AGRICULTURA PREDICTIVA

TRABAJO FINAL

ESTUDIANTES

Maria Fernanda Moreno De La Espriella

Jhon Sahian Alvarez Pantoja

FECHA

04/11/2025

INGENIERÍA AGRONÓMICA

UNIVERSIDAD EAFIT

MEDELLIN

2025

PARTE 1. Clasificación de la aptitud para el establecimiento del cultivo de cacao (Theobroma cacao L.) en Colombia a partir de variables climáticas mediante modelos de aprendizaje automático.

El presente trabajo tuvo como propósito determinar qué modelo de aprendizaje supervisado describe mejor la relación entre variables climáticas y la aptitud del terreno para el cultivo de cacao (*Theobroma cacao* L.) en Colombia. Se implementaron cuatro enfoques representativos: Extreme Gradient Boosting (XGBoost), Random Forest (RF), Red Neuronal Artificial (ANN) y Regresión Logística. Cada modelo se configuró con diferentes combinaciones de hiperparámetros, definidas mediante una búsqueda en cuadrícula (GridSearchCV) y validadas con una validación cruzada estratificada de cinco particiones, lo que permitió garantizar una comparación justa y confiable entre algoritmos.

El conjunto de datos provino de Kaggle e integró información de producción agrícola y condiciones de cultivo. Las variables climáticas predictoras se obtuvieron de la base NASA POWER, con una resolución espacial de 0.5° × 0.5°, adecuada para análisis regionales. Se consideraron indicadores clave del ambiente: radiación solar incidente (ALLSKY_SFC_SW_DWN), precipitación total corregida (PRECTOTCORR), radiación fotosintéticamente activa (CLRSKY_SFC_PAR_TOT), humedad relativa (RH2M), velocidad del viento (WS2M) y temperaturas máxima y mínima (T2M_MAX, T2M_MIN).

Para cada una de estas variables se incluyeron dos medidas: el promedio (*average*) y la desviación estándar (*std*). Usar ambas es importante porque el promedio refleja las condiciones medias del clima en cada zona, mientras que la desviación estándar muestra qué tan variables son esas condiciones. Esto permite al modelo tener en cuenta no solo si un lugar tiene, por ejemplo, buena temperatura o radiación en promedio, sino también si esas condiciones se mantienen estables o cambian mucho a lo largo del tiempo, lo cual puede afectar el desarrollo del cultivo.

El dataset incluyó información de los años 2019, 2022 y 2023, con 57,659 registros por año. Los tres conjuntos se unieron en una sola base de 288,291 registros, y se eliminó la variable “año” para evitar que el modelo aprendiera patrones por periodos específicos. Se corrigieron pequeñas inconsistencias en las coordenadas y se conservaron los valores extremos, ya que representan condiciones reales del territorio colombiano.

Durante el preprocesamiento, las variables se estandarizaron con *StandardScaler* (media 0, desviación 1), la variable objetivo se codificó con *One-Hot Encoding*, y se eliminaron predictores con alta correlación (> 0.8) o muy baja variabilidad (< 1e−6).

El análisis de balance de clases mostró una ligera diferencia entre categorías (92,426 alta, 131,519 media y 64,345 baja), lo que significa que el modelo podía inclinarse a predecir con más frecuencia la clase mayoritaria. Para evitar este problema, se aplicaron pesos balanceados durante el entrenamiento *con class weight = balanced.*

Esta técnica ajusta la importancia de cada clase según la cantidad de datos que tenga: las clases con menos ejemplos reciben un peso mayor, y las más frecuentes un peso menor. De esta forma, el modelo “presta más atención” a las clases minoritarias y aprende a reconocerlas con la misma relevancia. Esto ayuda a mejorar la precisión general y evita que el modelo tenga sesgos hacia las clases con más observaciones, sin necesidad de modificar artificialmente el tamaño del conjunto de datos (como ocurriría con el sobremuestreo o submuestreo).

Los mejores valores de hiperparámetros obtenidos para cada modelo se presentan en la Tabla 1 y las métricas obtenidas por modelo se presentan en la Tabla 2.

