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

Andrea C. Gamboa Ariza1, Paula A. Cáceres Ortiz2 Henry Lamos Díaz, PhD3, David E. Puentes Garzón4,

Escuela de Estudios Industriales y Empresariales -EEIE

Grupo de Investigación Ópalo

Universidad Industrial de Santander

andrea1496gamboa@gmail.com1, paulacaceresor@gmail.com2, hlamos@uis.edu.co3, dpuentesgarzon@gmail.com4

Resumen

Las herramientas de aprendizaje automático representan una buena alternativa para el sector agrícola, dado que permiten apoyar a los agricultores, gobierno y demás actores del sector en la toma de decisiones a partir de pronósticos en los rendimientos de los cultivos [1], proporcionando así información para mejorar su productividad. Por este motivo, en el presente trabajo se tiene como objeto de estudio un cultivo experimental de cacao en Santander, ubicado en el centro de investigación La Suiza, donde se utiliza el Modelo Lineal Generalizado (GLM) y las Máquinas de Soporte Vectorial (SVM) como herramientas para predecir el rendimiento del cultivo y donde a partir de ellas se consiguen identificar las variables: Diámetro del tronco, Fósforo (P), Magnesio (Mg), %Arena, %Hum/Grav, Radiación, Temperatura, Humedad y Lluvias acumuladas como variables representativas para predecir el rendimiento del cacao.

**Palabras clave:** Aprendizaje Automático, predicción, cultivos, Santander.

.

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 ha corroborado esta afirmación, 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% lo que representó un récord para el país en este sector [2].

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 [3].

A pesar de los volúmenes de producción y el área destinada al cultivo, se ha observado un escenario no tan favorable cuando se evalúan los rendimientos agrícolas (producción / área sembrada) [4], por lo que es preciso emplear alternativas que permitan mejorar esta productividad como las de aprendizaje automático.

Las herramientas de aprendizaje automático han sido aplicadas para la predicción de diferentes cultivos agrícolas en trabajos como los de Chen et al. [5]; en donde se utilizaron variables explicativas tales como temperatura media, humedad relativa, horas de sol y días de lluvia para los cultivos de arroz con el fin de predecir el rendimiento de dicho cultivo. Por otro lado, los autores Chattopadhyay & Mitra [6], emplean las herramientas de aprendizaje automático y las variables de entrada: Lluvias estacionales y área de producción de granos, para realizar un estudio con el fin de predecir la producción de granos alimenticios y a su vez, encontrar las variables más influyentes dentro de cultivos ubicados en India. Diferentes cultivos han sido materia de investigación para diferentes autores; los cultivos de uva lo fueron para los investigadores Cunha, Ribeiro, & Abreu [7], los cuales realizaron pronósticos para su producción. “La confiabilidad y la capacidad de los modelos realizados con herramientas de aprendizaje automático justifican su uso para apoyar en el proceso de toma de decisiones” es lo que afirman dichos autores luego de su investigación.

De acuerdo con lo anterior, predecir el rendimiento para los cultivos y determinar variables influyentes en ellos, es de gran ayuda para que agricultores, empresas y gobierno puedan comprender las condiciones más influyentes y, en consecuencia, mejorar la adaptabilidad de los cultivos tomando decisiones acertadas anticipadamente con el objetivo de obtener mayor rendimiento en los cultivos. Es por este motivo que para la presente investigación se utilizan las Máquinas de Soporte Vectorial y el Modelo Lineal Generalizado, como técnicas de Aprendizaje Automático para identificar los elementos que influyen en los rendimientos de cultivos de cacao y a su vez realizar pronósticos.

En la sección 2 se muestra la metodología tomada para el desarrollo de la investigación. Posteriormente, en la sección 3, se muestran los resultados alcanzados con la aplicación de los modelos. En la sección 4, se pueden observar las métricas de ajuste implementadas para cada uno de los modelos. Mientras que en la sección 5, se exponen las conclusiones asociadas a los hallazgos. Finalmente, en la sección 6 se presentan algunas recomendaciones dirigidas a los diferentes actores del sector agrícola.

