Revista Especializada de Ingeniería y Ciencias de la Tierra

ISSN: L2805 -1874

Vol: 1 N°1 Julio - Diciembre 2021

Clasificación de zonas afectadas por la marchitez en banano: una aplicación con algoritmos de Machine Learning en Venezuela

Classification of areas affected by banana wilt: an application with Machine Learning algorithms in Venezuela

Barlin Orlando Olivares[1*], Andrés Vega[2], M. Angélica Rueda Calderón[2], Juan Carlos Rey[3], Deyanira Lobo[4].

[1] Universidad de Córdoba. Campus Rabanales, Programa de Doctorado en Ingeniería Agraria, Alimentaria, Forestal y del Desarrollo Rural Sostenible. Córdoba, España. Correo: barlinolivares@gmail.com. Orcid: https://orcid.org/0000-0003-2651-570X.
[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina. Correo: andresvega@agro.unc.edu.ar ORCID: https://orcid.org/0000-0002-0558-6114, Correo: angelica8511@gmail.com ORCID: https://orcid.org/0000-0002-9607-5677
[3]Instituto Nacional de Investigaciones Agrícolas (INIA), Maracay, Venezuela. Correo: jcrey67@gmail.com ORCID: https://orcid.org/0000-0001-7271-3606
[4]Universidad Central de Venezuela, Facultad de Agronomía, Venezuela. Correo: lobo.deyanira@gmail.com ORDID: https://orcid.org/0000-0002-0444-6080

Resumen

Los sistemas de producción agrícola cuentan con millones de datos que la Inteligencia Artificial (IA) puede transformar en información para favorecer la exactitud en la toma de decisiones del productor y, así, maximizar la producción de forma sostenible. El objetivo de este trabajo es clasificar zonas afectadas por la marchitez en banano de Venezuela mediante algoritmos de Machine Learning tales como: Random Forest (RF), máquinas de soporte vectorial (Support Vector Machines, SVM), árboles de clasificación (CART), el algoritmo de árboles de decisión (C5.0) y análisis discriminante lineal (ADL), así mismo se aplicaron diferentes técnicas de re-muestreo: submuestreo, sobremuestreo, sobremuestreo aleatorio (ROSE) y técnica de sobremuestreo de minorías sintéticas (SMOTE). Para ello, se realizó un muestreo de suelo sistemático en los 39 lotes de bananos y se evaluó la incidencia durante los años 2016 y 2017. Los resultados indican que RF mediante la técnica de submuestreo puede ser un algoritmo eficaz para tomar decisiones en áreas bananeras afectadas por enfermedades como la marchitez del banano. Los estadísticos de sensibilidad, especificidad, exactitud y coeficiente de Kappa fueron de 1.0, 0.94, 0.96 y 0.91 respectivamente, sin técnica de remuestreo. RF ayudaría a prevenir y reducir el efecto de enfermedades bananeras y su impacto en la producción. En conclusión, el Machine Learning en la agricultura podría ofrecer un avance que garantizaría la toma de decisiones con el objetivo de alcanzar la sostenibilidad.

Palabras clave: Algoritmos, banana, inteligencia artificial, suelos, sostenibilidad

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

Abstract

Agricultural production systems have millions of data that Artificial Intelligence (AI) can transform into information to promote accuracy in the producer's decision-making and, thus, maximize production in a sustainable way. The objective of this work is to classify areas affected by wilt in banana in Venezuela using Machine Learning algorithms such as: Random Forest (RF), Support Vector Machines (SVM), classification trees (CART), the Decision trees algorithm (C5.0) and linear discriminant analysis (ADL), likewise different resampling techniques were applied: subsampling, oversampling, random oversampling (ROSE) and synthetic minority oversampling technique (SMOTE). To do this, a systematic soil sampling was carried out in the 39 banana lots and the incidence was evaluated during the years 2016 and 2017. The results indicate that RF through the subsampling technique can be an effective algorithm to make decisions in affected banana areas. from diseases such as banana wilt. The sensitivity, specificity, accuracy and Kappa coefficient statistics were 1.0, 0.94, 0.96 and 0.91 respectively, without the resampling technique. RF would help prevent and reduce the effect of banana diseases and their impact on production. In conclusion, Machine Learning in agriculture could offer an advance that would guarantee decision-making with the aim of achieving sustainability.

Keyword: Algorithms, banana, artificial intelligence, soils, sustainability

Introducción

El banano (*Musa* spp) representa un cultivo de suma importancia para la economía de Venezuela, basada predominantemente en el petróleo. Durante los últimos 20 años la producción bananera ha sufrido ciertos cambios atribuidos principalmente al desabastecimiento de agroinsumos (semillas, fertilizantes, agroquímicos), problemas de acceso a divisas para satisfacer la demanda interna, y a la inadecuada gestión de las políticas agrícolas, así como la afectación por sequía, plagas y enfermedades (FAO, 2019).

Entre los principales factores limitantes de la producción de bananos en la región central de Venezuela se encuentran, la ocurrencia de sequías meteorológicas (Olivares, 2018), el ataque del gorgojo negro del banano (*Cosmopolites sordidus* Germar) (Rey et al. 2006; 2009); y la severa afectación de superficies de bananos a causa de las enfermedades como la sigatoka negra (*Mycosphaerella fijiensis* Morelet), sigatoka amarilla (*Mycosphaerella musicola* Leach et Mulder) (Martínez et. al. 2020) y la marchitez del banano (MB) generada por un complejo

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

hongo-bacteria, los cuales han afectado el potencial del cultivo como producto para la exportación (Martínez et al. 2016).

En las zonas bananeras del estado Aragua de Venezuela, la MB está diezmando la producción de banano Cavendish desde el 2006, recrudeciendo sus consecuencias en el rendimiento del rubro desde el 2010 (Martínez et al. 2016; Rey et al. 2016). La MB es un desorden fisiológico y metabólico; en cuyo origen podría jugar un papel importante la unión de factores bióticos y abióticos, como las condiciones físicas, químicas y los microorganismos del suelo potencialmente patógenos (Domínguez et al. 2001). Pero lamentablemente hasta la fecha no se ha podido descubrir el agente causal de la enfermedad, por lo que su control y prevención es complicada.

Actualmente, la MB puede ser confundida con la marchitez por *Fusarium oxysporum* f.sp. *cubense* (Foc) considerada como una de las más destructivas de banano a nivel mundial (Dita et al. 2018). Específicamente, la Raza 4 Tropical (TR4) ha causado graves pérdidas en el Sudeste Asiático, afectando gravemente la subsistencia de los pequeños y medianos productores, extendiéndose al continente africano y al Medio Oriente, causando preocupación de su propagación al subcontinente indio y América Latina (Pérez-Vicente & Porras, 2015), con consecuencias que pudieran ser devastadoras para los productores bananeros, y sobre todo para la cadena de valor del banano.

Los diversos estudios sobre la epidemiología de la enfermedad establecen necesario considerar los efectos del tipo del suelo (Dita et al. 2018). Estas observaciones demuestran que la propagación de la marchitez en bananos en algunas zonas bananeras fue más rápida en unas regiones que en otras, lo que lleva a correlacionar la incidencia de ese tipo de enfermedad con las propiedades específicas del suelo, como textura (Deltour et al. 2017), pH (Li et al. 2018), capacidad de intercambio de cationes (Bosman, 2016), sales solubles totales, nutrientes disponibles (Segura et al. 2015), materia orgánica y drenaje (Lahav & Israeli, 2000).

