma \leq 0{,}999$, $0 \leq c \leq 10$, $d \in \{1,2,3,4,5\}$.
  - ENR: $0{,}01 \leq \alpha \leq 1$, $0 \leq \lambda \leq 1$.

2.11. Determinación de métricas

- M del *epb* en proceso: $\{R^2, RMSE\}$

3. **Etapa de preparación de los datos**

3.1. Generación de patrones

- F contiene 7 matrices de patrones, a continuación se indican los vectores columna de los atributos incluidos en cada matriz: [$[\overline{T}_a, \overline{Sr}, P, BW]$, $[\overline{T}_a, \overline{Sr}, BW]$, $[\overline{T}_a, P, BW]$, $[\overline{Sr}, P, BW]$, $[\overline{T}_a, BW]$, $[\overline{Sr}, BW]$, $[P, BW]$]
- Dado que la cantidad de elementos en Pat es 70 y la cantidad de matrices de patrones en F es 7, la cantidad de elementos en S del *epb* es: $70 \cdot 7 = 490$

3.2. Normalización de datos

- Se procede a normalizar S del *epb*

4. **Etapa de entrenamiento y validación**

4.1. Aplicación de técnicas

- Utilizando algoritmos genéticos, se calcula y completa el atributo O del *epb* en experimentación
- Se realiza la aplicación de técnicas según la Sección 3.4.1
- Dado que se aplican todas las técnicas en T, cuya cantidad de elementos es 6, en cada uno de los patrones en Pat, cuya cantidad de elementos es 490, la cantidad total de resultados en R del *epb* es 2940 ($7 \cdot 490 = 2940$)
- La tabla A.2 muestra un resumen de los resultados en R del *epb*

Tabla A.2: Resumen de los resultados al aplicar las técnicas (BW 28millas)

Técnica	ENR	OLSR	SVR/G	SVR/L	SVR/P	SVR/S
			$RMSE$			
Cardinalidad	490	490	490	490	490	490
Promedio	5.15	2.85	3.58	10.94	2.75	44369.4
Mediana	5.14	3.14	3.03	3.71	2.99	3.19
Desviación estándar	0.88	0.97	2.93	40.63	0.94	957473.4
Mínimo	1.1	0.99	0.98	0.98	0.98	0.98
Máximo	9.81	4.23	47.8	646.87	4.27	21189604.11
Rango	8.72	3.23	46.83	645.88	3.29	21189603.13
Coeficiente de variación	0.17	0.34	0.82	3.71	0.34	21.58
			R^2			
Cardinalidad	490	490	490	490	490	490
Promedio	11.88 %	66.8 %	65.17 %	59.32 %	71.84 %	63.18 %
Mediana	7.36 %	63.52 %	66.73 %	58.74 %	67.67 %	64.57 %
Desviación estándar	14.32 %	16.73 %	22.54 %	19.98 %	13.18 %	24.21 %
Mínimo	1.02 %	44.77 %	3.85 %	4.08 %	50.07 %	1.75 %
Máximo	94.66 %	95.65 %	95.74 %	95.72 %	95.59 %	95.85 %
Rango	93.64 %	50.88 %	91.89 %	91.64 %	45.52 %	94.1 %
Coeficiente de variación	120.53 %	25.05 %	34.58 %	33.68 %	18.35 %	38.31 %

4.2. Análisis estadístico

- La tabla A.3 muestra los resultados del análisis de varianza de los resultados en R del *epb*
- Se actualiza AE del *epb*

4.3. Frente de Pareto

- La figura A.1 gráfica el frente de Pareto entre $RMSE$ y R^2
- La tabla A.4 muestra las configuraciones del frente de Pareto
- La tabla A.5 muestra configuraciones del frente de Pareto para valores adicionales de a
- Se actualiza U del *epb*

5. **Etapa Conclusiva**

Tabla A.3: Resultados del diseño de experimentos con una confianza del 95 % (BW 28millas)

Factor	Métrica	H_0	Método	Cumplimiento de Supuestos		
				Normalidad	Independencia	Homocedasticidad
Técnicas	R^2	Rechazada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Rechazada	Kruskal-Wallis H-test	No	Si	NA
p	R^2	Aceptada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Aceptada	Kruskal-Wallis H-test	No	Si	NA
a	R^2	Rechazada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Rechazada	Kruskal-Wallis H-test	No	Si	NA
Variables	R^2	Aceptada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Aceptada	Kruskal-Wallis H-test	No	Si	NA

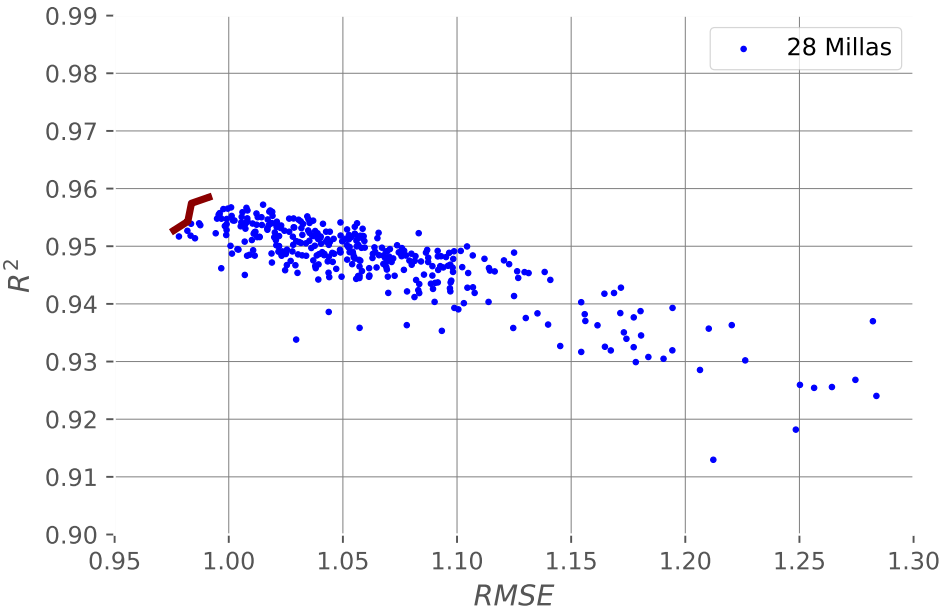


Figura A.1: Frente de Pareto (BW 28millas)

5.1. Análisis final de los resultados obtenidos

- En la sección 4.3 se mostró parte del análisis de resultados, acá se agrega solo información adicional
- La figura A.2 muestra histograma de los resultados obtenidos en cuanto a R^2

Tabla A.4: Frente de Pareto entre R^2 y $RMSE$ (BW 28millas).

Variables	$p \rightarrow a$	**Técnica**	**Parámetros**	**RMSE**	R^2
$\overline{T}_a$	$16 \rightarrow 1$	SVR/G	$C = 213{,}99$, $\epsilon = 0{,}08$, $\gamma = 0{,}35$	0,98	95,28 %
Ta-P-Sr	$1 \rightarrow 1$	SVR/G	$C = 100000{,}0$, $\epsilon = 0{,}01$, $\gamma = 0{,}06$	0,98	95,43 %
$\overline{Sr}$ P	$16 \rightarrow 1$	SVR/G	$C = 109955{,}45$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$	0,98	95,74 %
P	$13 \rightarrow 1$	SVR/S	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}03$, $c = 0$	0,99	95,85 %

Tabla A.5: Frente de Pareto para valores adicionales de a (BW 28 Millas)