Tabla 1. *Parámetros evaluados mediante Grid Search y parámetros óptimos utilizados en cada modelo de clasificació.*

Modelo	Parámetros en Grid Search	Parámetros utilizados
Regresión Logística	C =[0.01, 0.1, 1, 10, 100] penalty = [L1, L2]	C= 1, Penalty = L2
Random Forest	n_estimators= [50, 100, 150] max_depth = [10, 15, 20] criterion = [gini]	Max_depth = 20, n_estimators = 50
Red Neuronal	Hidden_layer_sizes = [(50,), (100,)] Activation = relu Solver = [adam] Alpha = [0.0001, 0.001]	Alpha = 0.0001, Hidden_layer_sizes = (100,)
XGBoost	N/A	objective='multi:softprob' learning_rate=0.1, n_estimators=300,

Tabla 2. *Desempeño de los modelos de clasificación según las métricas de F1-Score por clase y Accuracy general.*

Modelo	Aptitud Alta F1-Score	Aptitud Media F1-Score	Aptitud Baja F1-Score	Accuracy
Regresión Logística	0.42	0.29	0.41	0.37
Random Forest	0.88	0.9	0.87	0.88
Red Neuronal	0.66	0.71	0.61	0.67
XGBoost	0.84	0.86	0.83	0.84

Los modelos de aprendizaje supervisado evaluados permitieron estimar la aptitud del terreno para el cultivo de cacao en Colombia a partir de variables climáticas obtenidas del conjunto de datos de NASA POWER y la zonificación de la UPRA. La **Regresión Logística**, tras la optimización mediante *Grid Search*, presentó su mejor desempeño con una penalización L2 y un nivel intermedio de regularización, lo que favoreció un equilibrio entre flexibilidad y control de los coeficientes (Sperandei, 2014). No obstante, su naturaleza lineal limitó la capacidad del modelo para capturar relaciones complejas entre las variables climáticas, lo que se reflejó en un desempeño bajo (F1-score promedio = 0.38). Este comportamiento coincide con lo reportado por Talero-Sarmiento et al. (2025), quienes indican que los modelos lineales tienden a infrarepresentar la interacción entre variables ambientales no lineales en la predicción de la aptitud del cacao.

Por su parte, el modelo **Random Forest** alcanzó el mejor rendimiento general, con un *accuracy* y F1-score promedio de 0.88. Este resultado se asocia con la capacidad del modelo para manejar relaciones no lineales y la interacción entre múltiples variables climáticas, así como su robustez ante clases desbalanceadas gracias al uso de *class_weight = balanced* (Belgiu & Drăguţ, 2016). De manera similar, Talero-Sarmiento et al. (2025) encontraron que Random Forest ofrece ventajas significativas para la clasificación de la aptitud del cacao, debido a su habilidad para integrar información de temperatura, humedad y velocidad del viento sin sobreajuste.

El modelo **XGBoost**, configurado con una tasa de aprendizaje media y un número intermedio de estimadores, alcanzó un rendimiento comparable (F1-score = 0.84). Su naturaleza basada en boosting le permitió mejorar

progresivamente las predicciones corrigiendo los errores de iteraciones anteriores, reforzando su capacidad para modelar relaciones complejas. Esto coincide con los hallazgos de Gawdiya et al. (2024), quienes reportan que los modelos de boosting son especialmente eficaces en contextos agrícolas con datos heterogéneos y correlacionados.

En contraste, la **Red Neuronal Artificial (ANN)** mostró un rendimiento intermedio (accuracy = 0.67). Aunque la activación ReLU y el optimizador Adam favorecieron la convergencia, la arquitectura con una sola capa oculta resultó insuficiente para representar patrones jerárquicos complejos. Talero-Sarmiento et al. (2025) destacan que redes con mayor profundidad o combinadas con técnicas de ensamblaje pueden mejorar notablemente la capacidad predictiva en escenarios agroclimáticos.

Los resultados confirman que los modelos de tipo *ensemble*, particularmente Random Forest y XGBoost, son los más adecuados para la clasificación de la aptitud del terreno con base en variables climáticas. Su fortaleza radica en su capacidad para modelar interacciones no lineales y capturar la heterogeneidad ambiental (Meshram et al., 2021). En comparación con estudios previos, como el de Talero-Sarmiento et al. (2025), la ausencia de variables edáficas como la humedad del suelo en distintos niveles podría explicar por qué la precisión fue ligeramente inferior. Dichas variables, al reflejar la disponibilidad hídrica del perfil del suelo, se consideran determinantes para el desarrollo del cacao (Acheampong et al., 2019; Niether et al., 2020).