El aprendizaje automático o Machine Learning permite tanto identificar patrones entre una cantidad considerable de datos que pueden ser de diferente naturaleza como predecir comportamientos a través de algoritmos capaces de aprender y evolucionar basándose en su

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

propia experiencia (Ma et al 2017; Ye et al., 2020). Con miras a anticipar la enfermedad de MB y conseguir un diagnóstico más preciso, surge el objetivo de este trabajo el cual es clasificar las zonas afectadas por la marchitez en banano de Venezuela mediante los algoritmos: Random Forest (RF), máquinas vectoriales de soporte con Kernel Lineal (SVMkl), máquinas vectoriales de soporte con Kernel Radial (SVMkr), árboles de clasificación (CART), el algoritmo de árbol de decisión (C5.0) y análisis discriminante lineal (ADL), con el propósito de generar información valiosa capaz de detectar alteraciones en las plantas de banano afectadas por la incidencia de la marchitez y su correlación con las variables de suelo de manera objetiva y rigurosa.

Esta información resultaría muy valiosa para comprender en un sentido más amplio el problema fitosanitario que representa esta enfermedad en las plantaciones de bananos y sus relaciones con los demás componentes que conforman su sistema de producción. La determinación o separación de áreas permitiría ejecutar protocolos de prevención de la enfermedad, enfocados en las áreas críticas para evitar su propagación en la finca, considerando variables de suelo de fácil medición y económicas para el agricultor.

Método y materiales

Área de estudio

El área de estudio abarca una superficie de 180 ha de banano (*Musa paradisiaca* L.) subgrupo Cavendish, ubicada en el municipio Libertador del estado Aragua, Venezuela. El clima es tropical de sabana con una precipitación promedio anual de 1100 mm y una evaporación promedio anual entre 1800-2200 mm. Las lluvias son estacionales con 5 a 6 meses húmedos ubicados entre los meses de mayo-junio y octubre-noviembre. Los suelos en su gran mayoría son de origen lacustrino, con texturas medias, alta disponibilidad de nutrimentos, altos pH, y con condiciones salinas en forma localizada (Olivares et al. 2020).

Muestreo de suelos

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

Se realizó un muestreo de suelo sistemático siguiendo los lineamientos de Lozano et al. (2004), con una distancia aproximada de 150 m entre sitios de muestreo, comprendiendo 114 puntos de muestreo distribuidos en los 39 lotes de bananos de la finca. Se obtuvieron muestras compuestas en cada uno de los lotes de banano. Las muestras fueron sometidas a análisis de suelos con fines de fertilidad, determinando la proporción en porcentaje de arena (a), limo (L) y arcilla (A) (Gee & Or, 2002), la reacción del suelo (pH), conductividad eléctrica (CE) (dS/m) en suspensión 1:2 (suelo: agua) (Soil Survey Staff, 2014), MO: materia orgánica (%) (Heanes, 1994); los contenidos de K: potasio (mg/kg); Ca: calcio (mg/kg); Mn: manganeso (mg/kg); Fe: hierro (mg/kg); Zn: zinc (mg/kg), Cu: cobre (mg/kg); S: azufre (mg/kg) y fósforo (P) (mg/kg) (Mehlich, 1984).

Incidencia de marchitez en bananos (IMB)

Se evaluó la incidencia de la enfermedad, calculada como la proporción entre el número de plantas enfermas y el número total de plantas observadas (Akter et al. 2013) en los 39 lotes de terreno cultivado, durante los años 2016 y 2017.

Agrupamiento de las variables del suelo

Se evaluaron dos algoritmos de agrupamiento, el Método de Grupo de Pares No Ponderados con Media Aritmética, UPGMA (por sus siglas en inglés) y el K-means, utilizando las variables de suelo estandarizadas como variables de entrada. UPGMA utilizó la distancia euclidea como índice de similitud. El número óptimo de conglomerados se determinó con base en 18 índices para agrupar las variables de suelo mediante el paquete NbClust (Charrad et al. 2014). Estos índices se describen en Ostengo et al. (2020). Por lo tanto, el número de conglomerados sugerido para el agrupamiento de las variables de suelo es el que presente una mayor frecuencia de los índices evaluados.

Validación de los agrupamientos

Para comparar los conglomerados obtenidos por estos algoritmos de agrupamiento, se utilizaron los siguientes índices de validación interna: el índice de conectividad, el cual está relacionado con la distancia entre objetos en un mismo conglomerado, mientras más bajo sea el valor de

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

este índice es mejor (Kassambara 2017); el índice de ancho de silueta, el cual mide la confianza con la que una observación es asignada a un conglomerado (Sekula et al. 2017), y el índice Dunn, que es el cociente entre la mínima distancia entre dos objetos que no pertenecen a un mismo conglomerado y la máxima distancia entre dos objetos de un mismo conglomerado, combina la compactación (homogeneidad dentro del conglomerado) con el grado de separación entre conglomerados. (Dalton et al. 2009). Estos índices fueron calculados a través del paquete optCluster en R (Sekula et al. 2017).

Técnicas de re-muestreo

Para abordar el tema de desbalanceo presente en la variable categórica IMB (diferente cantidad de clases de incidencia alta o baja), se aplicaron técnicas de pre-procesamiento de datos conocida como re-muestreo o muestreo. Estas técnicas consisten en realizar modificaciones directas sobre los elementos de la base de datos con el fin de generar un equilibrio entre las distintas clases que la componen. Se usaron en este trabajo cuatro variantes de remuestreo: submuestreo (*undersampling*), sobremuestreo (*oversampling*) y sobremuestreo aleatorio (ROSE) y la técnica de sobremuestreo de minorías sintéticas (SMOTE). La técnica *undersampling* trabaja sobre la clase mayoritaria del conjunto de datos y consiste en la eliminación de muestras dentro de esta clase de tal forma que se genere un equilibrio entre clases. Por otro lado, la técnica *oversampling*, que en contra posición a la previamente nombrada, realiza su trabajo sobre la clase minoritaria del conjunto de datos, añadiendo muestras a esta clase de tal forma que se equilibre la diferencia entre clases. El sobremuestreo aleatorio (ROSE) produce una muestra sintética, posiblemente equilibrada, de datos simulados según el enfoque de remuestreo suavizado Bootstrap. Por último, SMOTE es una técnica basada en vecinos más cercanos juzgados por la distancia euclidiana entre puntos de datos en el espacio de características (Lunardon et al., 2014; Guo et al. 2017)

Algoritmos de clasificación

En este estudio, se utilizaron los siguientes algoritmos para diagnosticar la incidencia de marchitez del banano en la finca de Venezuela: Randon Forest (RF), árboles de clasificación y regresión (CART), las máquinas de vectores de soporte de núcleo de función de base radial

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

(RSVM), las máquinas de vector de soporte de núcleo lineal (LSVM), el algoritmo de árbol de decisión (C5.0) y el análisis discriminante lineal (LDA).

a. Random Forest (RF): es un conjunto de muchos árboles de regresión y clasificación individuales independientes (CART) y se define como la ecuación 1. donde, h representa el clasificador de RF, x es la variable de entrada y {θk} representa la independiente de forma idéntica variables predictoras aleatorias distribuidas, que se utilizan para generar cada árbol (Breiman, 2001). La respuesta final de la RF se calcula en función de la salida de todos los árboles de decisión.