Variables	$p \rightarrow a$	**Técnica**	**Parámetros**	**RMSE**	R^2
$\overline{T}_a$ $\overline{Sr}$	$22 \rightarrow 4$	SVR/P	$C = 10{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	1,72	87,82 %
	$16 \rightarrow 7$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 10$, $d = 2$	2,28	79,71 %
$\overline{T}_a$ P	$19 \rightarrow 10$	SVR/G	$C = 170{,}85$, $\epsilon = 0{,}55$, $\gamma = 0{,}98$	2,72	72,06 %
	$28 \rightarrow 13$	SVR/G	$C = 100{,}0$, $\epsilon = 0{,}1$, $\gamma = 0{,}51$	3,0	68,68 %
$\overline{Sr}$	$10 \rightarrow 16$	SVR/G	$C = 1{,}57$, $\epsilon = 0{,}03$, $\gamma = 1{,}84$	3,31	57,1 %
	$10 \rightarrow 16$	SVR/P	$C = 10{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 10$, $d = 3$	3,41	61,33 %
	$13 \rightarrow 16$	SVR/P	$C = 10{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}3$, $c = 10$, $d = 2$	3,51	61,94 %
	$28 \rightarrow 16$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	3,56	63,24 %
Ta-P-Sr	$16 \rightarrow 19$	SVR/G	$C = 93{,}19$, $\epsilon = 0{,}0$, $\gamma = 1{,}43$	3,22	64,15 %
$\overline{T}_a$ $\overline{Sr}$	$10 \rightarrow 19$	SVR/G	$C = 717{,}91$, $\epsilon = 0{,}0$, $\gamma = 1{,}13$	3,35	64,19 %

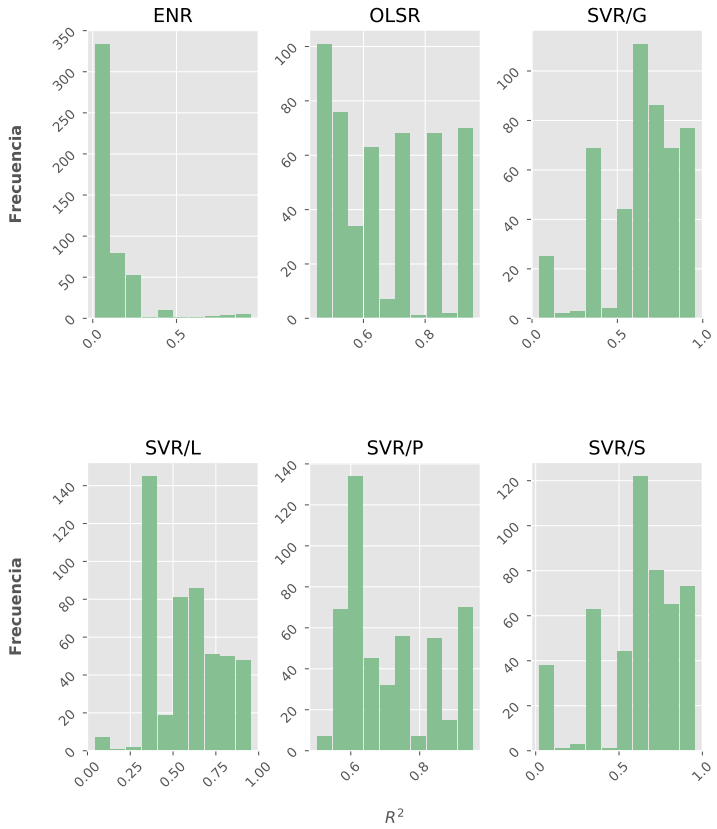


Figura A.2: Histograma de los resultados obtenidos en cuanto a R^2 (BW_28millas)

- La figura A.3 muestra histograma de los resultados obtenidos en cuanto a $RMSE$
- Se actualiza *observaciones* del *epb*

5.2. Traslado de resultados para recomendaciones agronómicas

- Se presentan los resultados a los expertos del dominio para que tomen las medidas agronómicas que correspondan
- Se podrá actualizar el atributo *observaciones* del *epb* en proceso con cual-

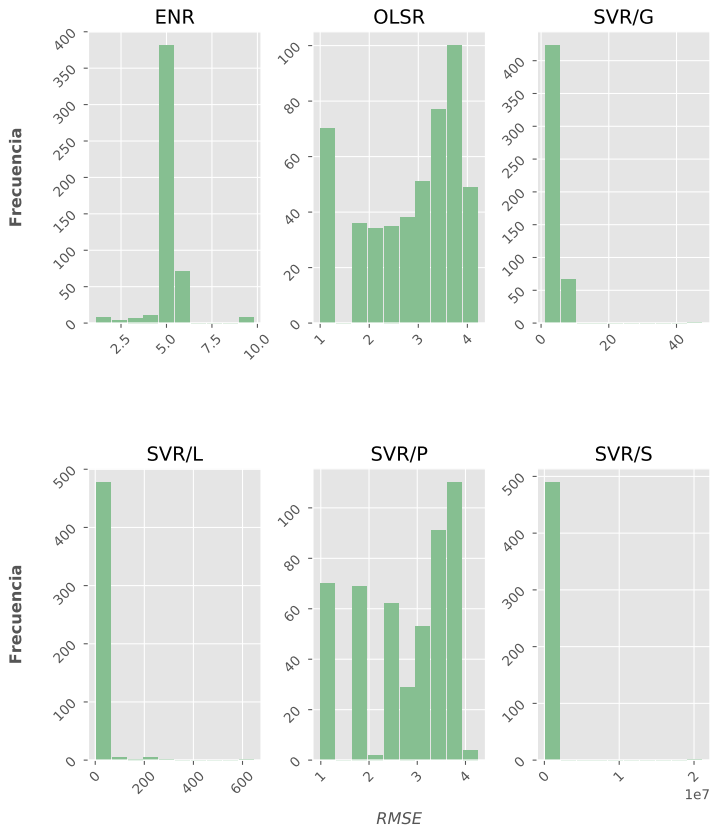


Figura A.3: Histograma de los resultados obtenidos en cuanto a $RMSE$ (BW_28millas)

quier recomendación agronómica que indiquen los expertos del dominio

5.3. Actualización del RCA

- VCA:
  - vca00001Ta:
    - vcadat00001
  - vca00002Sr:
    - vcadat00002
  - vca00003P:
    - vcadat00003
  - vca00004BW:
    - vcadat00004
- EPB:
  - epb00001
- CA:
  - ca00001
  - ca00002

Apéndice B

Detalle de resultados: Sigatoka negra

Tabla B.1: Resumen de los resultados obtenidos en cuanto R^2 y $RMSE$ para el estado de evolución de la Sigatoka negra (28 Millas)