A pesar de su bajo desempeño, la Regresión Logística permitió establecer una línea base interpretativa y evidenciar la naturaleza no lineal del fenómeno. La Red Neuronal, aunque menos precisa, mostró potencial de mejora con arquitecturas más profundas o la incorporación de normalización y regularización avanzadas.

En conjunto, los resultados demuestran que los modelos Random Forest y XGBoost ofrecen las mejores métricas de precisión y generalización en la clasificación de la aptitud del cacao a partir de variables climáticas. Ambos lograron capturar las interacciones complejas entre temperatura, humedad y radiación solar, factores determinantes en la fisiología del cultivo. Estos hallazgos coinciden con los de Talero-Sarmiento et al. (2025), quienes evidencian que los modelos ensamblados permiten obtener recomendaciones agrícolas más precisas al integrar múltiples fuentes de información ambiental.

Asimismo, la comparación con la literatura sugiere que la inclusión de variables edáficas y microclimáticas mejoraría significativamente la capacidad predictiva. Incorporar parámetros como la humedad del suelo y la radiación fotosintéticamente activa (PAR) indicadores críticos para la productividad del cacao (Saavedra et al., 2020) fortalecería la identificación de zonas aptas bajo condiciones de cambio climático.

En conclusión, los modelos de tipo *ensemble* se consolidan como herramientas robustas para la planificación agroclimática y la toma de decisiones basadas en datos, permitiendo desarrollar estrategias más sostenibles y resilientes frente a la variabilidad ambiental que afecta al cacao en Colombia.

PARTE 2. Evaluación de la sensibilidad de modelos de clasificación ante variaciones en la estructura de los datos mediante experimentos sintéticos.

Para el desarrollo del segundo punto se implementó un conjunto de tres experimentos con el fin de analizar la sensibilidad de distintos modelos de clasificación (SVM, KNN Y RF) ante variaciones en la estructura de los datos. Se utilizaron datos sintéticos generados artificialmente a partir de distribuciones normales multivariadas, con cuatro clases y cinco variables numéricas continuas, sin ningún significado físico o aplicado. La construcción de estos datos permitió controlar de manera precisa la escala, la dispersión y la proporción de observaciones por clase, de modo que se pudieran simular distintos escenarios de comportamiento de los modelos sin depender de un conjunto real.

Los algoritmos evaluados fueron Máquina de Vectores de Soporte (SVM) con kernel radial, K-Nearest Neighbors (KNN) con cinco vecinos y Random Forest (RF) con 300 árboles, todos implementados con sus configuraciones estándar en la librería scikit-learn. Los tres modelos se entrenaron y evaluaron bajo las mismas condiciones, utilizando una división de 70 % para entrenamiento y 30 % para prueba, y las métricas de desempeño consideradas fueron exactitud (Accuracy), precisión, Recall y F1 macro, complementadas con la visualización de matrices de confusión.

El primer experimento consistió en analizar el efecto de la escala de las variables sobre el rendimiento de los modelos, comparando los resultados antes y después de aplicar la estandarización con StandardScaler. El objetivo fue observar qué tan dependientes son los clasificadores de las magnitudes originales de los datos. Posteriormente, en el segundo experimento se evaluó el impacto de la presencia de valores atípicos (outliers). Para ello, se añadieron muestras anómalas en el conjunto de entrenamiento, alterando de forma aleatoria algunas dimensiones de cada clase, y se analizaron tres condiciones: datos sin escalar con outliers, datos estandarizados limpios y datos estandarizados pero contaminados con los mismos outliers. Finalmente, el tercer experimento se centró en el desbalanceo de clases, generando tres escenarios con diferentes proporciones de observaciones (balanceado, desbalance medio y desbalance extremo), a fin de determinar cómo las diferencias en la representación de cada clase afectan el desempeño general y la capacidad de los modelos para reconocer categorías minoritarias.