$$\{h(x, \theta_k), k = 1,2, \dots i \dots\} \qquad (1)$$

b. Análisis discriminante lineal (LDA): es un método de clasificación supervisado que se utiliza para crear modelos de aprendizaje automático. Estos modelos basados en la reducción de la dimensionalidad se utilizan en la detección de enfermedades de las plantas (Xanthopoulos et al. 2013). Utilizando el teorema de Bayes, LDA estima la probabilidad de que una observación, dado un valor específico de los predictores, pertenezca a cada una de las clases de la variable (Ecuación 2). Finalmente, la observación se asigna a la clase k para la cual la probabilidad predicha es mayor.

$$P(\mathrm{Y} = k|\mathrm{X} = x) \quad (2)$$

c. Árboles de clasificación y regresión (CART): la representación utilizada para CART es un árbol binario. Las predicciones se hacen con CART atravesando el árbol binario dado un nuevo registro de entrada. El árbol se aprende utilizando un algoritmo codicioso en los datos de entrenamiento para seleccionar divisiones en el árbol. Los criterios de detención definen cuánto aprende el árbol y la poda se puede utilizar para mejorar un árbol aprendido (Quinlan, 2007).

d. Kernel de función de base radial Support Vector Machines (RSVM): es un clasificador de aprendizaje estadístico supervisado no paramétrico. El mayor rendimiento del clasificador SVM lo convierte en una alternativa preferida para la detección de enfermedades de las plantas. SVM considera el principio de minimización de riesgos

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

estructurales (SRM) para maximizar el margen de separación de clases para un mejor rendimiento de generalización de SVM. (Vapnik, 1995). Hay dos parámetros que deben configurarse cuando se usa un clasificador SVM con el kernel de función de base radial, es decir, la función de costo (C) y el parámetro de ancho del kernel (γ). El parámetro C compensa la clasificación errónea de los ejemplos de entrenamiento con la simplicidad de la superficie de decisión. El γ afecta la suavidad del hiperplano divisor de clases (Ye et al. 2020). La ecuación 3 muestra la Definición Matemática del Núcleo de Base Radial, donde x, x´ son puntos vectoriales en cualquier espacio de dimensión fija.

$$K(x, x') = exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \qquad (3)$$

e. Máquinas de vectores de soporte de kernel lineal (LSVM): es un modelo paramétrico (Cortes y Vapnik, 1995). Dado un conjunto de muestras x_i (i = 1,2,… M), donde M es el número de muestras. El conjunto tiene dos clases, las de clase positiva y clase negativa. Denotamos $y_i = 1$ para la clase positiva y $y_i = -1$ para la clase negativa, respectivamente. Es posible encontrar un hiperplano $f(x) = 0$ que clasifica el conjunto de datos dado (ecuación 4), donde w es un vector de dimensión M y b es un escalar, y se utilizan para definir el hiperplano (Lei, 2017).

$$f(x) = w^{\mathrm{T}}x + b = \sum_{j=1}^{M} w_j\, x_j + b = 0 \qquad (4)$$

f. Algoritmo de árbol de decisión C5.0: Los dos modos principales para este modelo son: un modelo básico basado en árboles y un modelo basado en reglas (Quinlan, 2014) C5.0 puede crear un modelo de árbol inicial y luego descomponer la estructura del árbol en un conjunto de reglas mutuamente excluyentes. Estas reglas pueden luego podarse y modificarse en un conjunto más pequeño de reglas potencialmente superpuestas (Kuhn, et al. 2018). C5.0 utiliza el concepto de entropía para medir la pureza. La entropía de una muestra de datos indica cuán mezclados son los valores de clase; el valor mínimo de 0 indica que la muestra es completamente homogénea, mientras que 1 indica la cantidad máxima de desorden. La definición de entropía se puede especificar en la ecuación 5, para un segmento dado de datos (S), el término c se refiere al número de

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

diferentes niveles de clase y p_i se refiere a la proporción de valores que caen en el nivel de clase i.

$$Entropy(S) = \sum_{I=1}^{C} -p_i \, log_2(p_i) \qquad (5)$$

Modelación estadística

Para el conjunto de datos se realizaron 30 submuestras aleatorias de la siguiente forma: 60% para entrenamiento, 15% para validación, 25% para testeo. Esto se realizó con el fin de identificar el mejor tipo re-muestreo acorde al conjunto de datos en el estudio. Para medir el desempeño de los algoritmos evaluados se consideró: el área bajo la curva ROC (Característica Operativa del Receptor), la sensibilidad, la especificidad y la precisión mediante el estadístico Kappa de Cohen aplicando una validación cruzada de 5 veces (Feuerman and Miller, 2008).