Variables	$p \rightarrow a$	**Técnica**	**Parámetros**	**RMSE**	R^2	**Experimento**
Ta-H-W-P	$10 \rightarrow 1$	ENR	$\lambda = 1$, $\alpha = 0{,}22$	397,69	57,69 %	Validación cruzada
$\overline{T}_a$ $\overline{H}$ P	$7 \rightarrow 1$	SVR/S	$C = 753002{,}97$, $\epsilon = 2{,}79$, $\gamma = 0{,}01$, $c = 0{,}0$	397,95	58,61 %	Validación cruzada
	$9 \rightarrow 1$	SVR/S	$C = 2417872{,}37$, $\epsilon = 0{,}02$, $\gamma = 0{,}0$, $c = 0{,}18$	398,01	59,73 %	Validación cruzada
	$9 \rightarrow 1$	SVR/L	$C = 11724{,}42$, $\epsilon = 0{,}14$	399,2	59,84 %	Validación cruzada
$\overline{T}_a$ $\overline{H}$	$4 \rightarrow 1$	SVR/G	$C = 100000{,}0$, $\epsilon = 0{,}01$, $\gamma = 0{,}51$	405,46	60,18 %	Validación cruzada
$\overline{H}$	$9 \rightarrow 1$	SVR/G	$C = 10000{,}0$, $\epsilon = 0{,}01$, $\gamma = 0{,}26$	414,12	59,65 %	Tr:LR-SS / Te:28-SS
P	$4 \rightarrow 1$	SVR/S	$C = 82560{,}09$, $\epsilon = 13{,}64$, $\gamma = 0{,}07$, $c = 0{,}1$	418,34	60,54 %	Tr:LR-SS / Te:28-SS
$\overline{T}_a$ P $\overline{W}$	$6 \rightarrow 1$	ENR	$\lambda = 1$, $\alpha = 0{,}64$	419,2	60,73 %	Tr:LR-SS / Te:28-SS
$\overline{T}_a$ $\overline{H}$ $\overline{W}$	$5 \rightarrow 1$	SVR/S	$C = 1000000{,}0$, $\epsilon = 100{,}0$, $\gamma = 0{,}01$, $c = 0{,}0$	419,92	61,41 %	Tr:LR-SS / Te:28-SS
P	$4 \rightarrow 1$	SVR/L	$C = 10000000{,}0$, $\epsilon = 10{,}0$	420,39	61,01 %	Tr:LR-SS / Te:28-SS
$\overline{T}_a$ $\overline{H}$	$5 \rightarrow 1$	OLSR		423,21	60,41 %	Tr:LR-SS / Te:28-SS
$\overline{T}_a$	$6 \rightarrow 1$	SVR/L	$C = 10000{,}0$, $\epsilon = 100{,}0$	424,05	61,63 %	Tr:LR-SS / Te:28-SS
$\overline{W}$	$4 \rightarrow 1$	SVR/S	$C = 5158139{,}55$, $\epsilon = 7{,}9$, $\gamma = 0{,}01$, $c = 0{,}01$	424,98	60,91 %	Tr:LR-SS / Te:28-SS
$\overline{T}_a$ P $\overline{W}$	$6 \rightarrow 1$	ENR	$\lambda = 1$, $\alpha = 0{,}64$	425,09	62,16 %	Tr:LR-CS / Te:28-SS
$\overline{T}_a$ $\overline{H}$	$4 \rightarrow 1$	SVR/L	$C = 1000000{,}0$, $\epsilon = 0{,}1$	429,66	62,5 %	Tr:LR-SS / Te:28-SS
	$5 \rightarrow 1$	OLSR		431,43	61,48 %	Tr:LR-CS / Te:28-SS
$\overline{H}$	$9 \rightarrow 1$	SVR/G	$C = 10000{,}0$, $\epsilon = 0{,}01$, $\gamma = 0{,}26$	441,98	63,07 %	Tr:LR-CS / Te:28-SS
$\overline{T}_a$	$6 \rightarrow 1$	SVR/L	$C = 10000{,}0$, $\epsilon = 100{,}0$	443,94	63,91 %	Tr:LR-CS / Te:28-SS
$\overline{T}_a$ $\overline{H}$ $\overline{W}$	$5 \rightarrow 1$	SVR/S	$C = 1000000{,}0$, $\epsilon = 100{,}0$, $\gamma = 0{,}01$, $c = 0{,}0$	444,74	63,6 %	Tr:LR-CS / Te:28-SS
$\overline{T}_a$ $\overline{H}$	$4 \rightarrow 1$	SVR/L	$C = 1000000{,}0$, $\epsilon = 0{,}1$	447,59	63,57 %	Tr:LR-CS / Te:28-SS
P	$4 \rightarrow 1$	SVR/L	$C = 10000000{,}0$, $\epsilon = 10{,}0$	453,53	63,63 %	Tr:LR-CS / Te:28-SS
$\overline{W}$	$4 \rightarrow 1$	SVR/S	$C = 5158139{,}55$, $\epsilon = 7{,}9$, $\gamma = 0{,}01$, $c = 0{,}01$	455,44	63,82 %	Tr:LR-CS / Te:28-SS
P	$4 \rightarrow 1$	SVR/S	$C = 82560{,}09$, $\epsilon = 13{,}64$, $\gamma = 0{,}07$, $c = 0{,}1$	1479,38	54,68 %	Tr:LR-CS / Te:28-SS

Tabla B.2: Resumen de los resultados obtenidos en cuanto R^2 y $RMSE$ para el estado de evolución de la Sigatoka negra (La Rita)

Variables	$p \to a$	Técnica	Parámetros	RMSE	R^2	Experimento
$\overline{H}$	$9 \to 1$	SVR/G	$C = 10000{,}0$, $\epsilon = 0{,}01$, $\gamma = 0{,}26$	679,32	68,79 %	Validación cruzada
$\overline{T}_a\ P\ \overline{W}$	$6 \to 1$	ENR	$\lambda = 1$, $\alpha = 0{,}64$	679,5	68,8 %	Validación cruzada
$\overline{T}_a\ \overline{H}$	$5 \to 1$	OLSR		681,48	69,18 %	Validación cruzada
$\overline{T}_a\ \overline{H}\ \overline{W}$	$5 \to 1$	SVR/S	$C = 1000000{,}0$, $\epsilon = 100{,}0$, $\gamma = 0{,}01$, $c = 0{,}0$	682,03	69,31 %	Validación cruzada
$\overline{T}_a$	$6 \to 1$	SVR/L	$C = 10000{,}0$, $\epsilon = 100{,}0$	682,14	69,69 %	Validación cruzada
P	$4 \to 1$	SVR/S	$C = 82560{,}09$, $\epsilon = 13{,}64$, $\gamma = 0{,}07$, $c = 0{,}1$	682,82	69,99 %	Validación cruzada
$\overline{T}_a\ \overline{H}$	$4 \to 1$	SVR/L	$C = 1000000{,}0$, $\epsilon = 0{,}1$	685,08	70,41 %	Validación cruzada
P	$4 \to 1$	SVR/L	$C = 10000000{,}0$, $\epsilon = 10{,}0$	685,81	70,72 %	Validación cruzada
$\overline{W}$	$4 \to 1$	SVR/S	$C = 5158139{,}55$, $\epsilon = 7{,}9$, $\gamma = 0{,}01$, $c = 0{,}01$	687,61	70,85 %	Validación cruzada
$\overline{T}_a\ \overline{H}\ P$	$9 \to 1$	SVR/L	$C = 11724{,}42$, $\epsilon = 0{,}14$	710,28	65,82 %	Tr:28-CS / Te:LR-SS
	$7 \to 1$	SVR/S	$C = 753002{,}97$, $\epsilon = 2{,}79$, $\gamma = 0{,}01$, $c = 0{,}0$	711,0	66,08 %	Tr:28-CS / Te:LR-SS
	$9 \to 1$	SVR/S	$C = 2417872{,}37$, $\epsilon = 0{,}02$, $\gamma = 0{,}0$, $c = 0{,}18$	724,33	64,15 %	Tr:28-CS / Te:LR-SS
$\overline{T}_a\ \overline{H}$	$4 \to 1$	SVR/G	$C = 100000{,}0$, $\epsilon = 0{,}01$, $\gamma = 0{,}51$	725,33	61,33 %	Tr:28-SS / Te:LR-SS
$\overline{T}_a\ \overline{H}\ \overline{W}\ P$	$10 \to 1$	ENR	$\lambda = 1$, $\alpha = 0{,}22$	726,39	59,81 %	Tr:28-CS / Te:LR-SS
$\overline{T}_a\ \overline{H}$	$4 \to 1$	SVR/G	$C = 100000{,}0$, $\epsilon = 0{,}01$, $\gamma = 0{,}51$	735,07	62,5 %	Tr:28-CS / Te:LR-SS
$\overline{T}_a\ \overline{H}\ P$	$9 \to 1$	SVR/L	$C = 11724{,}42$, $\epsilon = 0{,}14$	742,63	58,41 %	Tr:28-SS / Te:LR-SS
	$9 \to 1$	SVR/S	$C = 2417872{,}37$, $\epsilon = 0{,}02$, $\gamma = 0{,}0$, $c = 0{,}18$	747,34	57,68 %	Tr:28-SS / Te:LR-SS
	$7 \to 1$	SVR/S	$C = 753002{,}97$, $\epsilon = 2{,}79$, $\gamma = 0{,}01$, $c = 0{,}0$	754,84	55,83 %	Tr:28-SS / Te:LR-SS
$\overline{T}_a\ \overline{H}\ \overline{W}\ P$	$10 \to 1$	ENR	$\lambda = 1$, $\alpha = 0{,}22$	779,24	52,56 %	Tr:28-SS / Te:LR-SS