2. Metodología e implementación de modelos

Para la implementación de las herramientas de Aprendizaje Automático en los cultivos de cacao, se hace uso del lenguaje de programación Python, en el cual se realizó la construcción de modelos tanto de Máquinas de Soporte Vectorial como de Modelo Lineal Generalizado. Para los dos modelos, se optó por la partición de los conjuntos de datos 80/20; el 80% del conjunto de datos fue utilizado para el entrenamiento de los algoritmos y el otro 20% se utilizó para la prueba de estos últimos. A continuación, se describe de forma detallada la implementación de cada uno de los modelos.

2.1. Datos

El conjunto de datos empleado para el análisis fue suministrado por AGROSAVIA (Corporación Colombiana de Investigación Agropecuaria) a partir de un cultivo experimental ubicado en el centro de investigación La Suiza en Rio Negro, Santander, 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 (SCC-19, 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 a sol o a sombra. Considerando estos 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. Adicionalmente, se obtiene información de variables climáticas y se acude a fuentes secundarias como el IDEAM para obtener información sobre la influencia de las lluvias en el cultivo.

Cabe resaltar que, para los datos perdidos, se realiza una imputación de datos utilizando la media, consolidando de esta manera el Dataset y posteriormente realizando a este su debido preprocesamiento; con una normalización sobre las variables independientes, que son las que conforman los conjuntos de datos de fotosíntesis, morfometría, características físicas y químicas del suelo y de variables climáticas, y teniendo como variable a predecir el rendimiento del cultivo de cacao.

2.2. Selección de variables

Se toma cada uno de los conjuntos de datos previamente analizados, a los cuales se les realiza un análisis de correlación con el fin de seleccionar aquellas variables que tengan baja correlación y así evitar problemas de multicolinealidad en los modelos.

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 [8] propusieron por primera vez el modelo lineal generalizado que construye una extensión de los modelos lineales y considera distribuciones no normales de la variable dependiente [9]. Los Modelos Lineales Generalizados son una alternativa cuando la variable respuesta pertenece a la familia de distribuciones exponenciales (Binomiales, Poisson, gamma, etc.) con varianzas no constantes; donde la media está relacionada con las variables explicativas.

Un modelo GLM está constituido por un componente aleatorio (término error: ε), variables explicativas (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 de la variable respuesta con las variables explicativas x:

g(µ) = x! β (2)

Considerando la variable respuesta: rendimiento del cultivo de cacao, se plantea una prueba de hipótesis con el fin de evaluar la distribución de probabilidad que sigue, encontrándose que con un valor de *p* de 0.106 > 0.05, se afirma que la variable respuesta sigue unadistribución de probabilidad tipo Gamma, motivo por el cual se ratifica el uso adecuado del Modelo Lineal Generalizado. 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 Logístico se suele considerar como más útil [10], motivo por el que se decide tomar este último. Cabe resaltar que para el modelo GLM se tomó inicialmente el 100% de los datos con el fin de identificar aquellas variables que tenían mayor significancia dentro del modelo y posteriormente se optó la partición 80/20 de los datos para entrenamiento y prueba del modelo.

2.3. Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial o también llamadas SVMs por sus siglas en ingles “Support Vector Machines”, fueron desarrolladas originalmente por Vapnik [11]. Inicialmente las máquinas de soporte vectorial eran utilizadas para problemas de clasificación, pero con el paso del tiempo se vio necesario que también fueran aplicadas en problemas de regresión. La principal idea de las SVMs es construir un hiperplano como superficie de decisión tal que el margen de separación entre los ejemplos positivos y negativos sea máximo [12], esto se conoce como hiperplano óptimo de separación.

Para efectos de la presente investigación, se implementan las SVMs, de modo que la predicción de la variable respuesta, depende de la función del kernel evaluada en un subconjunto de los puntos de datos de entrenamiento. Las variables de entrada para el modelo de Máquinas de Soporte Vectorial son las mismas variables previamente identificadas como significativas en el modelo GLM, de modo que para los dos modelos se entrenó y validó con las mismas variables y conjunto de datos.

