TRABAJO FINAL

UNIVERSIDAD EAFIT - AGRICULTURA PREDICTIVA

MARÍA FERNANDA MORENO DE LA ESPRIELLA
JHON SAHIAN ALVAREZ PANTOJA

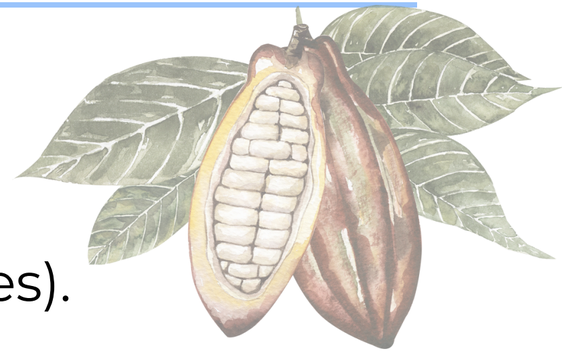
01

CLASIFICACIÓN DE LA **APTITUD** PARA EL ESTABLECIMIENTO DEL **CULTIVO DE CACAO** EN COLOMBIA A PARTIR DE VARIABLES CLIMÁTICAS USANDO MODELOS DE **MACHINE LEARNING.**

Evaluar la **aptitud** del terreno para el **cultivo de cacao en Colombia** a partir de **variables climáticas** mediante la aplicación y comparación de diferentes modelos de **machine learning**, con el propósito de determinar cuáles ofrecen el mejor desempeño para la clasificación de aptitud y apoyar la **planificación agrícola del cultivo**.

DATASET

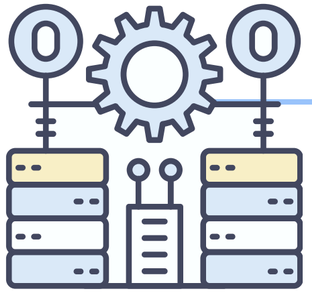
Obtenido de: Kaggle - hallbartfinaldataset



- NASA POWER, resolución espacial de 0.5° × 0.5° (análisis regionales).
- Aptitud - Upra (Unidad de Planificación Rural Agropecuaria)
- Radiación solar incidente, precipitación total corregida, radiación fotosintéticamente activa, humedad relativa, velocidad del viento y temperaturas máxima y mínima .

- Promedio (average) y la desviación estándar (std).
- Promedio refleja las condiciones medias del clima en cada zona
- Desviación estándar muestra qué tan variables son esas condiciones. (Condiciones estables ?)

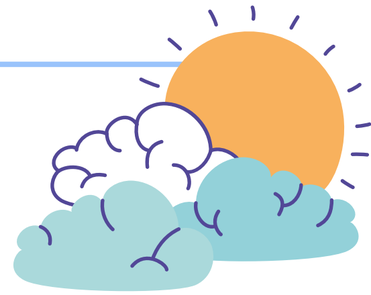
Talero-Sarmiento, L., Roa-Prada, S., Caicedo-Chacón, L., & Gavanzo-Cárdenas, O. (2025). A data-driven approach to improve cocoa crop establishment in Colombia: Insights and agricultural practice recommendations from an ensemble machine learning model. AgriEngineering, 7(6). https://doi.org/10.3390/agriengineering7010006



PROCESAMIENTO DE DATOS

- Dataset (2019, 2022 y 2023), con 57,659 registros por año.
- Dataset de 288,291 registros, se eliminó la variable “año”
- Se corrigieron inconsistencias en las coordenadas y se conservaron outliers.

- Variables se estandarizaron con StandardScaler (media 0, desviación 1)
- Variable objetivo se codificó con One-Hot Encoding
- Se Eliminaron predictores con alta correlación (> 0.8) o muy baja variabilidad (< 1e−6).
- Clases mostraron diferencia entre categorías (92,426 alta, 131,519 media y 64,345 baja)



Búsqueda de hiperparámetros

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.

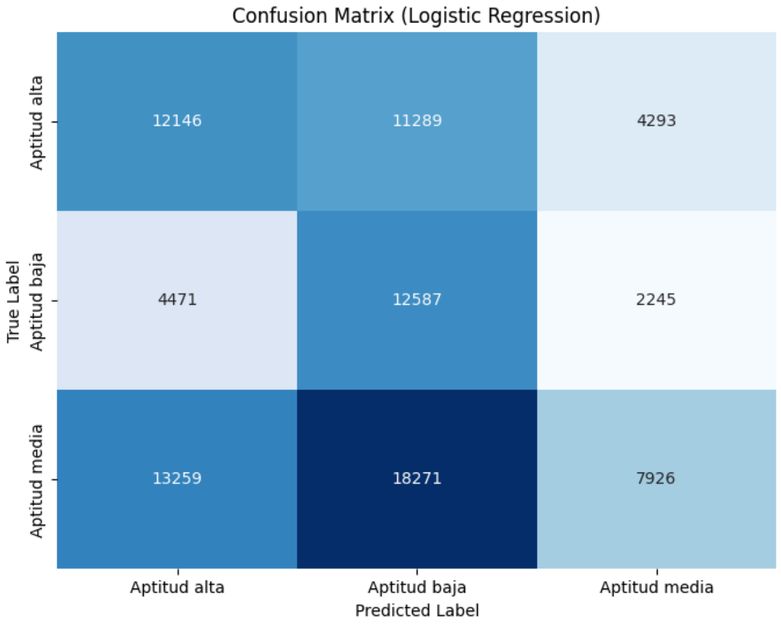
class weight = balanced.