Tabla B.3: Resumen de los resultados al aplicar las técnicas (Sigatoka negra - 28 Millas)

Técnica	ENR	OLSR	SVR/G	SVR/L	SVR/P	SVR/S
			$RMSE$			
Cardinalidad	576	576	576	576	576	576
Promedio	565.74	462.74	480.12	486.15	458.59	467.52
Mediana	514.59	474.23	479.61	484.27	468.86	476.28
Desviación estándar	247.79	38.77	52.32	58.34	37.99	43.03
Mínimo	397.69	400.24	400.22	399.2	398.97	397.95
Máximo	2283.72	534.83	649.12	651.76	535.61	607.12
Rango	1886.03	134.58	248.89	252.56	136.64	209.18
Coeficiente de variación	0.44	0.08	0.11	0.12	0.08	0.09
			R^2			
Cardinalidad	576	576	576	576	576	576
Promedio	25.24 %	45.08 %	36.72 %	36.23 %	42.76 %	39.52 %
Mediana	22.3 %	43.54 %	38.7 %	39.06 %	40.21 %	39.26 %
Desviación estándar	13.82 %	8.8 %	16.3 %	17.06 %	10.12 %	12.78 %
Mínimo	4.59 %	24.56 %	1.05 %	0.83 %	18.02 %	1.08 %
Máximo	58.15 %	59.15 %	60.18 %	59.93 %	58.5 %	59.73 %
Rango	53.56 %	34.59 %	59.13 %	59.1 %	40.48 %	58.65 %
Coeficiente de variación	54.77 %	19.52 %	44.4 %	47.09 %	23.67 %	32.34 %

Tabla B.4: Resumen de los resultados al aplicar las técnicas (Sigatoka negra - La Rita)

Técnica	ENR	OLSR	SVR/G	SVR/L	SVR/P	SVR/S
			$RMSE$			
Cardinalidad	576	576	576	576	576	576
Promedio	935.4	820.04	908.73	867.55	812.85	882.97
Mediana	921.74	833.11	901.87	843.12	827.56	858.48
Desviación estándar	178.15	94.15	166.59	142.13	94.3	151.93
Mínimo	679.5	681.48	679.32	681.95	681.79	681.39
Máximo	2480.34	997.61	1248.83	1239.61	939.89	1242.27
Rango	1800.84	316.13	569.51	557.66	258.1	560.89
Coeficiente de variación	0.19	0.11	0.18	0.16	0.12	0.17
			R^2			
Cardinalidad	576	576	576	576	576	576
Promedio	35.1 %	56.19 %	42.11 %	47.11 %	53.82 %	44.19 %
Mediana	35.35 %	55.19 %	47.88 %	50.52 %	53.37 %	47.58 %
Desviación estándar	18.27 %	9.15 %	23.15 %	19.69 %	10.22 %	20.46 %
Mínimo	5.9 %	42.46 %	0.33 %	0.36 %	37.37 %	0.47 %
Máximo	68.8 %	69.18 %	70.67 %	70.72 %	68.92 %	70.85 %
Rango	62.9 %	26.72 %	70.34 %	70.37 %	31.55 %	70.38 %
Coeficiente de variación	52.04 %	16.28 %	54.98 %	41.79 %	18.99 %	46.29 %

Tabla B.5: Resultados obtenidos en el diseño experimental para el proceso biológico Sigatoka negra

Factor	Métrica	H_0	Método	**Cumplimiento de Supuestos**		
				Normalidad	Independencia	Homocedasticidad
Técnicas	R^2	Rechazada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Rechazada	Wilcoxon signed-rank test	No	No	NA
p	R^2	Aceptada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Aceptada	Kruskal-Wallis H-test	No	Si	NA
a	R^2	Rechazada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Rechazada	Kruskal-Wallis H-test	No	Si	NA
Variables	R^2	Rechazada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Aceptada	Kruskal-Wallis H-test	No	Si	NA

Apéndice C

Detalle de resultados: roya

Tabla C.1: Resultados con las configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para la incidencia de la roya (San Carlos)

Variables	$p \to a$	Técnica	Parámetros	RMSE	R^2	Experimento
$\overline{At}\ \overline{T}_a\ \overline{W}\ \overline{Hdd}$	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	1,72	77,52 %	Validación cruzada
$\overline{T}_a\ \overline{W}\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	2,16	71,68 %	Tr:F-CS / Te:SC-SS
$\overline{At}\ \overline{T}_a\ P\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	2,69	64,75 %	Tr:D-SS / Te:SC-SS
$\overline{W}\ P\ \overline{Hdd}$	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	3,17	65,35 %	Tr:D-SS / Te:SC-SS
$P\ \overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	3,58	28,6 %	Tr:D-CS / Te:SC-SS
$\overline{At}\ \overline{T}_a\ P\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	3,84	55,55 %	Tr:D-CS / Te:SC-SS
$\overline{At}$	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	4,33	37,41 %	Tr:F-SS / Te:SC-SS
$\overline{T}_a\ \overline{W}\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	4,69	64,48 %	Tr:F-SS / Te:SC-SS
$\overline{W}\ P\ \overline{Hdd}$	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	5,12	45,44 %	Tr:D-CS / Te:SC-SS
$P\ \overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	6,35	48,57 %	Tr:D-SS / Te:SC-SS
All	$1 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}9$, $\gamma = 0{,}5$, $c = 0$, $d = 2$	7,39	43,08 %	Tr:P-SS / Te:SC-SS
	$1 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}9$, $\gamma = 0{,}5$, $c = 0$, $d = 2$	7,43	39,07 %	Tr:P-CS / Te:SC-SS
$\overline{At}$	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	10,03	46,53 %	Tr:F-CS / Te:SC-SS
$\overline{T}_a\ \overline{H}\ \overline{W}\ P$	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	10,11	56,41 %	Tr:B-CS / Te:SC-SS
$\overline{H}\ \overline{W}\ \overline{Hdd}$	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	12,16	48,28 %	Tr:B-CS / Te:SC-SS
$\overline{At}\ \overline{T}_a$	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	12,39	47,89 %	Tr:C-SS / Te:SC-SS
$\overline{H}\ \overline{W}\ \overline{Hdd}$	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	12,8	48,34 %	Tr:B-SS / Te:SC-SS
$\overline{T}_a\ \overline{H}\ \overline{W}\ P$	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	13,13	55,52 %	Tr:B-SS / Te:SC-SS
	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	17,89	49,36 %	Tr:B-CS / Te:SC-SS
$\overline{At}\ \overline{T}_a$	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	21,57	49,4 %	Tr:C-CS / Te:SC-SS
$\overline{T}_a\ \overline{H}\ \overline{W}\ P$	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	24,76	50,23 %	Tr:B-SS / Te:SC-SS
$\overline{At}\ \overline{T}_a\ \overline{H}\ P$	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	28,2	50,51 %	Tr:SV-SS / Te:SC-SS
	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	48,01	49,22 %	Tr:SV-CS / Te:SC-SS

Tabla C.2: Resultados con las configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para la incidencia de la roya (SanVito)