RESULTADOS Y DISCUSIÓN

Tabla 3. *Métricas de desempeño (Accuracy y F1-score por clase) obtenidas en los tres escenarios de balanceo de datos (balanceado, desbalanceo medio y desbalanceo extremo) para los modelos SVM, KNN y Random Forest.*

Escenario	Modelo	Accuracy	F1_macro	F1_clase_0	F1_clase_1	F1_clase_2	F1_clase_3
Balanceado	SVM	0.879	0.878	1.0	0.968	0.768	0.775
Balanceado	KNN	0.854	0.851	1.0	0.949	0.736	0.718
Balanceado	RF	0.865	0.864	1.0	0.965	0.733	0.757
Desbalanceo medio	SVM	0.974	0.922	1.0	0.984	0.906	0.8
Desbalanceo medio	KNN	0.95	0.82	1.0	0.965	0.828	0.488
Desbalanceo medio	RaF	0.964	0.89	1.0	0.981	0.868	0.712
Desbalanceo extremo	SVM	0.978	0.782	1.0	0.983	0.828	0.316
Desbalanceo extremo	KNN	0.971	0.801	1.0	0.952	0.729	0.522
Desbalanceo extremo	RF	0.978	0.838	1.0	0.975	0.817	0.56

Tabla 4. *Promedio de métricas globales (Accuracy, F1 macro, Precision macro y Recall macro) calculadas por escenario de balanceo en los tres modelos evaluados.*

Accuracy	F1_macro	Precision_macro	Recall_macro
0.866	0.864	0.866	0.866
0.976	0.807	0.865	0.791
0.963	0.877	0.914	0.865

Los resultados del Experimento 1 muestran que la estandarización mejoró el rendimiento de SVM y KNN, mientras que Random Forest permaneció prácticamente igual. Esto coincide con la literatura: los modelos basados en distancia son sensibles a la escala de los datos, ya que las variables de mayor magnitud dominan el cálculo de distancias, afectando la frontera de decisión (Hsu et al., 2016). En contraste, los algoritmos de árboles como Random Forest no dependen de distancias y por tanto no requieren normalización para mantener su desempeño (Pedregosa et al., 2011; Kuhn & Johnson, 2013).

En el Experimento 2 se observó que la presencia de valores atípicos afecta de forma distinta a los tres modelos evaluados. En SVM, la exactitud disminuyó significativamente con datos contaminados (Acc = 0.757) y mejoró tras la estandarización limpia (Acc = 0.898), evidenciando su alta sensibilidad a outliers, tal como señalan Hodge & Austin (2004) y Sohrab et al. (2021), quienes describen que los puntos extremos distorsionan el hiperplano de decisión. De manera similar, KNN mostró ligeras variaciones (Acc ≈ 0.83–0.87), coherentes con su vulnerabilidad a muestras ruidosas en el vecindario (García et al., 2015; Wilson & Martínez, 2000). En contraste, Random Forest mantuvo un desempeño estable (Acc ≈ 0.88–0.89), confirmando su robustez ante ruido y datos contaminados, atribuida a la agregación de múltiples árboles y su carácter no paramétrico (Breiman, 2001; Fernández-Delgado et al., 2014).

Los resultados del Experimento 3 (tabla 3 y 4) mostraron que, a medida que aumenta el desbalance de clases, la exactitud general mejora (0.86 a 0.97), pero el desempeño equilibrado medido por F1 macro disminuye, evidenciando que los modelos tienden a favorecer las clases mayoritarias (Johnson & Khoshgoftaar, 2019). Este patrón se refleja en los valores de F1 por clase, donde la clase 1, la que tiene más muestras, mantiene valores cercanos a 1, mientras que las clases 2 y 3 muestran caídas progresivas, especialmente bajo desbalance extremo. Esto indica que el aprendizaje del modelo se concentra en las categorías con más datos y disminuye su capacidad para reconocer correctamente las clases con menos ejemplos, fenómeno descrito en estudios sobre sesgo de distribución en clasificadores tradicionales (Fernández et al., 2018).