Resultados y discusión

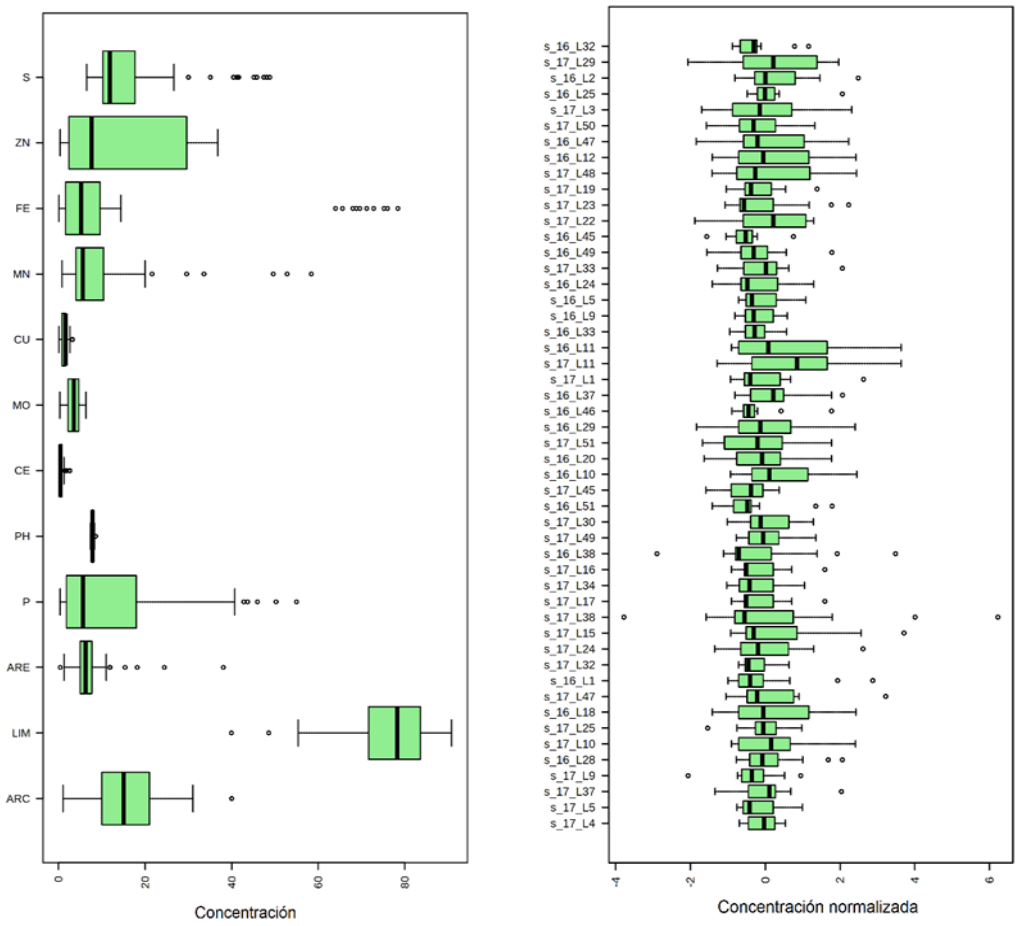
Características del suelo bananero

Los resultados del análisis descriptivo (figura 1), indicaron importantes diferencias entre las características de los suelos de los lotes, ocurriendo suelos de textura franca a franco limosa, suelos afectados ligeramente por sales (Lotes 1-5), así como la variabilidad en los contenidos de nutrimentos (fósforo y potasio) (figura 2) y las amplias diferencias en contenidos de calcio y magnesio (figura 2). En algunos lotes el pH fue ligeramente más alto que en otros, lo cual se relacionó con altos niveles de Na y bajos niveles de Fe y Mn. Por otra parte, las características particulares del material parental de los suelos generan niveles muy altos de Ca, lo cual se asocia con muy altas relaciones Ca/Mg; Ca/K (datos no mostrados) e incluso existen altas relaciones Mg/K (datos no mostrados), que provocarían desbalance nutricional y dificultad en la absorción de estos elementos.

En la mayoría de los lotes predominan las texturas francas (FL/F – FL), con predominancia de partículas con diámetro equivalente entre 2 y 50 µm, con valores bajos de densidad aparente (0,45 a 0.89 $Mg.m^{-3}$), lo cual es de esperarse, ya que se trata de suelos de origen lacustrino, cuyo

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

material parental le proporciona tal característica. Por esta razón, los valores de Porosidad Total son bastante altos (63 a 78 %). Estas características permiten que estos suelos tengan una tasa de infiltración moderada a moderadamente alta (Olivares et al. 2020). La Figura 2 está indicando la variabilidad que existe entre los lotes evaluados y que puede estar generando una incidencia diferente de la MB.



(a) (b)

Figura 1. Diagramas de caja y antes y después de la normalización. (a) Concentración de algunas variables bajo estudio: ARC: arcilla (%); LIM: Limo (%); ARE: Arena (%); P: fósforo (mg/kg); pH: reacción del suelo; CE: conductividad eléctrica (ds/m); MO: materia orgánica (%); Mn: manganeso (mg/kg); Fe: hierro (mg/kg); Zn: zinc (mg/kg); Cu: cobre (mg/kg); S:

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

azufre (mg/kg). (b) Concentración de lotes de bananos muestreados normalizados. En la figura se presentan 50 muestras (o sitios de muestreo) como referencia.

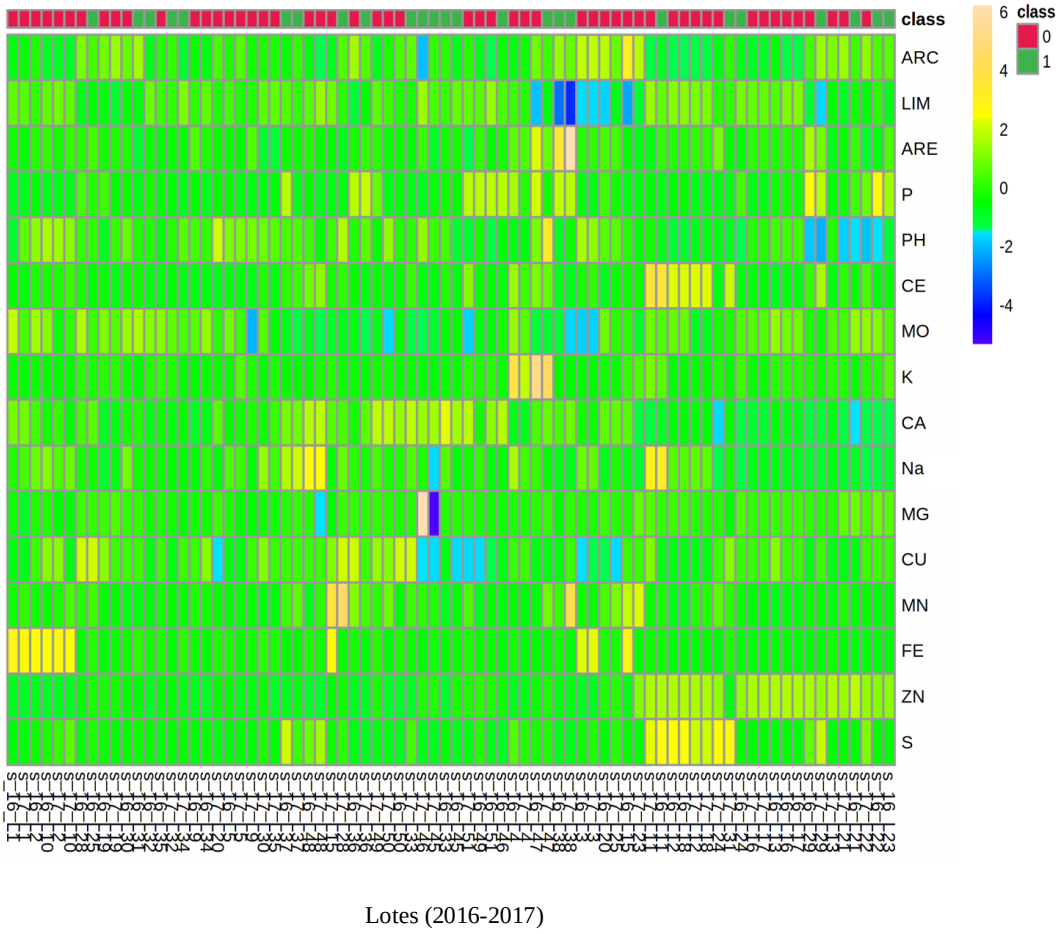


Figura 2. Mapa de calor generado a partir de datos de suelo de los lotes bananeros evaluados en el 2016 y 2017, el cual representa los valores de concentración de las variables de suelo (color azul a amarillo) para el periodo de estudio.

Incidencia de la marchitez del banano (IMB)

El transcurso de la incidencia de la marchitez del banano en el sitio evaluado se presenta en la figura 3, donde puede observarse que para el año 2016, la mayor incidencia ocurrió en el lote 38 con 5.57%, y el menos afectado fue el lote 3 con 0.54%. Con relación al año2017, la mayor proporción de incidencia se registró en el lote 36 con 8.47%, mientras que el lote 17 presentó una baja incidencia (0.11%).