Variables	$p \to a$	Técnica	Parámetros	RMSE	R^2	Experimento
$\overline{At}\ \overline{T_a}\ \overline{H}\ P$	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	11,7	84,62 %	Validación cruzada
$\overline{W}\ P\ \overline{Hdd}$	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	23,02	70,4 %	Tr:D-SS / Te:SV-SS
$\overline{H}\ \overline{W}\ \overline{Hdd}$	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	23,49	66,66 %	Tr:B-SS / Te:SV-SS
$\overline{W}\ P\ \overline{Hdd}$	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	30,51	71,61 %	Tr:D-CS / Te:SV-SS
$\overline{At}\ \overline{T_a}\ P\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	32,48	56,63 %	Tr:D-SS / Te:SV-SS
$\overline{H}\ \overline{W}\ \overline{Hdd}$	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	34,04	50,45 %	Tr:B-CS / Te:SV-SS
$\overline{At}\ \overline{T_a}\ \overline{Hdd}\ \overline{Lw1}$	$2 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 0$, $d = 3$	39,64	64,76 %	Tr:D-SS / Te:SV-SS
$P\ \overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	43,25	50,36 %	Tr:D-CS / Te:SV-SS
$\overline{T_a}\ \overline{H}\ \overline{W}\ P$	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	51,73	48,41 %	Tr:B-CS / Te:SV-SS
	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	67,22	58,33 %	Tr:B-SS / Te:SV-SS
$\overline{At}\ \overline{T_a}\ \overline{W}\ \overline{Hdd}$	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	67,68	44,54 %	Tr:SC-CS / Te:SV-SS
$P\ \overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	68,18	53,81 %	Tr:D-SS / Te:SV-SS
$\overline{At}\ \overline{T_a}\ \overline{W}\ \overline{Hdd}$	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	87,24	44,5 %	Tr:SC-SS / Te:SV-SS
$\overline{T_a}\ \overline{H}\ \overline{W}\ P$	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	90,87	57,1 %	Tr:B-SS / Te:SV-SS
$\overline{T_a}\ \overline{W}\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	97,21	55,67 %	Tr:F-SS / Te:SV-SS
	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	114,48	56,0 %	Tr:F-CS / Te:SV-SS
$\overline{At}\ \overline{T_a}\ \overline{Hdd}\ \overline{Lw1}$	$2 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 0$, $d = 3$	124,19	53,97 %	Tr:D-CS / Te:SV-SS
$\overline{At}\ \overline{H}\ \overline{Hdd}\ \overline{Lw1}$	$5 \to 1$	SVR/P	$C = 100000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}3$, $c = 10$, $d = 2$	166,31	50,29 %	Tr:C-SS / Te:SV-SS
$\overline{At}\ \overline{T_a}\ P\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	174,7	52,32 %	Tr:D-CS / Te:SV-SS
Ta-H-W-P	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	456,03	50,87 %	Tr:B-CS / Te:SV-SS
$\overline{At}$	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	539,16	49,72 %	Tr:F-SS / Te:SV-SS
$\overline{At}\ \overline{T_a}$	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	598,64	49,9 %	Tr:C-SS / Te:SV-SS
$\overline{At}\ \overline{H}\ \overline{Hdd}\ \overline{Lw1}$	$5 \to 1$	SVR/P	$C = 100000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}3$, $c = 10$, $d = 2$	748,76	49,73 %	Tr:C-CS / Te:SV-SS
$\overline{At}$	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	976,88	48,96 %	Tr:F-CS / Te:SV-SS
$\overline{At}\ \overline{T_a}$	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	1147,96	50,19 %	Tr:C-CS / Te:SV-SS

Tabla C.3: Resultados con las configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para la incidencia de la roya (Barva)

Variables	$p \to a$	Técnica	Parámetros	RMSE	R^2	Experimento
$\overline{H}\ \overline{W}\ \overline{Hdd}$	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	14,73	76,62 %	Validación cruzada
$\overline{T_a}\ \overline{H}\ \overline{W}\ P$	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	14,86	77,69 %	Validación cruzada
	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	15,18	80,74 %	Validación cruzada
$\overline{W}\ P\ \overline{Hdd}$	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	23,08	69,25 %	Tr:D-SS / Te:B-SS
	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	23,86	76,58 %	Tr:D-CS / Te:B-SS
All	$1 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}9$, $\gamma = 0{,}5$, $c = 0$, $d = 2$	27,82	53,73 %	Tr:P-SS / Te:B-SS
$\overline{At}\ \overline{T_a}\ P\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	33,65	60,79 %	Tr:D-SS / Te:B-SS
$\overline{At}\ \overline{T_a}\ \overline{H}\ P$	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	38,53	52,87 %	Tr:SV-CS / Te:B-SS
$P\ \overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	39,3	48,15 %	Tr:D-CS / Te:B-SS
$\overline{At}\ \overline{T_a}\ \overline{H}\ P$	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	39,87	58,55 %	Tr:SV-SS / Te:B-SS
$\overline{At}\ \overline{T_a}\ \overline{W}\ \overline{Hdd}$	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	58,29	42,29 %	Tr:SC-CS / Te:B-SS
$P\ \overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	62,38	52,75 %	Tr:D-SS / Te:B-SS
$\overline{At}\ \overline{T_a}\ \overline{W}\ \overline{Hdd}$	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	74,17	42,13 %	Tr:SC-SS / Te:B-SS
$\overline{T_a}\ \overline{W}\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	79,86	56,84 %	Tr:F-SS / Te:B-SS
All	$1 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}9$, $\gamma = 0{,}5$, $c = 0$, $d = 2$	82,85	39,88 %	Tr:P-CS / Te:B-SS
$\overline{T_a}\ \overline{W}\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	93,26	56,3 %	Tr:F-CS / Te:B-SS
$\overline{At}\ \overline{T_a}\ P\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	147,04	52,76 %	Tr:D-CS / Te:B-SS
$\overline{At}$	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	465,08	49,43 %	Tr:F-SS / Te:B-SS
$\overline{At}\ \overline{T_a}$	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	574,05	49,86 %	Tr:C-SS / Te:B-SS
$\overline{At}$	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	868,55	49,14 %	Tr:F-CS / Te:B-SS
$\overline{At}\ \overline{T_a}$	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	1072,0	50,04 %	Tr:C-CS / Te:B-SS

Tabla C.4: Resultados con las configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para la incidencia de la roya (Carrizal)

Variables	$p \to a$	Técnica	Parámetros	RMSE	R^2	Experimento
$\overline{At}\ \overline{T}_a$	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	2,47	76,79 %	Validación cruzada
$\overline{At}\ \overline{H}\ \overline{Hdd}\ \overline{Lw1}$	$5 \to 1$	SVR/P	$C = 100000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}3$, $c = 10$, $d = 2$	2,52	80,64 %	Validación cruzada
$\overline{At}\ \overline{T}_a\ \overline{Hdd}\ \overline{Lw1}$	$2 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 0$, $d = 3$	3,4	68,26 %	Tr:D-SS / Te:C-SS
$\overline{At}\ \overline{T}_a\ P\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	3,55	63,64 %	Tr:D-SS / Te:C-SS
$\overline{At}\ \overline{T}_a\ \overline{Hdd}\ \overline{Lw1}$	$2 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 0$, $d = 3$	4,18	54,19 %	Tr:D-CS / Te:C-SS
$\overline{W}\ P\ \overline{Hdd}$	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	4,39	67,37 %	Tr:D-SS / Te:C-SS
$\overline{At}\ \overline{T}_a\ P\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	4,59	47,89 %	Tr:D-CS / Te:C-SS
$\overline{W}\ P\ \overline{Hdd}$	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	4,75	59,01 %	Tr:D-CS / Te:C-SS
$\overline{At}$	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	5,82	47,35 %	Tr:F-SS / Te:C-SS
$\overline{T}_a\ \overline{W}\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	6,82	60,96 %	Tr:F-CS / Te:C-SS
$P\ \overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	7,16	33,57 %	Tr:D-CS / Te:C-SS
$\overline{At}\ \overline{T}_a\ \overline{W}\ \overline{Hdd}$	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	10,43	40,12 %	Tr:SC-SS / Te:C-SS
	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	11,29	46,81 %	Tr:SC-CS / Te:C-SS
$\overline{H}\ \overline{W}\ \overline{Hdd}$	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	11,45	46,93 %	Tr:B-CS / Te:C-SS
	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	11,71	46,93 %	Tr:B-SS / Te:C-SS
$P\ \overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	12,47	46,0 %	Tr:D-SS / Te:C-SS
$\overline{T}_a\ \overline{H}\ \overline{W}\ P$	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	14,58	55,8 %	Tr:B-CS / Te:C-SS
Ta-H-W-P	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	15,4	58,06 %	Tr:B-SS / Te:C-SS
$\overline{At}$	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	18,02	48,47 %	Tr:F-CS / Te:C-SS
$\overline{T}_a\ \overline{W}\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	21,58	50,88 %	Tr:F-SS / Te:C-SS
$\overline{T}_a\ \overline{H}\ \overline{W}\ P$	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	23,49	51,02 %	Tr:B-CS / Te:C-SS
$\overline{At}\ \overline{T}_a\ \overline{H}\ P$	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	27,04	51,75 %	Tr:SV-SS / Te:C-SS
$\overline{T}_a\ \overline{H}\ \overline{W}\ P$	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	34,08	50,71 %	Tr:B-SS / Te:C-SS
$\overline{At}\ \overline{T}_a\ \overline{H}\ P$	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	48,3	49,32 %	Tr:SV-CS / Te:C-SS

