**Mode****lo predictivo para el rendimiento de cultivos de cacao en Santander basado en herramientas de aprendizaje automático supervisado**

PhD. Henry Lamos Díaz1, David E. Puentes Garzón2, Andrea C. Gamboa Ariza3, Paula A. Cáceres Ortiz4

Escuela de Estudios Industriales y Empresariales -EEIE

Grupo de Investigación Ópalo

Universidad Industrial de Santander

hlamos@uis.edu.co1, dpuentesgarzon@gmail.com2, andrea1496gamboa@gmail.com3, paulacaceresor@gmail.com4

Abstract

En el presente artículo se exponen dos modelos para la predicción de rendimientos de cacao en Santander: Modelo Lineal Generalizado (GLM) y Máquinas de Soporte Vectorial (SVM). Los GLM son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (binomiales, Poisson, gamma, etc.) y varianzas no constantes. Estos modelos son una alternativa a la transformación de la variable respuesta y a la falta de normalidad en el modelo, mientras que una Máquina de Soporte Vectorial (SVM) aprende la superficie decisión de dos clases distintas de los puntos de entrada. Diámetro del tronco, Fósforo (P), Magnesio (Mg), %Arena, %Hum/Grav, Radiación, Temperatura, Humedad y Lluvias acumuladas fueron las variables que, tanto en el Modelo Lineal Generalizado como en las Máquinas de soporte Vectorial, fueron definidas como las variables que influyen significativamente en el rendimiento del cultivo de cacao. Así mismo, dichos modelos se evaluaron con diferentes métricas de ajuste obtenidas a partir de validación cruzada, obteniendo resultados de R^2:(0.1534) RMSE: (144.1014) MAE: (854.4128) para SVM Y R^2: (0.2336) RMSE: (143.9239) MAE: (928.545) para GLM; resultados importantes al momento de predecir rendimientos de cacao.

1. Introducción

La agricultura es una de las actividades de mayor contribución al crecimiento económico de la población en Colombia. Particularmente, el cultivo de cacao contribuye en gran medida a dicho crecimiento, debido a que fue el cultivo que más creció porcentualmente en producción; de 2016 a 2017 se produjo un incremento en la producción del 6.6% el cual creó un récord para el país en este sector [1].

Del mismo modo, el sector cacaotero está catalogado como uno de los sectores estratégicos en el departamento de Santander con un porcentaje de participación en el área nacional sembrada del 24% y una producción de aproximadamente 26.431,64 Ton/Año [2].

En aras de contribuir con uno de los objetivos del Plan de Desarrollo Departamental, el cual busca “fortalecer la agricultura familiar de tal forma que se garantice la seguridad alimentaria”, se origina el presente trabajo en el que por medio de las metodologías de aprendizaje automático: Modelo lineal Generalizado y Máquinas de Soporte Vectorial, se pretende predecir el rendimiento de cultivos de cacao.

Las herramientas de aprendizaje automático han sido aplicadas para la predicción de diferentes cultivos agrícolas en trabajos como los de Chen et al. (2016) & Chattopadhyay & Mitra (2018), en los que se utilizó gran cantidad de datos de entrada para predecir la variable respuesta para cultivos de arroz y granos respectivamente.

Tanto Máquinas de Soporte Vectorial como Modelo Lineal Generalizado, han sido técnicas de Aprendizaje Automático que han ayudado a encontrar influencias, dentro de los cultivos, de factores climáticos, morfológicos y de suelo, y con ello, han facilitado a los diferentes actores a formular estrategias adecuadas para hacerle frente a las variables con mayor influencia dentro de los cultivos.

En ese orden de ideas, predecir el rendimiento para los cultivos alimenticios y determinar variables influyentes en ellos, es de gran ayuda para que agricultores, empresas y gobierno puedan comprender las condiciones mas influyentes y en consecuencia, mejorar la adaptabilidad de los cultivos tomando decisiones acertadas anticipadamente con el objetivo de obtener mayor rendimiento en los cultivos.