Otra propiedad importante de las SVMs es la determinación de los parámetros del modelo. La constante C, Gamma y kernel son parámetros con los que cuenta este modelo. La constante C es conocida como un parámetro de penalización en términos del error, es decir controla la compensación entre los errores de entrenamiento y los márgenes rígidos, creando así un margen que permita algunos errores en la clasificación a la vez que los penaliza [13]. Por otra parte, el parámetro gamma define hasta qué punto llega la influencia de un solo ejemplo de entrenamiento, con valores bajos que significan "lejos" y valores altos que significan "cerca". Para mayor comprensión, un Gamma bajo, significa que los puntos están alejados de la línea de separación, mientras que un gamma con valor alto presenta cercanía de los puntos a la línea de separación [14]. De la importancia de la identificación de hiperparámetros adecuados y luego de escalar las variables, se utilizó la validación cruzada para la identificación de parámetros y para ello la función *GridSearch* del lenguaje Python para el presente modelo, de modo que se encontró que los parámetros más adecuados fueron: 1000 para la constante C, 0.01 para gamma y un kernel lineal, los cuales fueron implementados tanto en el entrenamiento como en la prueba del presente modelo.

3. Resultados

3.1. GLM

Para la implementación del Modelo Lineal Generalizado, se realiza el respectivo entrenamiento y prueba tomando como variables explicativas aquellas previamente identificadas como significativas. En la tabla 1 se muestran dichas variables, donde se observan valores *p* inferiores a 0.05, lo cual indica que cada uno de sus coeficientes de regresión es diferente de 0 y, por tanto, son variables consideradas como significativas dentro del Modelo Lineal Generalizado.

|  |  |  |
| --- | --- | --- |
| *Variable* | *Coef* | *p* |
| *1. Diámetro tronco* | 0.062 | 0.000 |
| *2. P* | 1.0123 | 0.000 |
| *3. Mg* | 2.5123 | 0.000 |
| *4. %A* | -0.2323 | 0.000 |
| *5. %Hum/Grav* | 0.3997 | 0.000 |
| *6. Radiación* | 4.662e-07 | 0.000 |
| *7. Temperatura* | -0.2630 | 0.000 |
| *8. Humedad* | -34.5760 | 0.000 |
| *9. Lluvias acumuladas* | 0.0083 | 0.000 |

Tabla 1. Resultados modelo GLM

De la misma manera, se evidencia que las variables 1, 2, 3, 5, 6 y 9; como se muestra en la tabla 1, presentan coeficientes positivos, es decir, que estas influyen de manera positiva sobre los rendimientos del cultivo de cacao y permiten explicar la variabilidad de la variable respuesta. Por otro lado, el resto de las variables con coeficiente negativo pueden estar impactando de manera negativa el rendimiento del cacao.

Por otro lado, en la figura 1, se puede observar cómo se ajustan los datos del modelo GLM. Cabe resaltar que el coeficiente de regresión toma valores bajos de ajuste porque se está dando prioridad a explicar la variabilidad de la variable respuesta.

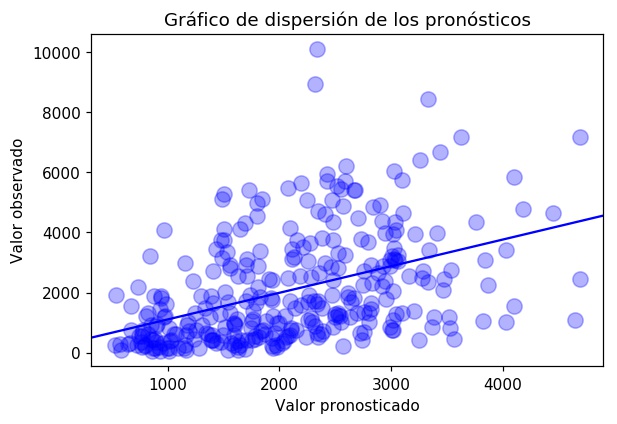


Figura 1. Ajuste pronósticos modelo GLM

3.1. SVM

El modelo de Máquinas de Soporte Vectorial se entrenó y validó con las mismas variables independientes seleccionadas previamente en el modelo GLM. Los coeficientes generados en las Máquinas de Soporte Vectorial para cada una de las variables se muestran a continuación en la tabla 2.