Entre los modelos, SVM y KNN mostraron una mayor degradación del F1 macro, mientras que Random Forest mantuvo un rendimiento más estable (F1, 0.84), coherente con su robustez derivada del voto agregado de múltiples árboles (Chen, 2004). En conjunto, los resultados confirman que el desbalance puede generar una aparente mejora en *accuracy* sin reflejar un verdadero aumento del rendimiento general, subrayando la importancia de analizar métricas como F1 y los resultados por clase para identificar posibles sesgos hacia las categorías más representadas.

UNIVERSIDAD EAFIT

REFERENCIAS

Acheampong, K., Daymond, A. J., Adu-Yeboah, P., & Hadley, P. (2019). *Improving field establishment of cacao (Theobroma cacao) through mulching, irrigation and shading. Experimental Agriculture, 55*(6), 898–912. https://doi.org/10.1017/S0014479718000479

Belgiu, M., & Drăguț, L. (2016). *Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114,* 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

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

Chen, C. (2004). *Using random forest to learn imbalanced data.* Department of Statistics, University of California, Berkeley. https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). *Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research (JMLR), 15,* 3133–3181. https://doi.org/10.5555/2627435.2697065

Fernandez, A., Garcia, S., Galar, M., Prati, R., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets.* Springer International Publishing. https://doi.org/10.1007/978-3-319-98074-4

Garcia, S., Luengo, J., & Herrera, F. (2014). *Data preprocessing in data mining* (2015a ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-10247-4

Gawdiya, S., Kumar, D., Ahmed, B., & Sharma, R. K. (2024). *Field scale crop yield prediction using ensemble machine learning techniques. Smart Agricultural Technology, 9,* 100543. https://doi.org/10.1016/j.atech.2024.100543

Hodge, V., & Austin, J. (2004). *A survey of outlier detection methodologies. Artificial Intelligence Review, 22*(2), 85–126. https://www.researchgate.net/publication/220638052_A_Survey_of_Outlier_Detection_Methodologies

Hsu, C., Chang, C., & Lin, C. (2009). *A practical guide to support vector classification.* National Taiwan University. https://eecs.csuohio.edu/~sschung/DSA460/SVM_guide.pdf

Johnson, J., & Khoshgoftaar, T. (2019). *Survey on deep learning with class imbalance. Journal of Big Data, 6*(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* Springer. https://doi.org/10.1007/978-1-4614-6849-3

Meshram, V., Patil, K., & Hanchate, D. (2021). *Machine learning in agriculture domain: A state-of-art survey. Artificial Intelligence in Life Sciences, 1,* 100010. https://doi.org/10.1016/j.ailsci.2021.100010

Niether, W., Jacobi, J., Blaser, W. J., Andres, C., & Armengot, L. (2020). *Cocoa agroforestry systems versus monocultures: A multidimensional meta-analysis. Environmental Research Letters, 15,* 104085. https://doi.org/10.1088/1748-9326/abb053

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Vanderplas, J. (2011). *Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12,* 2825–2830. https://dl.acm.org/doi/pdf/10.5555/1953048.2078195

Saavedra, F., Peña, E. J., Schneider, M., & Naoki, K. (2020). *Effects of environmental variables and foliar traits on the transpiration rate of cocoa (Theobroma cacao L.) under different cultivation systems. Agroforestry Systems, 94,* 2021–2031. https://doi.org/10.1007/s10457-020-00522-5

Sohrab, F., Raitoharju, J., Iosifidis, A., & Gabbouj, M. (2021). *Multimodal subspace support vector data description. Pattern Recognition, 110,* 107648. https://doi.org/10.1016/j.patcog.2020.107648

Sperandei, S. (2014). *Understanding logistic regression analysis. Biochemia Medica, 24*(1), 12–18. https://doi.org/10.11613/BM.2014.003

Talero-Sarmiento, L., Roa-Prada, S., Caicedo-Chacón, L., & Gavanzo-Cárdenas, O. (2025). *A data-driven approach to improve cocoa crop establishment in Colombia: Insights and agricultural practice recommendations from an ensemble machine learning model. AgriEngineering, 7*(6). https://doi.org/10.3390/agriengineering7010006

Wilson, D., & Martinez, T. (2000). *Reduction techniques for instance-based learning algorithms. Machine Learning, 38*(3), 257–286. https://doi.org/10.1023/a:1007626913721