2. Metodología

El conjunto de datos es el insumo fundamental de todo análisis predictivo, el cual es necesario para encontrar la variable respuesta, puesto que es la fuente que contiene la información de variables y su relación entre ellas. El conjunto de datos empleado para el presente trabajo fue suministrado por AGROSAVIA (Corporación Colombiana de Investigación Agropecuaria) a partir de un cultivo experimental ubicado en el centro de investigación La Suiza para los años 2015, 2016 y 2017. Acto seguido, se utilizó el programa PYTHON para la construcción y posterior análisis de los dos modelos predictivos con herramientas de aprendizaje automático: Modelo Lineal Generalizado y Máquinas de Soporte Vectorial. Las variables contenidas en los conjuntos de datos de fotosíntesis, morfometría, características físicas y químicas del suelo y variables climáticas se utilizaron como variables independientes o explicativas, y la variable dependiente a predecir fue el rendimiento de los cultivos de cacao.

2.1. Datos

Para efectos de la presente investigación, el conjunto de datos de entrada que se emplearán para el desarrollo y validación del modelo fueron suministrados por AGROSAVIA (Corporación Colombiana de Investigación Agropecuaria) los cuales fueron tomados a partir de un cultivo experimental ubicado en el centro de investigación La Suiza para los años 2015, 2016 y 2017. El cultivo experimental de cacao está compuesto por 3 factores: fertilización, clon y exposición. El primero de ellos cuenta con 3 niveles: fertilización al 50%, 100%, 150%. En segundo lugar, se tienen los 10 tipos de clones más representativos de Santander clasificados en 5 regionales (SCC19, SCC-52, SCC-61, SCC-64 Y SCC-83) y 5 universales (ICS-95, CNN-51, ETT-8, TSH565 y ICS-1). Y, por último, la exposición del cultivo es a sol o a sombra. Considerando estos factores y niveles se cuenta con un total de 60 tratamientos. A cada uno de los tratamientos mencionados anteriormente, le corresponden 15 plantas, por lo que se tiene un total de 900 de ellas a las cuales se les miden sus características fotosintéticas, morfológicas, físicas y químicas del suelo. De igual manera, el cacao producido por estas plantas se evalúa en términos de altura de la planta, número de ramas, diámetro del tronco, peso de la almendra, peso de la mazorca completa, entre otros. En primer lugar, de las características fotosintéticas fueron tomadas 3 muestras, 1 vez al año durante el periodo de 2015-2017 para los 60 tratamientos, para un total de 540 observaciones. Por otro lado, se cuenta con 2160 observaciones para las características morfológicas de la planta tomadas una vez por semestre, realizando 3 repeticiones para 2 árboles por repetición para los 60 tratamientos. A su vez, se cuenta con 7239 observaciones de condiciones ambientales del cultivo experimental, las cuales fueron medidas con ayuda de sensores que realizaron recuentos para el periodo comprendido entre 2015 y 2017.

Para el presente trabajo, la variable a predecir; la misma variable dependiente, es el rendimiento de los cultivos, y las variables independientes o variables explicativas son las que conforman los conjuntos de datos de fotosíntesis, morfometría, características físicas y químicas del suelo y de variables climáticas.

2.2. Selección de variables

Partiendo de que una correlación es la forma numérica en la que la estadística ha podido evaluar la relación recíproca o mutua entre dos o más variables, es decir, se mide la dependencia de una variable con respecto de otra variable independiente, se realizó una correlación para cada conjunto de datos con el fin de descartar las variables con mayor similitud entre ellas (mayor coeficiente de correlación) y así poder analizar para cada modelo aquellas variables con menor coeficiente de correlación, es decir, aquellas con menor similitud entre ellas.

En base a la correlación del conjunto de datos de las variables de fotosíntesis, morfometría, características físicas y químicas del suelo y de variables climáticas, fueron seleccionadas las variables: Transpiración, Uso eficiente del agua, Diámetro del tronco, Materia Orgánica (MO), Sodio (Na), Fósforo (P), Magnesio (Mg), %Arena, %Hum/Grav, Radiación, Temperatura, Humedad y Lluvias acumuladas como variables de entrada para los modelos.