Tabla C.5: Resultados con las configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para la incidencia de la roya (Dota)

Variables	$p \to a$	Técnica	Parámetros	RMSE	R^2	Experimento
$\overline{W}\ P\ \overline{Hdd}$	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	1,64	75,42 %	Validación cruzada
$\overline{At}\ \overline{T}_a\ \overline{Hdd}\ \overline{Lw1}$	$2 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 0$, $d = 3$	1,66	78,2 %	Validación cruzada
$P\ \overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	1,67	78,3 %	Validación cruzada
$\overline{At}\ \overline{T}_a\ P\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	1,81	78,9 %	Validación cruzada
$\overline{At}$	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	3,58	40,43 %	Tr:F-SS / Te:D-SS
	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	3,71	41,35 %	Tr:F-CS / Te:D-SS
$\overline{At}\ \overline{T}_a\ \overline{W}\ \overline{Hdd}$	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	4,94	55,41 %	Tr:SC-CS / Te:D-SS
$\overline{At}\ \overline{H}\ \overline{Hdd}\ \overline{Lw1}$	$5 \to 1$	SVR/P	$C = 100000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}3$, $c = 10$, $d = 2$	4,95	56,49 %	Tr:C-SS / Te:D-SS
$\overline{T}_a\ \overline{W}\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	5,01	56,24 %	Tr:F-CS / Te:D-SS
$\overline{At}\ \overline{T}_a\ \overline{W}\ \overline{Hdd}$	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	7,21	53,28 %	Tr:SC-SS / Te:D-SS
$\overline{T}_a\ \overline{W}\ \overline{Hdd}$	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	12,37	54,01 %	Tr:F-SS / Te:D-SS
Ta-H-W-P	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	12,85	53,69 %	Tr:B-SS / Te:D-SS
$\overline{H}\ \overline{W}\ \overline{Hdd}$	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	14,57	48,44 %	Tr:B-SS / Te:D-SS
$\overline{T}_a\ \overline{H}\ \overline{W}\ P$	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	16,28	50,6 %	Tr:B-SS / Te:D-SS
$\overline{H}\ \overline{W}\ \overline{Hdd}$	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	16,97	47,6 %	Tr:B-CS / Te:D-SS
$\overline{T}_a\ \overline{H}\ \overline{W}\ P$	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	17,14	51,9 %	Tr:B-CS / Te:D-SS
Ta-H-W-P	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	18,51	48,87 %	Tr:B-CS / Te:D-SS
$\overline{At}\ \overline{T}_a$	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	20,86	48,23 %	Tr:C-SS / Te:D-SS
$\overline{At}\ \overline{T}_a\ \overline{H}\ P$	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	27,1	50,34 %	Tr:SV-SS / Te:D-SS
$\overline{At}\ \overline{T}_a$	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	32,3	49,98 %	Tr:C-CS / Te:D-SS
$\overline{At}\ \overline{H}\ \overline{Hdd}\ \overline{Lw1}$	$5 \to 1$	SVR/P	$C = 100000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}3$, $c = 10$, $d = 2$	37,11	49,52 %	Tr:C-CS / Te:D-SS
$\overline{At}\ \overline{T}_a\ \overline{H}\ P$	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	51,2	49,49 %	Tr:SV-CS / Te:D-SS

Tabla C.6: Resultados con las configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para la incidencia de la roya (Frailes)

Variables	$p \to a$	Técnica	Parámetros	RMSE	R^2	Experimento
$\overline{At}$	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	1,27	73,07 %	Validación cruzada
$\overline{T_a}$ $\overline{W}$ $\overline{Hdd}$	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	1,56	80,19 %	Validación cruzada
$\overline{W}$ P $\overline{Hdd}$	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	2,66	68,5 %	Tr:D-SS / Te:F-SS
$\overline{At}$ $\overline{T_a}$ P $\overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	3,25	64,02 %	Tr:D-SS / Te:F-SS
P $\overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	3,47	35,3 %	Tr:D-CS / Te:F-SS
$\overline{At}$ $\overline{T_a}$ P $\overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	3,78	53,96 %	Tr:D-CS / Te:F-SS
$\overline{At}$ $\overline{T_a}$ $\overline{W}$ $\overline{Hdd}$	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	4,22	48,78 %	Tr:SC-CS / Te:F-SS
	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	4,46	41,2 %	Tr:SC-SS / Te:F-SS
$\overline{W}$ P $\overline{Hdd}$	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	4,85	57,09 %	Tr:D-CS / Te:F-SS
All	$1 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}9$, $\gamma = 0{,}5$, $c = 0$, $d = 2$	5,13	39,25 %	Tr:P-CS / Te:F-SS
P $\overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	6,12	51,58 %	Tr:D-SS / Te:F-SS
All	$1 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}9$, $\gamma = 0{,}5$, $c = 0$, $d = 2$	6,68	46,51 %	Tr:P-SS / Te:F-SS
$\overline{H}$ $\overline{W}$ $\overline{Hdd}$	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	11,37	49,06 %	Tr:B-CS / Te:F-SS
	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	11,93	49,22 %	Tr:B-SS / Te:F-SS
$\overline{T_a}$ $\overline{H}$ $\overline{W}$ P	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	13,13	53,01 %	Tr:B-CS / Te:F-SS
Ta-H-W-P	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	19,84	52,39 %	Tr:B-SS / Te:F-SS
$\overline{At}$ $\overline{T_a}$ $\overline{H}$ P	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	25,78	50,55 %	Tr:SV-SS / Te:F-SS
$\overline{At}$ $\overline{T_a}$	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	27,73	49,16 %	Tr:C-SS / Te:F-SS
Ta-H-W-P	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	28,17	49,75 %	Tr:B-CS / Te:F-SS
$\overline{T_a}$ $\overline{H}$ $\overline{W}$ P	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	29,1	50,02 %	Tr:B-SS / Te:F-SS
$\overline{At}$ $\overline{T_a}$	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	53,1	49,94 %	Tr:C-CS / Te:F-SS
$\overline{At}$ $\overline{T_a}$ $\overline{H}$ P	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	55,5	49,8 %	Tr:SV-CS / Te:F-SS

Tabla C.7: Resultados con las configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para la incidencia de la roya (Poas)