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

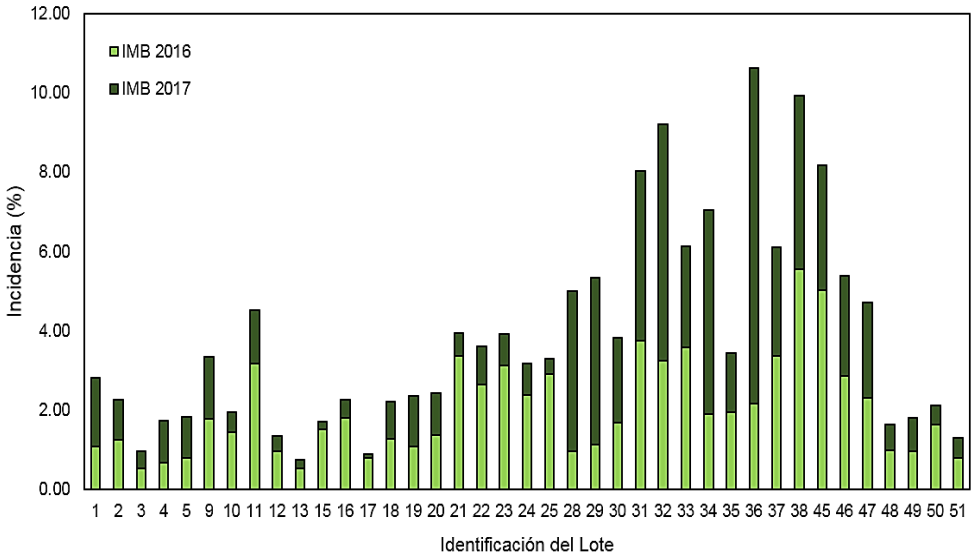


Figura 3. Transcurso de la incidencia de la marchitez del banano (IMB) durante el periodo 2016-2017 en la zona bananera.

Agrupamiento de las variables del suelo

En general, para los métodos UPGMA y K-means, los índices de validación identificaron 2 conglomerados para el agrupamiento de las variables de suelo. Ambos métodos evaluados sugieren que la conectividad más baja para UPGMA fue de 4.77 y K-means de 8.80; y los valores de silueta más altos (UPGMA: 0.37; K-means: 0.35) se obtuvieron para k = 2. Sin embargo, el índice de Dunn no fue sensible para identificar el número óptimo de conglomerados, dando valores diferentes para los dos métodos de agrupamiento, Siendo 0.47 para UPGMA y de 0.44 para K-means.

Modelación estadística

Se observó que cuando no se utilizó algún tipo de remuestreo a los datos originales (*Raw data set*) se obtuvieron, en general, valores bajos para cada uno de los criterios estadísticos usados para medir el desempeño de los algoritmos aplicados en este estudio. El uso de las técnicas de remuestreo permitió obtener mejores resultados que al considerar los datos originales. Los diferentes algoritmos y tipos de remuestreos utilizados en este estudio, se presentan en la tabla 1. En general, se observó un comportamiento similar en cuanto a que el remuestreo

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

oversampling fue el que presentó una mayor área bajo la curva ROC para todos los algoritmos evaluados. Además, se observó que el tipo de remuestreo ROSE mostró un menor desempeño para el conjunto de datos y para los algoritmos evaluados, destacándose por presentar, en general, valores bajos del área bajo la curva ROC, tanto para el conjunto de validación como el de testeo, baja precisión, baja sensibilidad y especificidad (Tabla 1). El algoritmo RF presentó, en promedio, valores altos para cada uno de los criterios usados AUC: área bajo la curva ROC, SN: sensibilidad, SP: especificidad y AC: precisión; siendo este, el algoritmo con el mejor desempeño respecto a los otros algoritmos usados en este trabajo. Por otra parte, los algoritmos LDA, LSVM, RSVM y CART fueron algoritmos que presentaron, en promedio, valores bajos en la mayoría de los criterios.

Tabla 1. Evaluación del rendimiento de diferentes algoritmos de aprendizaje automático en las pruebas: análisis discriminante lineal (LDA), árboles de clasificación y regresión (CART), bosque aleatorio (RF), algoritmo de árbol de decisión (C5.0), máquinas de soporte vectorial con kernel lineal (LSVM) y máquinas de soporte vectorial con base radial (RSVM).

Modelos	Remuestreo	Sensibilidad	Especificidad	Precisión	Precisión IC 95%	Kappa
LDA	Raw dataset	0.00	0.82	0.61	(0.39-0.80)	-0.21
	Oversampling	0.67	0.82	0.78	(0.56-0.93)	0.47
	Undersampling	0.83	0.88	0.87	(0.66-0.97)	0.68
	SMOTE	1.00	0.71	0.78	(0.56-0.93)	0.56
	ROSE	0.33	0.82	0.70	(0.47-0.87)	0.17
CART	Raw dataset	0.50	0.82	0.74	(0.52-0.90)	0.32
	Oversampling	0.50	0.24	0.30	(0.13-0.53)	-0.17
	Undersampling	0.50	0.82	0.74	(0.52-0.90)	0.32
	SMOTE	1.00	0.65	0.74	(0.52-0.90)	0.49
	ROSE	1.00	0.18	0.39	(0.20-0.61)	0.10
RF	Raw sataset	0.50	0.88	0.78	(0.56-0.93)	0.40
	Oversampling	0.83	0.71	0.74	(0.52-0.90)	0.44
	Undersampling	1.00	1.00	1.00	(0.85-1.00)	1.00
	SMOTE	1.00	0.88	0.91	(0.72-0.99)	0.80
	ROSE	0.50	0.65	0.61	(0.39-0.80)	0.13
C5.0	Raw dataset	0.50	0.82	0.74	(0.52-0.90)	0.32
	Oversampling	0.50	0.41	0.43	(0.23-0.66)	-0.06
	Undersampling	0.83	1.00	0.96	(0.78-1.00)	0.88
	SMOTE	1.00	0.76	0.83	(0.61-0.95)	0.63
	ROSE	0.67	0.65	0.65	(0.43-0.84)	0.26
LSVM	Raw dataset	0.33	0.53	0.48	(0.27-0.69)	-0.11

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

	Oversampling	0.50	0.88	0.78	(0.56-0.93)	0.40
	Undersampling	0.67	0.88	0.83	(0.61-0.95)	0.55
	SMOTE	1.00	0.71	0.78	(0.56-0.93)	0.56
	ROSE	0.33	0.82	0.70	(0.47-0.87)	0.17
RSVM	Raw dataset	0.00	1.00	0.74	(0.52-0.90)	0.00
	Oversampling	0.00	1.00	0.74	(0.52-0.90)	0.00
	Undersampling	0.50	0.88	0.78	(0.56-0.93)	0.40
	SMOTE	0.83	0.82	0.83	(0.61-0.95)	0.59
	ROSE	0.00	1.00	0.74	(0.52-0.90)	0.00

El área bajo la curva de la característica operativa del receptor (ROC) de bosque aleatorio de análisis discriminante lineal (figura 4a), bosque aleatorio (figura 4b) y máquinas vectoriales de soporte de núcleo de función de base radial (figura 4f) fueron 0,80, 0,90 y 0,84, respectivamente, con la técnica sintética de sobremuestreo minoritario (SMOTE). Mientras que el algoritmo de árboles de clasificación y regresión (figura 4c) y el algoritmo de árbol de decisión C5.0 (figura 4d) presentaron valores de área bajo la curva de 0,84 y 0,90, respectivamente, con sobremuestreo aleatorio (ROSE). Por otro lado, el algoritmo de máquinas vectoriales de soporte de kernel lineal (figura 4e) mostró un área bajo la curva de 0,81 con la técnica de submuestreo. En general, el conjunto de datos sin procesar presentó valores bajos para los modelos ajustados, lo que indica que es necesario considerar el enfoque de remuestreo para nuestro conjunto de datos.