2.3. Modelo Lineal Generalizado

Nelder y Baker propusieron por primera vez el modelo lineal generalizado que construye una extensión flexible de los modelos lineales, y considera distribuciones no normales de la variable dependiente [3]. Los GLM son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (binomiales, Poisson, gamma, etc.) y varianzas no constantes, donde la media está relacionada con las variables explicativas. Estos modelos son una alternativa a la transformación de la variable respuesta y a la falta de normalidad en el modelo.

Un modelo GLM está constituido por un componente aleatorio (término error: ε), una función de los factores de diseño (x) y coeficientes β:

y ≈ β0 + β1x1 + β2x2 + β3x3 + ε (1)

Y por una función enlace g (µ) que determina cómo se relaciona la media con a las variables explicativas x:

g(µ) = x! β (2)

De la importancia de determinar el comportamiento que sigue la variable respuesta para determinar la función enlace, se realizó una prueba de hipótesis para los valores observados de la variable respuesta. De la figura 1, se observa el valor p el cual es significativo para rechazar la hipótesis nula, el cual indica una distribución Gamma de la variable respuesta.

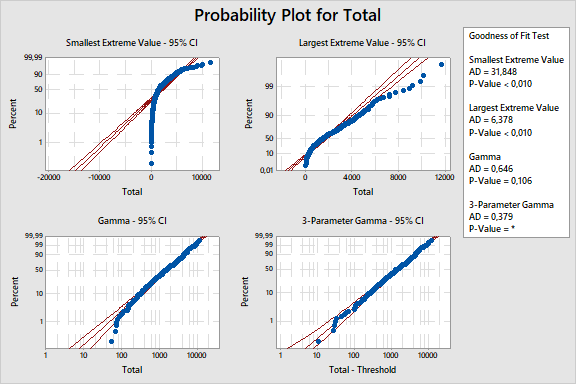


Figura 2. Distribución variable respuesta

Así mismo, de la figura 2, se puede observar una distribución característica Gamma; de variables continuas no negativas y sesgadas a la derecha con grandes valores en la cola superior

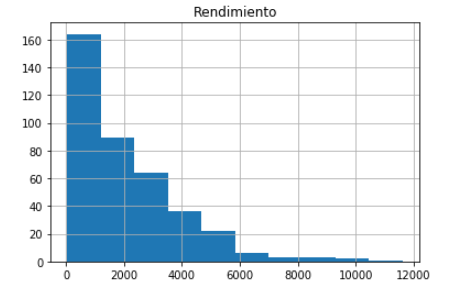


Figure 2. Histograma variable respuesta

Dado lo anterior, de la función enlace Gamma, se tiene la función de probabilidad G (µ, ν):

f (y) = , y>0

E(y) = µ, Var(y) = µ2 ν

Cabe resaltar que el enlace canónico para la distribución Gamma es la función inversa, sin embargo, ya que los parámetros de un modelo con enlace inverso son difíciles de interpretar, el enlace LOG se suele considerar como más útil [4].

2.3. Máquinas de Soporte Vectorial

El análisis de la máquina de soporte vectorial es una herramienta popular de aprendizaje automático para la clasificación y la regresión, identificada por primera vez por Vladimir Vapnik y sus colegas en 1992. La regresión SVM se considera una técnica no paramétrica porque se basa en las funciones del kernel.

Statistics and Machine Learning Toolbox ™ implementa una regresión SVM (ε-SVM) insensible a épsilon lineal, que también se conoce como pérdida L 1. En la regresión ε -SVM, el conjunto de datos de entrenamiento incluye variables predictoras y valores de respuesta observados. El objetivo es encontrar una función f (x) que se desvíe de y n en un valor no mayor que ε para cada punto de entrenamiento x, y al mismo tiempo sea lo más plano posible [5].