Variables	$p \to a$	Técnica	Parámetros	RMSE	R^2	Experimento
All	$1 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}9$, $\gamma = 0{,}5$, $c = 0$, $d = 2$	2,85	60,59 %	Validación cruzada
$\overline{W}$ P $\overline{Hdd}$	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	4,82	44,91 %	Tr:D-SS / Te:P-SS
$\overline{At}$ $\overline{T_a}$ $\overline{W}$ $\overline{Hdd}$	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	8,73	39,79 %	Tr:SC-SS / Te:P-SS
P $\overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	10,08	45,9 %	Tr:D-CS / Te:P-SS
$\overline{W}$ P $\overline{Hdd}$	$4 \to 1$	SVR/S	$C = 114774{,}13$, $\epsilon = 0{,}1$, $\gamma = 0{,}02$, $c = 0$	11,05	54,71 %	Tr:D-CS / Te:P-SS
$\overline{T_a}$ $\overline{W}$ $\overline{Hdd}$	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	11,9	48,41 %	Tr:F-CS / Te:P-SS
$\overline{H}$ $\overline{W}$ $\overline{Hdd}$	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	15,17	51,64 %	Tr:B-CS / Te:P-SS
	$10 \to 1$	SVR/P	$C = 1{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}5$, $c = 1$, $d = 3$	16,49	50,6 %	Tr:B-SS / Te:P-SS
P $\overline{Hdd}$	$12 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$, $c = 1$, $d = 3$	16,68	47,61 %	Tr:D-SS / Te:P-SS
Ta-H-W-P	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	17,26	50,76 %	Tr:B-SS / Te:P-SS
$\overline{At}$	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	18,22	50,74 %	Tr:F-SS / Te:P-SS
$\overline{T_a}$ $\overline{H}$ $\overline{W}$ P	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	18,84	51,21 %	Tr:B-CS / Te:P-SS
$\overline{At}$ $\overline{T_a}$ $\overline{W}$ $\overline{Hdd}$	$11 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 3$	19,07	49,74 %	Tr:SC-CS / Te:P-SS
$\overline{T_a}$ $\overline{W}$ $\overline{Hdd}$	$3 \to 1$	SVR/P	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 10$, $d = 2$	21,63	49,48 %	Tr:F-SS / Te:P-SS
$\overline{At}$ $\overline{T_a}$ $\overline{H}$ P	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	21,86	50,26 %	Tr:SV-SS / Te:P-SS
	$2 \to 1$	SVR/P	$C = 100{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}97$, $c = 5$, $d = 3$	25,13	50,77 %	Tr:SV-CS / Te:P-SS
$\overline{At}$ $\overline{T_a}$ P $\overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	30,15	49,83 %	Tr:D-SS / Te:P-SS
$\overline{T_a}$ $\overline{H}$ $\overline{W}$ P	$1 \to 1$	SVR/G	$C = 1000000{,}0$, $\epsilon = 10{,}0$, $\gamma = 0{,}26$	30,34	50,13 %	Tr:B-SS / Te:P-SS
	$1 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 1$, $d = 3$	36,39	49,94 %	Tr:B-CS / Te:P-SS
$\overline{At}$ $\overline{T_a}$ P $\overline{Hdd}$	$3 \to 1$	SVR/P	$C = 1000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}7$, $c = 5$, $d = 2$	87,8	49,96 %	Tr:D-CS / Te:P-SS
$\overline{At}$	$11 \to 1$	SVR/P	$C = 10000000{,}0$, $\epsilon = 0{,}4$, $\gamma = 0{,}1$, $c = 0$, $d = 3$	303,87	50,02 %	Tr:F-CS / Te:P-SS
$\overline{At}$ $\overline{T_a}$	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	542,98	50,02 %	Tr:C-SS / Te:P-SS
	$10 \to 2$	SVR/P	$C = 1000000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}1$, $c = 5$, $d = 2$	871,61	49,99 %	Tr:C-CS / Te:P-SS

Tabla C.8: Resultados obtenidos en el diseño experimental por pares de niveles para el proceso biológico roya

Factor	Métrica	H_0	Método	Cumplimiento de Supuestos		
				Normalidad	Independencia	Homocedasticidad
Técnicas	R^2	Aceptada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Rechazada	Wilcoxon signed-rank test	No	No	NA
p	R^2	Aceptada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Aceptada	Kruskal-Wallis H-test	No	Si	NA
a	R^2	Aceptada	ANOVA	Si	Si	Si
	$RMSE$	Rechazada	Wilcoxon signed-rank test	No	No	NA
Lugar	R^2	Aceptada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Aceptada	Kruskal-Wallis H-test	No	Si	NA
Variables	R^2	Aceptada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Aceptada	Kruskal-Wallis H-test	No	Si	NA

Tabla C.9: Resultados al aplicar tSNE en los conjuntos de datos de la roya, ordenados por *pnca* (7 variables) - Parte 1

Orden	**Conjunto de datos** (ca)	**Norma euclidiana** (nca)	**Distancia** ($dnca$)	**Distancia relativa** ($pnca$)
1	Dota	158,5761	0,0007	0,0002 %
	Carrizal	158,5768		
2	Carrizal	158,5768	2,4873	0,7782 %
	Frailes	161,0641		
3	Dota	158,5761	2,488	0,7784 %
	Frailes	161,0641		
4	Poas	461,3744	59,3891	6,8788 %
	Barva	401,9853		
5	SanCarlos	189,1821	28,118	8,0281 %
	Frailes	161,0641		
6	SanCarlos	189,1821	30,6053	8,8007 %
	Carrizal	158,5768		
7	Dota	158,5761	30,606	8,8009 %
	SanCarlos	189,1821		
8	Poas	461,3744	124,7041	11,9055 %
	SanVito	586,0785		
9	SanVito	586,0785	184,0932	18,6317 %
	Barva	401,9853		
10	Barva	401,9853	212,8033	35,9971 %
	SanCarlos	189,1821		

Tabla C.10: Resultados al aplicar tSNE en los conjuntos de datos de la roya, ordenados por *pnca* (7 variables) - Parte 2

Orden	**Conjunto de datos** (*ca*)	**Norma euclidiana** (*nca*)	**Distancia** (*dnca*)	**Distancia relativa** (*pnca*)
11	Poas SanCarlos	461,3744 189,1821	272,1924	41,8399 %
12	Barva Frailes	401,9853 161,0641	240,9212	42,7886 %
13	Barva Carrizal	401,9853 158,5768	243,4085	43,4222 %
14	Barva Dota	401,9853 158,5761	243,4093	43,4224 %
15	Poas Frailes	461,3744 161,0641	300,3103	48,2474 %
16	Poas Carrizal	461,3744 158,5768	302,7976	48,8422 %
17	Poas Dota	461,3744 158,5761	302,7983	48,8423 %
18	SanVito SanCarlos	586,0785 189,1821	396,8965	51,1952 %
19	SanVito Frailes	586,0785 161,0641	425,0145	56,8853 %
20	SanVito Carrizal	586,0785 158,5768	427,5017	57,4093 %
21	SanVito Dota	586,0785 158,5761	427,5025	57,4095 %

Tabla C.11: Resultados al aplicar tSNE en los conjuntos de datos de la roya, ordenados por Divergencia KL - Parte 1

Orden	**Divergencia KL**	**ca**
1	0,0499	SanVito Frailes
2	0,0716	SanVito Carrizal
3	0,0745	SanVito SanCarlos
4	0,0746	SanVito Dota
5	0,079	Barva Frailes
6	0,092	Barva Carrizal
7	0,1142	Barva Dota
8	0.1229	Barva SanCarlos
9	0.1299	SanVito Barva
10	0.1348	Poas SanVito

Tabla C.12: Resultados al aplicar tSNE en los conjuntos de datos de la roya, ordenados por Divergencia KL - Parte 2

Orden	**Divergencia KL**	**ca**
11	0.1482	Poas SanCarlos
12	0.1617	Poas Dota
13	0.1654	Poas Frailes
14	0.1792	SanCarlos Carrizal
15	0.1892	Poas Carrizal
16	0.1914	Poas Barva
17	0.2097	Dota Frailes
18	0.2261	Dota SanCarlos
19	0.2325	Carrizal Frailes
20	0.2496	SanCarlos Frailes
21	0.2563	Dota Carrizal