En nuestro estudio (Tabla 2), el algoritmo RF fue significativamente más preciso que el algoritmo de árbol de decisión C5.0 con valores de coeficiente de kappa de 1.0 y 0.88, respectivamente, mediante la técnica de submuestreo (*undersampling*), así mismo RF con la técnica SMOTE obtuvo un coeficiente de Kappa de 0.80 para el conjunto de datos de prueba. Las máquinas de vectores de soporte de kernel de función de base radial se desempeñaron mejor que las máquinas de vector de soporte de kernel lineal y el algoritmo de árboles de clasificación y regresión. No hubo diferencia significativa entre SVM lineal y radial con SMOTE. Aunque los resultados no fueron significativamente diferentes, la facilidad de construcción del modelo fue mucho mayor para el bosque aleatorio que para las máquinas de vectores de soporte.

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

En el algoritmo RF, solo se ajusta un parámetro clave (número de árboles); Los modelos de máquinas vectoriales de soporte deben ajustar al menos 4-5 parámetros. Además, el significado de algunos parámetros es desconocido para los agricultores. Teniendo en cuenta la facilidad de construcción del modelo, el RF es un mejor modelo para su uso en el diagnóstico de marchitez en banano. El modelo arrojó que los suelos con alta incidencia de MB son salinos en profundidad con pH ≤ 7.2, mayor contenido de calcio (>16.000 mg/kg), zinc (>30 mg/kg) y bajos contenidos de hierro (<13 mg/kg) y azufre (< 17 mg/kg). Los resultados de la incidencia de la marchitez del banano (MB) coinciden con los reportados por González (2003), quien indica que la enfermedad se desarrolla en clima tropical y subtropical, con presencia de alta humedad y suelos con mal drenaje, con fuertes desequilibrios nutricionales. Cuando no hay aireación, la infección se produce en las raíces sanas por encontrar un exceso perjudicial de anhídrido carbónico originado por la respiración, y aunque la raíz principal es poco afectada, las raicillas laterales se enferman y quedan destruidas.

Por otra parte, Martínez et al. (2016) establecieron que las variables más relacionadas con la incidencia de la marchitez del banano en la zona de estudio son las relacionadas con la granulometría, conductividad eléctrica, carbono orgánico, nitrógeno total, contenidos de fósforo, calcio y magnesio intercambiables y la relación calcio – magnesio.

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

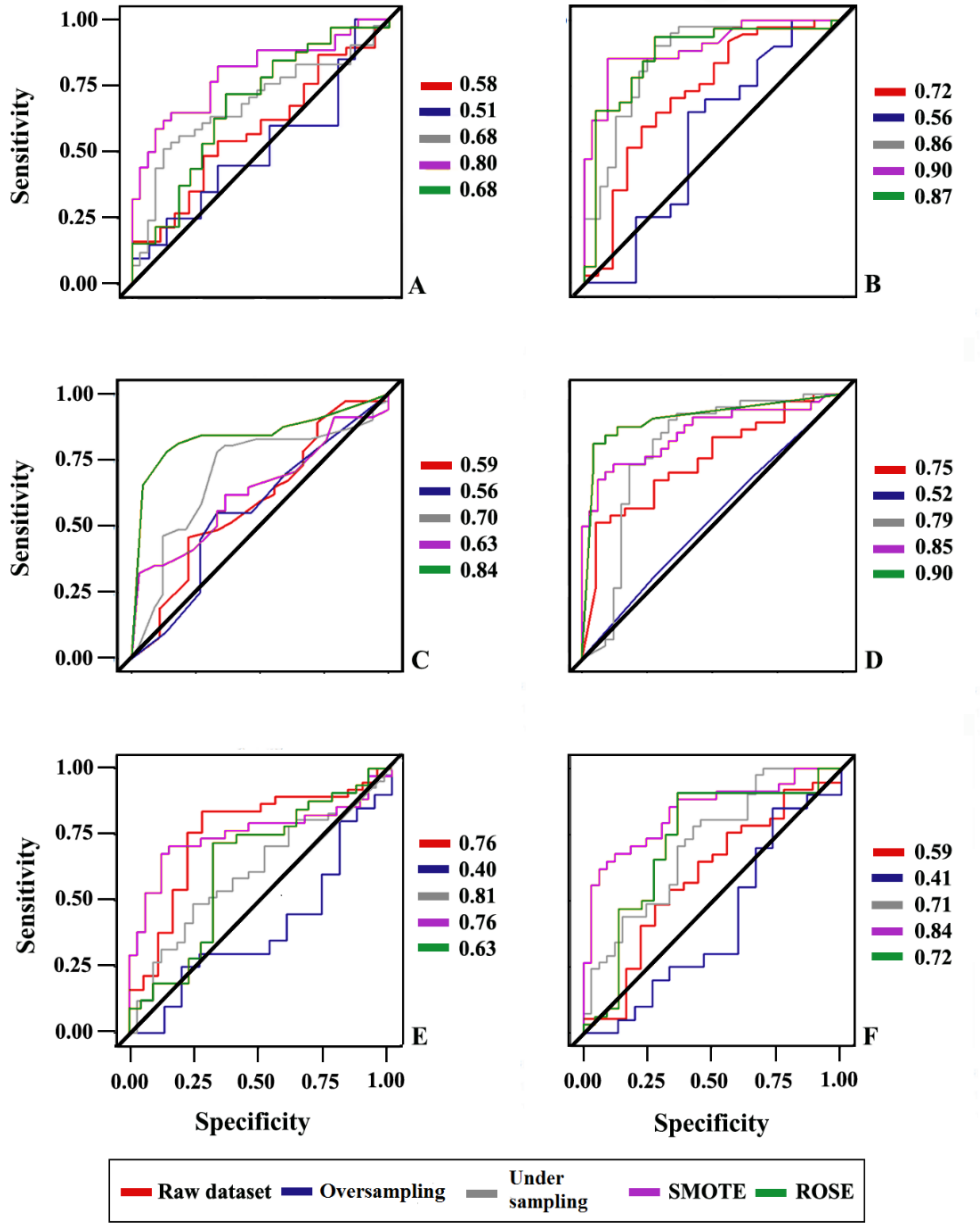


Figura 4. Curvas de características operativas del receptor (ROC) de análisis discriminante lineal (A), bosque aleatorio (B), árboles de clasificación y regresión (C), algoritmo de árbol de decisión C5.0 (D), máquinas de soporte vectoriales con kernel lineal (E) y máquinas de soporte vectoriales con kernel de base radial (F) para la sensibilidad y especificidad en la clasificación de zonas de marchitez del banano con las técnicas de remuestreo evaluados.

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

Tabla 2. Resultados de Sensibilidad, Especificidad, Precisión, el intervalo de confianza del 95% y el estadístico Kappa para los algoritmos de clasificación utilizando el criterio kappa (Kappa> 0,60) en las bases de datos original y de prueba.

Algoritmos[§]	Método de remuestreo	Sensibilidad		Especificidad		Exactitud		Exactitud IC 95%		Kappa	
		TD	RD	TD	RD	TD	RD	TD	RD	TD	RD
RF	Under Sampling	1.00	1.00	1.00	0.94	1.00	0.96	(0.85-1.00)	(0.89-0.99)	1.00	0.91
C5.0	Under Sampling	0.83	0.96	1.00	0.98	0.96	0.97	(0.78-1.00)	(0.91-1.00)	0.88	0.94
RF	SMOTE	1.00	0.92	0.88	0.80	0.91	0.83	(0.72-0.99)	(0.73-0.91)	0.80	0.65
LDA	Under Sampling	0.83	0.83	0.88	0.76	0.87	0.78	(0.66-0.97)	(0.67-0.87)	0.68	0.54
C5.0	SMOTE	1.00	0.88	0.76	0.74	0.83	0.78	(0.61-0.95)	(0.67-0.87)	0.63	0.55

[§] Algoritmos con kappa >0.60. TD: Conjunto de datos de prueba; RD: Conjunto de datos sin procesar.