2.3.1. Regresión de SVM lineal

Se necesita un conjunto de datos de entrenamiento donde x n es un conjunto multivariado de N observaciones con valores de respuesta observados y  *n*

Para encontrar la función lineal:

f (x ) = x ′ β + b ,

sujeto a todos los residuos que tengan un valor menor que ε; o, en forma de ecuación:

∀ n : yn - (xn ′ β + b) ≤ ε .

Es posible que no exista tal función f (x) para satisfacer estas restricciones para todos los puntos. Para hacer frente a las restricciones de otro modo inviables, se deben introducir las variables de holgura ξ n y ξ \* n para cada punto. Este enfoque es similar al concepto de "margen blando" en la clasificación SVM, porque las variables de holgura permiten que existan errores de regresión hasta el valor de ξ n y ξ \* n , y aun así satisfacen las condiciones requeridas.

Incluir variables de holgura conduce a la función objetivo, también conocida como la fórmula primordial

J ( β ) = 12 β′β+CNn=1(ξn + ξ∗n) ,

La constante C es la restricción de la caja, un valor numérico positivo que controla la penalización impuesta a las observaciones que se encuentran fuera del margen de épsilon ( ε ) y ayuda a evitar el sobreajuste (regularización). Este valor determina la compensación entre la planitud de f ( x ) y la cantidad hasta la cual se toleran las desviaciones mayores que ε.

2.3.1. Regresión de SVM no lineal

Algunos problemas de regresión no pueden describirse adecuadamente usando un modelo lineal. En tal caso, la formulación dual de Lagrange permite que la técnica descrita anteriormente se extienda a funciones no lineales. Se obtiene un modelo de regresión SVM no lineal reemplazando el producto de puntos x 1 ′ x 2 con una función de núcleo no lineal G ( x 1 , x 2 ) = < φ ( x 1 ), φ ( x 2 )> , donde φ ( x ) es una transformación que mapea x a un espacio de alta dimensión. Statistics and Machine Learning Toolbox proporciona las siguientes funciones integradas de kernel semidefinido [5].



3. Resultados

(Intro)

3.1. GLM

Para el modelo lineal generalizado se trabaja con las variables independientes seleccionadas, mencionadas anteriormente. En primera instancia, el modelo GLM se corrió con todas ellas, como se muestra en la figura 3.

Cada una de las variables tiene un valor p; el cual indica que las que tengan valores por debajo de 0.05, son aquellas variables que tienen mayor incidencia en la variable respuesta, las cuales son: 'Diámetro tronco (cm)', 'P', 'Mg', '%A', '%Hum/Grav', 'Radiación', 'Temperatura', 'Humedad', 'Lluvias acumuladas (mms)'.

Es por esto último que posteriormente se entrena y valida el modelo con las variables de valores p menores que 0.05; aquellas que mayor incidencia tienen en la variable respuesta. Los resultados se muestran en la figura 4.

Como se puede observar en la figura 4, de los valores de p menores a 0.05, las variables: 'Diámetro tronco (cm)', 'P', 'Mg', '%A', '%Hum/Grav', 'Radiación', 'Temperatura', 'Humedad', 'Lluvias acumuladas (mms)' son aquellas variables representativas dentro del modelo.

De cada uno de sus coeficientes se concluye que:

Para la variable 'Diámetro tronco (cm)', con un coeficiente positivo de 0.0622, se deduce que parece influir positivamente en la variable dependiente rendimiento. La variable ‘P', con un coeficiente positivo de 1.0123, parece influir positivamente en el rendimiento.

La variable ‘Mg', con un coeficiente positivo de 2.5123, parece influir positivamente en la variable dependiente.

La variable ‘%A', con un coeficiente negativo de -0.2323, parece influir negativamente en el rendimiento del cultivo.

De los coeficientes positivos de las variables ‘%Hum/Grav', 'Radiación' y 'Lluvias acumuladas, de 0.3997, 4.662e-07 y 0.0083 respectivamente, se puede deducir que estarían influyendo positivamente en la variable dependiente.