|  |  |
| --- | --- |
| *Variable* | *Coef* |
| *1. Diámetro tronco* | 502.09608563 |
| *2. P* | 1949.81204821 |
| *3. Mg* | 1009.58241061 |
| *4. %A* | -1330.99002137 |
| *5. %Hum/Grav* | 1056.58241061 |
| *6. Radiación* | 698.90374679 |
| *7. Temperatura* | -281.03554591 |
| *8. Humedad* | -1698.00205152 |
| *9. Lluvias acumuladas* | 1597.48227769 |

Tabla 2. Resultados modelo SVM

Como se puede observar de la tabla 2, las variables 1, 2, 3, 5, 6 y 9 presentan coeficientes positivos, con lo que se concluye que, al igual que con el Modelo Lineal Generalizado, estas variables influyen de manera positiva sobre los rendimientos del cultivo de cacao y permiten explicar la variabilidad de la variable respuesta. Por otro lado, el resto de las variables con coeficiente negativo pueden estar impactando de manera negativa el rendimiento del cacao.

4. Validación de modelos

Para la validación de los modelos de Máquinas de Soporte Vectorial y Modelo Lineal Generalizado, se utilizó la técnica Hold Out, la cual consiste en tomar el conjunto completo de datos disponibles y dividirlo aleatoriamente en dos subconjuntos; el primero de ellos son los datos aplicados a la fase de entrenamiento y el segundo para la fase de prueba. Generalmente, se consideran dos terceras partes para el conjunto de entrenamiento y la otra tercera parte para el conjunto de prueba [15]. El conjunto de entrenamiento se subdivide, a su vez, en dos conjuntos:

1. El conjunto de estimación, usando para seleccionar el algoritmo.
2. El conjunto de validación, usando para probar o validar el algoritmo.

La idea principal de la segunda división está en validar el algoritmo sobre un conjunto de datos diferente del empleado para estimar sus parámetros [16].

En primera instancia, para el modelo de Máquinas de Soporte Vectorial con esta técnica, se obtienen valores de 0,1058 para R2; 1758,7421 de RMSE (Root Mean Square Error) y de 926,2623 para MAE (Median Absolute Error). Mientras que, para el Modelo Lineal Generalizado, se tiene un valor de R2 de 0,1319; 1708,2904 de RMSE y de 1028,7805 para MAE.

A partir de la técnica Hold Out y de cada uno de los valores de error RMSE, se planteó una prueba de hipótesis sobre la igualdad de medias para los dos modelos, tal como se muestra en la tabla 3, encontrando un valor de *p* de 0.016 < 0,05, lo que indica que hay evidencia significativa para afirmar que hay diferencias entre los valores medios de RMSE y que, a su vez, el error RMSE del modelo GLM es menor al modelo SVM, lo que da a entender que existe mayor precisión en la predicción para el modelo GLM.

|  |  |  |  |
| --- | --- | --- | --- |
|  | RMSE | Intervalo de confianza | Valor p |
| SVM | 1758,7421 | [-132,30 ; -14,0] | 0,016 |
| GLM | 1700,0615 |

5. Conclusiones

De la investigación realizada, se puede concluir que las herramientas de Aprendizaje Automático son bastante apropiadas dentro de la ciencia para la predicción de la variable respuesta y a su vez, para encontrar las variables explicativas con mayor influencia en las investigaciones de interés. Los modelos realizados ayudan a reducir la incertidumbre y a su vez a comprender las condiciones más influyentes con el fin de tomar las decisiones mas acertadas. Para la presente investigación, los modelos mostraron que las variables: Diámetro del tronco, Fósforo (P), Magnesio (Mg), %Arena, %Hum/Grav, Radiación, Temperatura, Humedad y Lluvias acumuladas son las variables que están influyendo en mayor proporción de manera positiva y negativa sobre la variable respuesta: el rendimiento de los cultivos de cacao en Santander.