La incidencia de esta enfermedad es usualmente baja, aunque hasta el 60% de una plantación puede ser afectada. Ya que la marchitez puede ser confundida fácilmente con el marchitamiento por Fusarium, a menudo no es posible para el productor bananero distinguir entre las dos enfermedades. En las regiones afectadas por el marchitamiento por Fusarium, por lo tanto, la ocurrencia de la marchitez del banano puede ser subestimada.

En este estudio, se observó que el método de remuestreo ROSE presentó valores bajos de los criterios usados tales como: área bajo la curva ROC, sensibilidad, especificidad y precisión, esto puede deberse a que este tipo de remuestreo genera datos sintéticos que distan a la realidad de los datos tomados en campo (Lunardon et al. 2014). Los algoritmos como RF y las máquinas de soporte vectorial con kernel radial (RSVM) se usan generalmente cuando las clases de una variable no pueden ser separables linealmente (Bauckhage, 2019). En nuestro trabajo, se observó que los algoritmos que presentaron un mejor desempeño fueron los que permitían una separación no lineal de las clases de IMB (RF y RSVM). Esto indicaría que las clases de IMB deberían ser abordadas desde un enfoque no separable linealmente. Por otra parte, en nuestro trabajo se observó que el desempeño más bajo lo presentaron los algoritmos LDA y RSVM; siendo estos algoritmos comúnmente usados cuando las clases son separables linealmente.

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

Estos resultados son similares a los reportados por Gómez Selvaraj et al. (2020), quienes establecen que modelos de Machine Learning basados en imágenes aéreas tienen un gran potencial para proporcionar un sistema de apoyo a la toma de decisiones para las principales enfermedades del banano en África. También, el estudio desarrollado por Ye et al. (2020) concluyó que el clasificador de RF fue el más adecuado para la identificación y mapeo de la enfermedad del marchitamiento por Fusarium del banano a partir de imágenes de detección remota basadas en el uso de Vehículos Aéreos No Tripulados. La máquina de soporte vectorial (SVM), el Random Forest (RF) y el algoritmo basado en redes neuronales artificiales (ANN) han sido usados por Ye et al. (2020) y Aruraj et al. (2019) para identificar ubicaciones que estaban infestadas o no infestadas con marchitez por Fusarium.

Conclusiones

Se evidenció que el algoritmo Random Forest (RF) permitió clasificar la incidencia de marchitez en suelos lacustrinos de Venezuela con buena precisión, es decir, puede ser una herramienta eficaz para la toma de decisiones en campo. Además, la utilización de información de suelo en zonas bananeras de Venezuela permitió identificar lotes con alta y baja incidencia de la marchitez del banano a través de un algoritmo de aprendizaje automático como RF. El modelo arrojó que los suelos con alta incidencia de MB son salinos en profundidad con pH $\leq$ 7.2, mayor contenido de calcio y zinc con bajos contenidos de hierro y azufre. Este estudio, permitiría evidenciar que, se podría anticipar la predisposición al desarrollo de la enfermedad identificando las características del suelo de una finca bananera.

Obtener información relevante a partir de los datos, no solo podría ayudar a mejorar la experiencia del agricultor, sino también a tomar decisiones pertinentes desde las primeras etapas de la aparición de los síntomas a campo. En este sentido, los algoritmos aquí evaluados permitirían analizar, interpretar y predecir futuros escenarios que ayudan a prevenir y reducir la carga de enfermedades en lotes de banano de Venezuela y su impacto en la producción.

Referencias bibliográficas

Revista Especializada de Ingeniería y Ciencias de la Tierra
ISSN: L2805 -1874
Vol: 1 N°1 Julio - Diciembre 2021

Akter H., Hassan, Md Kamrul, Rabbani, Md, Al Mahmud, Abdullah. (2013). Effects of variety and postharvest treatments on shelf life and quality of banana. Journal of Environmental Science and Natural Resources, 6(2): 163 -175 doi:10.3329/jesnr.v6i2.22113.

Aruraj, A., Alex, A., Subathra, M. S. P., Sairamya, N. J., George, S. T., Ewards, S. V. (2019, March). Detection and Classification of Diseases of Banana Plant Using Local Binary Pattern and Support Vector Machine. In *2019 2nd International Conference on Signal Processing and Communication* (ICSPC) (pp. 231-235). IEEE. doi: 10.1109 / ICSPC46172.2019.8976582

Bauckhage, C. (2019). *Lecture Notes on Machine Learning: Binary Linear Classifiers*. B-IT, Germany, University of Bonn.

Bosman, M. (2016). Role of the environment on the incidence of Panama disease in bananas. MSc thesis - Soil Geography and Landscape Master Earth and Environment (MEE). Wageningen University. Netherlands

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. doi:10.1023/a:1010933404324

Charrad, M., N. Ghazzali, V. Boiteau, and A. Niknafs. (2014). NbClust: An R package for determining the relevant number of clusters in a data set." Journal of Statistical Software, 61 (6): 1–36. doi:10.18637/jss.v061.i06.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. doi:10.1007/BF00994018

Dalton, L., V. Ballarin, and M. Brun. (2009). Clustering algorithms: on learning, validation, performance, and applications to genomics. Current Genomics, 10 (6): 430–445. doi:10.2174/138920209789177601.

Deltour, P., Franca, S. C., Pereira, O. L., Cardoso, I., De Neve, S., Debode, J., and Hofte, M. (2017). Disease suppressiveness to Fusarium wilt of banana in an agroforestry system: Influence of soil characteristics and plant community. Agriculture Ecosystems & Environment, 239, 173-181. doi:10.1016/j.agee.2017.01.018

Dita, M., Barquero, M., Heck, D., Mizubuti, ESG. and Staver, CP. (2018). Fusarium wilt of banana: current knowledge on epidemiology and research needs toward sustainable disease management. Front. Plant Sci. 9 (1468), 1-21. doi: 10.3389/fpls.2018.01468

Domínguez, J., Negrin, M. A., and Rodríguez, C. M. (2001). Aggregate water-stability, particle-size and soil solution properties in conducive and suppressive soils to Fusarium wilt of banana from Canary Islands (Spain). Soil Biology & Biochemistry, 33(4-5), 449-455. doi:10.1016/s0038-0717(00)00184-x

FAO (2019) Banana Market Review and Banana Statistics 2018. Rome. http://www.fao.org/economic/est/est-commodities/bananas/en/Accessed jan122020

Feuerman, M., & Miller, A. R. (2008). Relationships between statistical measures of agreement: sensitivity, specificity and kappa. *J Eval Clin Pract, 14*(5), 930-933. doi:10.1111/j.1365-2753.2008.00984.x

Gee, G.W and D. Or. (2002). Particle-size analysis. En: J.H. Dane ans G.C. Topp (Ed.) Methods of Soil Analysis. Part 4. SSSA Book series Nº 5, SSSA, Madison, WI. p 255-293.