De las variables 'Temperatura' y 'Humedad' con coeficientes de -0.2630 y -34.5760 respectivamente, se podría estar concluyendo que estas variables en elevada proporción estarían influyendo negativamente al rendimiento del cultivo.

3.2. SVM

Basándonos en los conceptos teóricos y para efectos de evaluación de los dos modelos, el modelo de máquinas de soporte vectorial se entrenó y validó con las mismas variables independientes del modelo GLM; aquellas que mostraron mayor incidencia en la variable dependiente.

Luego de escalar las variables se encontraron los mejores parámetros para las máquinas de regresión obteniéndose un valor de 10000 para C y el modelo se entrena y valida con un kernel lineal; generando para él las predicciones para el rendimiento.

Los coeficientes generados en las máquinas de soporte vectorial para cada una de las variables se muestran a continuación en la figura n:

Donde se evidencia que el signo de los coeficientes para cada una de las variables del modelo SVM.

Se observa que la incidencia de las variables comparada con el modelo GLM es similar.

Luego de generar los coeficientes para cada variable se concluye que:

Para la variable “Diámetro del tronco”, con un coeficiente positivo de 460,47348037, se deduce que parece influir positivamente en la variable dependiente rendimiento. La variable Fósforo(P), con un coeficiente positivo de 1676,49904291, parece influir positivamente en el rendimiento.

La variable ‘Mg', con un coeficiente positivo de 739.47395887, parece influir positivamente en la variable dependiente.

La variable “%A”, con un coeficiente negativo de -1014,12586028, parece influir negativamente en el rendimiento del cultivo.



De los coeficientes positivos de las variables ‘%Hum/Grav',

'Radiación' y 'Lluvias acumuladas, de 885.0244371,

538.83755418 y 1468.66102623 respectivamente, se puede

deducir que estarían influyendo positivamente en la variable

dependiente

De las variables 'Temperatura' y 'Humedad' con coeficientes

de -278.53260164, -1549.99072426 respectivamente, se

podría estar concluyendo que estas variables en elevada

proporción, estarían influyendo negativamente al rendimiento

del cultivo.

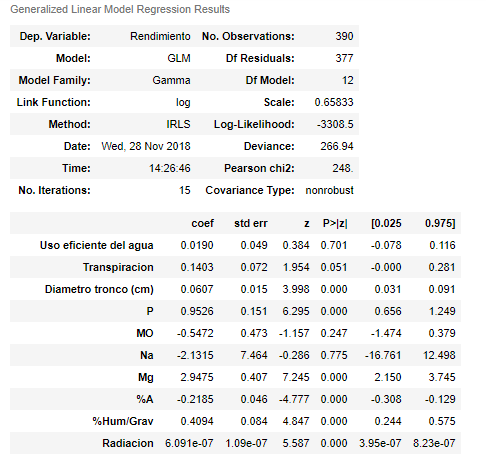
****

Figure 3. Resultados Modelo GLM

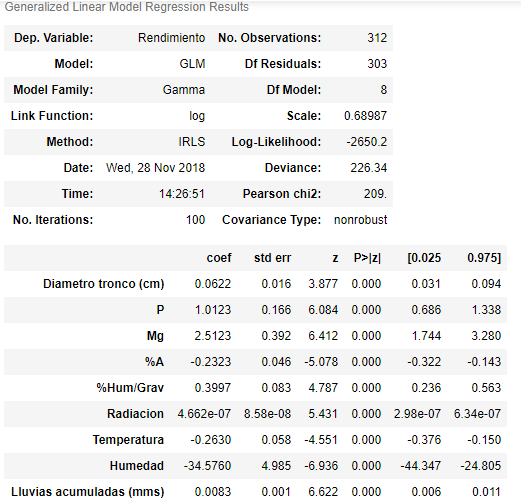
****

Figure 4. Resultados Modelo GLM