Apéndice D

Detalle de resultados: floración del banano

Tabla D.1: Resumen de los resultados obtenidos en cuanto R^2 y $RMSE$ para la floración (Las Valquirias)

Variables	$p \to a$	**Técnica**	**Parámetros**	**RMSE**	R^2	**Experimento**
$\overline{Sr}$ P	$4 \to 1$	SVR/P	$C = 10{,}0$, $\epsilon = 0{,}1$, $\gamma = 0{,}9$, $c = 0$, $d = 3$	0,86	96,6 %	Validación cruzada
P	$13 \to 1$	SVR/S	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}03$, $c = 0$	0,93	96,22 %	Tr:28-SS / Te:LV-SS
Ta-P-Sr	$1 \to 1$	SVR/G	$C = 100000{,}0$, $\epsilon = 0{,}01$, $\gamma = 0{,}06$	0,93	96,11 %	Tr:28-SS / Te:LV-SS
$\overline{Sr}$ P	$16 \to 1$	SVR/G	$C = 109955{,}45$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$	0,99	95,64 %	Tr:28-CS / Te:LV-SS
	$16 \to 1$	SVR/G	$C = 109955{,}45$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$	1,0	95,51 %	Tr:28-SS / Te:LV-SS
$\overline{T}_a$	$16 \to 1$	SVR/G	$C = 213{,}99$, $\epsilon = 0{,}08$, $\gamma = 0{,}35$	1,07	94,33 %	Tr:28-CS / Te:LV-SS
	$16 \to 1$	SVR/G	$C = 213{,}99$, $\epsilon = 0{,}08$, $\gamma = 0{,}35$	1,08	94,15 %	Tr:28-SS / Te:LV-SS
Ta-P-Sr	$1 \to 1$	SVR/G	$C = 100000{,}0$, $\epsilon = 0{,}01$, $\gamma = 0{,}06$	1,75	88,4 %	Tr:28-CS / Te:LV-SS
P	$13 \to 1$	SVR/S	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}03$, $c = 0$	16,2	45,92 %	Tr:28-CS / Te:LV-SS

Tabla D.2: Resumen de los resultados obtenidos en cuanto R^2 y $RMSE$ para la floración (28 Millas)

Variables	$p \to a$	**Técnica**	**Parámetros**	**RMSE**	R^2	**Experimento**
$\overline{T}_a$	$16 \to 1$	SVR/G	$C = 213{,}99$, $\epsilon = 0{,}08$, $\gamma = 0{,}35$	0,98	95,28 %	Validación cruzada
Ta-P-Sr	$1 \to 1$	SVR/G	$C = 100000{,}0$, $\epsilon = 0{,}01$, $\gamma = 0{,}06$	0,98	95,43 %	Validación cruzada
$\overline{Sr}$ P	$16 \to 1$	SVR/G	$C = 109955{,}45$, $\epsilon = 0{,}0$, $\gamma = 0{,}0$	0,98	95,74 %	Validación cruzada
P	$13 \to 1$	SVR/S	$C = 10000{,}0$, $\epsilon = 0{,}0$, $\gamma = 0{,}03$, $c = 0$	0,99	95,85 %	Validación cruzada
$\overline{Sr}$ P	$4 \to 1$	SVR/P	$C = 10{,}0$, $\epsilon = 0{,}1$, $\gamma = 0{,}9$, $c = 0$, $d = 3$	1,11	95,36 %	Tr:LV-CS / Te:28-SS
	$4 \to 1$	SVR/P	$C = 10{,}0$, $\epsilon = 0{,}1$, $\gamma = 0{,}9$, $c = 0$, $d = 3$	1,14	94,85 %	Tr:LV-SS / Te:28-SS

Tabla D.3: Configuraciones del frente de Pareto en cuanto R^2 y $RMSE$ para diferentes periodos adelante (a), en la etapa de validación cruzada para la floración del banano (28 Millas)

Variables	$p \rightarrow a$	**Técnica**	$RMSE$		R^2	
			mean	*stdev*	*mean*	*stdev*
$\overline{\overline{T}_a}$	$16 \rightarrow 1$	SVR/G	0,98	0,44	95,28 %	4,52 %
Ta-P-Sr	$1 \rightarrow 1$	SVR/G	0,98	0,4	95,43 %	3,4 %
$\overline{Sr}$ P	$16 \rightarrow 1$	SVR/G	0,98	0,4	95,74 %	2,81 %
P	$13 \rightarrow 1$	SVR/S	0,99	0,44	95,85 %	2,55 %
$\overline{\overline{T}_a}$ $\overline{Sr}$	$22 \rightarrow 4$	SVR/P	1,72	0,77	87,82 %	8,63 %
	$16 \rightarrow 7$	SVR/P	2,28	0,91	79,71 %	10,76 %
$\overline{\overline{T}_a}$ P	$19 \rightarrow 10$	SVR/G	2,72	0,45	72,06 %	9,83 %
	$28 \rightarrow 13$	SVR/G	3,0	0,78	68,68 %	8,4 %
$\overline{Sr}$	$10 \rightarrow 16$	SVR/G	3,31	0,76	57,1 %	9,26 %
	$10 \rightarrow 16$	SVR/P	3,41	0,93	61,33 %	10,93 %
	$13 \rightarrow 16$	SVR/P	3,51	0,87	61,94 %	9,65 %
	$28 \rightarrow 16$	SVR/P	3,56	0,47	63,24 %	6,28 %
Ta-P-Sr	$16 \rightarrow 19$	SVR/G	3,22	0,7	64,15 %	10,32 %
$\overline{\overline{T}_a}$ $\overline{Sr}$	$10 \rightarrow 19$	SVR/G	3,35	0,63	64,19 %	10,56 %

Tabla D.4: Configuraciones del frente de Pareto en cuanto R^2 y $RMSE$, para diferentes periodos adelante (a), en la etapa de validación cruzada para la floración del banano (Las Valquirias)

Variables	$p \rightarrow a$	**Técnica**	$RMSE$		R^2	
			mean	*stdev*	*mean*	*stdev*
$\overline{Sr}$ P	$4 \rightarrow 1$	SVR/P	0,86	0,27	96,6 %	1,56 %
$\overline{Sr}$	$10 \rightarrow 4$	SVR/G	1,31	0,28	91,53 %	4,02 %
$\overline{Sr}$ P	$16 \rightarrow 4$	SVR/P	1,31	0,28	92,26 %	2,58 %
	$13 \rightarrow 7$	SVR/G	1,51	0,43	88,12 %	8,91 %
$\overline{\overline{T}_a}$ $\overline{Sr}$	$13 \rightarrow 7$	SVR/G	1,53	0,25	89,62 %	3,45 %
$\overline{Sr}$	$13 \rightarrow 10$	SVR/G	1,68	0,24	86,81 %	2,36 %
$\overline{Sr}$ P	$10 \rightarrow 13$	SVR/G	1,93	0,21	82,84 %	4,25 %
	$7 \rightarrow 16$	SVR/G	2,06	0,27	80,31 %	3,91 %
Ta-P-Sr	$4 \rightarrow 19$	SVR/G	2,26	0,32	76,69 %	3,66 %

Tabla D.5: Resultados del diseño de experimentos con una confianza del 95 % (Floración)

Factor	Métrica	H_0	Método	**Cumplimiento de Supuestos**		
				Normalidad	Independencia	Homocedasticidad
Técnicas	R^2	Rechazada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Rechazada	Kruskal-Wallis H-test	No	Si	NA
p	R^2	Aceptada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Aceptada	Kruskal-Wallis H-test	No	Si	NA
a	R^2	Rechazada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Rechazada	Kruskal-Wallis H-test	No	Si	NA
Lugar	R^2	Rechazada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Rechazada	Kruskal-Wallis H-test	No	Si	NA
Variables	R^2	Aceptada	Kruskal-Wallis H-test	No	Si	NA
	$RMSE$	Aceptada	Kruskal-Wallis H-test	No	Si	NA