6. Recomendaciones

Se recomienda a los actores del sector tomar en consideración los resultados de este tipo de investigaciones para seguir mejorando los rendimientos de los diferentes cultivos. A su vez, se sugiere la toma de nuevas observaciones para aumentar la variabilidad en otros cultivos, clones y regiones del departamento. Se invita a los investigadores continuar trabajando en la construcción de modelos haciendo uso de las diferentes metodologías de Aprendizaje Automático, con el fin de mejorar la precisión. De igual forma, podría llegar a ser interesante un análisis más profundo enfocado en las variables climáticas; variables que mejor explican el modelo. Finalmente, continuar trabajando en conjunto con diferentes grupos de interés dentro de la investigación, para lograr cada vez mejores resultados.

7. Agradecimientos

Agradecemos el apoyo de la Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA) por el suministro de los datos para llevar a cabo la presente investigación y esperamos sea de apoyo para las diferentes federaciones de agricultores.

8. Referencias

[1] Hansen, J. W., Potgieter, A., & Tippett, M. K. (2004). Using a general circulation model to forecast regional wheat yields in northeast Australia. Agricultural and Forest Meteorology, 127(1–2), 77–92. https://doi.org/10.1016/j.agrformet.2004.07.005

[2] DANE - PIB 2017. (n.d.). Retrieved August 9, 2018, from http://www.dane.gov.co/ index.php/52-espanol/noticias/noticias/4505-pib-oferta-iv-trimestre-201

[3] FEDECACAO. (2018). Retrieved August 9, 2018, from http://www.fedecacao.com.co/portal/index.php/es/2015-04-23-20-00-33/551-en-2017colombia-alcanzo-nuevo-record-en-produccion-de-cacao

[4] CIAT. (2017). Retrieved August 13, 2018, from https://blog.ciat.cgiar.org/es/que-papel-puedejugar-el-cacao-para-la-paz-en-colombia/

[5] Chen, H., Wu, W., & Liu, H.-B. (2016). Assessing the relative importance of climate variables to rice yield variation using support vector machines. Theoretical and Applied Climatology, 126(1–2), 105–111. https://doi.org/10.1007/s00704-015-1559-y

[6] Chattopadhyay, M., & Mitra, S. K. (2018). Assessing the predictability of different kinds of models in estimating impacts of climatic factors on food grain availability in India. Opsearch, 55(1), 50–64. https://doi.org/10.1007/s12597-017-0314-9

[7] Cunha, M., Ribeiro, H., & Abreu, I. (2016). Pollen-based predictive modelling of wine production: application to an arid region. European Journal of Agronomy, 73, 42–54. https://doi.org/10.1016/j.eja.2015.10.008

[8] Nelder JA, Baker RJ (1972) Generalized linear models. Encyclopedia of Statistical Sciences

[9] Cayuela, L. (2010). Modelos lineales generalizados (GLM).

[10] P. de Jong and G. Z. Heller (2008). Generalized Linear Models for Insurance Data. 978-0-521-87914-9.

[11] Vapnik VN (1998) Statistical learning theory. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York

[12] Fu, X., Ying, Y., Xu, H., 2009. Quantitative analysis of fruit firmness by near infrared spectroscopy and least-squares support vector machine. American Society of Agricultural and Biological Engineers, St. Joseph, MI, ASAE paper number: 096475

[13] Gavidia Bovadilla Dirigida por, G. E., & Oñate Ibañez de Navarra Ing Eduardo Soudah Prieto, E. (n.d.). Clasificadores Basados en Máquinas de Soporte Vectorial para el Diagnóstico y Predicción de la Enfermedad de Alzheimer. Retrieved from https://web.cimne.upc.edu/users/esoudah/publications/Tesis\_Gio\_2012\_SVM\_Alzheimer.pdf

[14] Chapter 2: SVM (Support Vector Machine) — Theory – Machine Learning 101 – Medium. (n.d.). Retrieved December 8, 2018, from https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

[15] Referencia, M. De, & Mendoza, L. L. P. O. (2008). Índice General.

[16] Rapporti, R. A. I., & Dell, A. (2012). Ra ■ i rapporti annuali dell’arte.