Gómez Selvaraj, M., Vergara, A., Montenegro, F., Alonso Ruiz, H., Safari, N., Raymaekers, D., Ocimati, W., Ntamwira, J., Tits, L., Omondi, A. B., & Blomme, G. (2020). Detection of banana plants and their major diseases through aerial images and machine learning methods: A case study in DR Congo and Republic of Benin. *ISPRS Journal of Photogrammetry and Remote Sensing*, *169*, 110-124. doi:10.1016/j.isprsjprs.2020.08.025

González, S. (2003). Etiología y Epidemiología del "Falso Mal de Panamá" de La Platanera en Canarias. Tesis Doctoral. Instituto Canario de Investigaciones Agrarias. Tenerife, España. 286p.

Guo H., Li Y., Shang J., Mingyun G., Yuanyue H., Bing G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73: 220-239. doi:10.1016/j.eswa.2016.12.035.

Heanes D.L. (1984) Determination of total organic-C in soils by an improved chromic acid digestion and spectrophotometric procedure. Communications in Soil Science and Plant Analysis, 15, 1191-1213.

Kassambara, A. (2017). Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. STHDA. www.sthda.com

Kuhn M, Weston S, Culp M, Coulter N, Quinlan R. (2018). Paquete «C50». https://cran.r-project.org/web/packages/C50/C50.pdf

Lahav, E. & Israeli, Y. (2000). Mineral deficiencies of Banana. In Diseases of banana, Abacá and enset. (Ed. D. Jones). CABI Publishing: Wallingford, Oxon, UK. pp. 339-350.

Lei, Y. (2017). 3 - Individual intelligent method-based fault diagnosis. In Y. Lei (Ed.), *Intelligent Fault Diagnosis and Remaining Useful Life Prediction of Rotating Machinery* (pp. 67-174). Butterworth-Heinemann. doi:10.1016/B978-0-12-811534-3.00003-2

Li, Z., Deng, Z., Chen, S., Yang, H., Zheng, Y., Dai, L., Zhang, F., Wang, S., Hu, S. (2018). Contrasting physical and biochemical properties of orchard soils suppressive and conducive to Fusarium wilt of banana. Soil Use and Management, 34(1), 154-162. doi:10.1111/sum.12390

Lozano P., Z., Bravo C., Ovalles F., Hernández R.M., Moreno B., Piñango, L., Villanueva, J.G. (2004). Selección de un diseño de muestreo en parcelas experimentales a partir del estudio de la variabilidad espacial de los suelos. Bioagro, 16(1),61-72. https://n9.cl/nhvf

Lunardon, N.; Menardi, G.; Torelli, N. (2014). ROSE: A package for binary imbalanced learning. The R Journal, 6(1), 82-92.

Ma L., Fu T., Blaschke T., Li M., Tiede D., Zhou Z., et al. (2017). Evaluation of feature selection methods for object-based land cover mapping of unmanned aerial vehicle imagery using random forest and support vector machine classifiers. ISPRS Int. J. Geo-Inf, 6(2), 51. doi: 10.3390/ijgi6020051

Martínez, G. E., Rey-Brina, J.C., D. Rodríguez, Jiménez, C., Rodríguez, Y., Rumbos, R., R. Pargas-Pichardo, Martínez, E. (2020) Análisis de la situación fitopatológica actual de las musáceas comestibles en Venezuela. *Agro. Trop.* 70 1-20.

Martínez, G.; J. C. Rey; L. Castro; E. Micale; O. López; R. Pargas; y E. Manzanilla. (2016). Marchitez en banano Cavendish en la Región Central de Venezuela, asociado a un complejo hongo – bacteria. Reunión ACORBAT, 2016. Miami USA. Memorias.

Mehlich A. (1984). Mehlich 3 soil test extractant: a modification of Mehlich 2 extractant. Comm Soil Sci Plant Anal, 15:1409-1416

Olivares B.O. (2018). Tropical rainfall conditions in rainfed agriculture in Carabobo, Venezuela. Lgr Lif Sci J. 27:86-102. doi:10.17163/lgr.n27.2018.07

Olivares, B., Araya-Alman, M., Acevedo-Opazo, C. et al. (2020). Relationship Between Soil Properties and Banana Productivity in the Two Main Cultivation Areas in Venezuela. J Soil Sci Plant Nutr. 20 (3): 2512-2524. doi:10.1007/s42729-020-00317-8

Ostengo S, Rueda Calderón M. A, Bruno C, Cuenya M.I, Balzarini M. (2020) Selecting sugarcane genotypes (*Saccharum* spp.) according to sucrose accumulation, Journal of Crop Improvement, 34:2, 190-205. doi:10.1080/15427528.2019.1683783

Pérez-Vicente, L. and Porras, Á. (2015). Impacto potencial del cambio climático sobre las plagas de bananos y plátanos en Cuba. Fitosanidad, 19 (3), 201-211.

Quinlan, J. R. (2014). C4. 5: programs for machine learning. Amsterdam, Netherlands: Elsevier.

Quinlan, J.R. (2007). Decision Trees as Probabilistic Classifiers. In: Proceedings of the Fourth International Workshop on Machine Learning. Massachusetts, United States: Morgan Kaufmann Publishers, Inc.

Rey J.C., Chacín M., Sapuky M., Núñez M., Martínez G., Rodríguez G., Espinoza J., Arturo M., Pocasangre L., Delgado E., Rosales F. (2006). Aptitud de las tierras para banano en suelos de Venezuela y su relación con la productividad. XVII Reunión Internacional ACORBAT: Banano un negocio sustentable. Joinville. Santa Catarina, Brasil. Nov 15-20. p. 362.

Rey J.C., Martínez G., Lobo D., Trejos J., Pocasangre L., Rosales F. (2009). Aspectos sobre calidad y salud de suelos bananeros en Venezuela. Producción Agropecuaria, 2: 52-55

Rey, J. C.; G. Martínez; N. Pizzo; E. Micale; N. Fernández. (2016). Áreas susceptibles a la enfermedad falso mal de Panamá en banano Cavendish, en la Región Central de Venezuela. Reunión ACORBAT, 2016. Miami USA. Memorias.

Segura, R. A., Serrano, E., Pocasangre, L., Acuna, O., Bertsch, F., Stoorvogel, J. J., Sandoval, J. A. (2015a). Chemical and microbiological interactions between soils and roots in

commercial banana plantations (*Musa* AAA, cv. Cavendish). Scientia Horticulturae, 197, 66-71. doi:10.1016/j.scienta.2015.10.028

Sekula, M., S. Datta, S. Datta. (2017). OptCluster: An R package for determining the optimal clustering algorithm. Bioinformation, 13 (3): 101–103. doi:10.6026/bioinformation.

Soil Survey Staff. (2014). Kellogg Soil Survey Laboratory Methods Manual. Soil Survey Investigations Report No. 42, Version 5.0. R. Burt and Soil Survey Staff (ed.). U.S. Department of Agriculture, Natural Resources Conservation Service.

Vapnik. V. N. (1995). The nature of statistical learning theory. Berlin, Heidelberg: Springer-Verlag.

Xanthopoulos, P., Pardalos, P. M., Trafalis, T. B. (2013). Linear discriminant analysis. In: *Robust data mining* (pp. 27-33). Springer, New York, NY. doi:10.1007/978-1-4419-9878-1_4

Ye H. C., Huang W. J., Huang S. Y., Cui B., Dong Y. Y., Guo A. T., Ren Y., Jin Y. (2020). Identification of banana fusarium wilt using supervised classification algorithms with UAV-based multi-spectral imagery. Int J Agric & Biol Eng, 13(3): 136–142. DOI: 10.25165/j.ijabe.20201303.5524