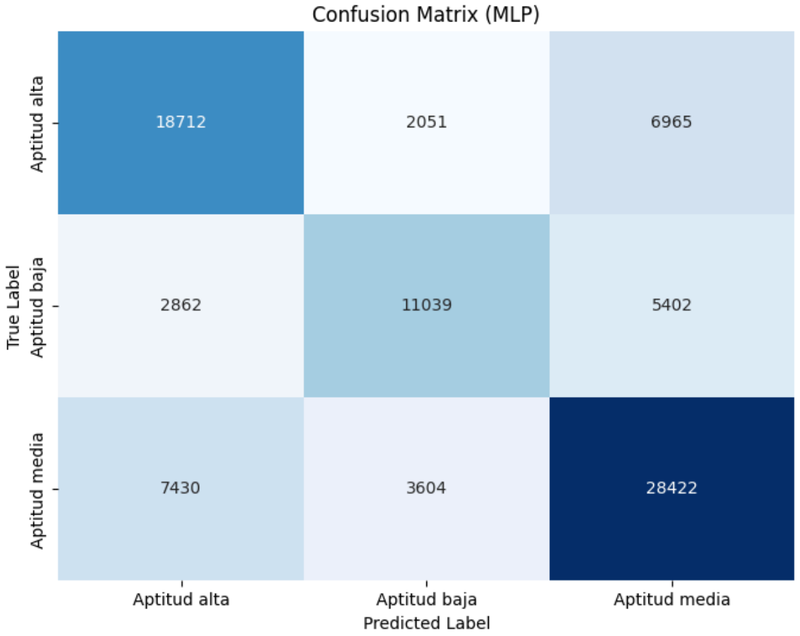
Modelo	Parámetros en Grid Search	Parámetros utilizados
Regresión Logística	C =[0.01, 0.1, 1, 10, 100] penalty = [L1, L2]	C= 1, Penalty = L2
Random Forest	n_estimators= [50, 100, 150] max_depth = [10, 15, 20] criterion = [gini]	Max_depth = 20, n_estimators = 50
Red Neuronal	Hidden_layer_sizes = [(50,), (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,

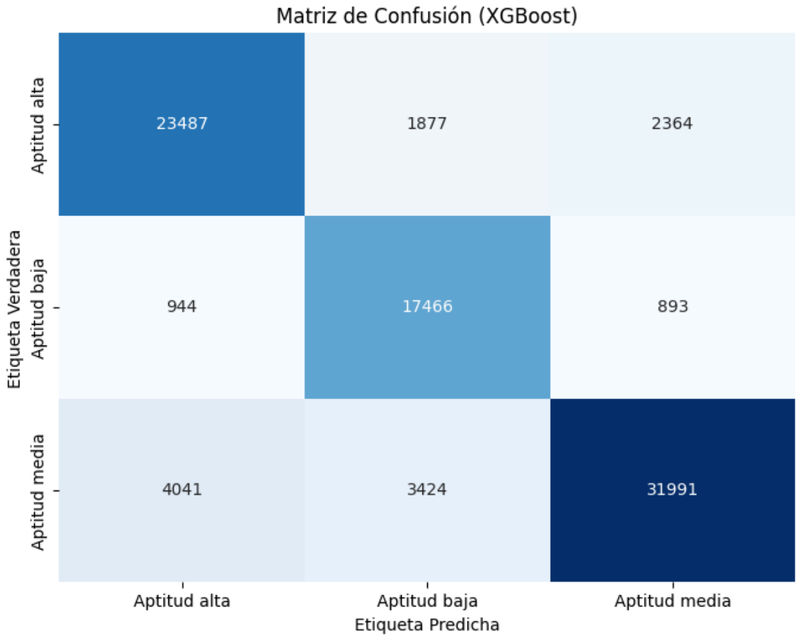
Tiempos:
RL =20 mins
RF = 30 mins
ANN = 90 mins
XGBoost= 10mins
SVM - N/A

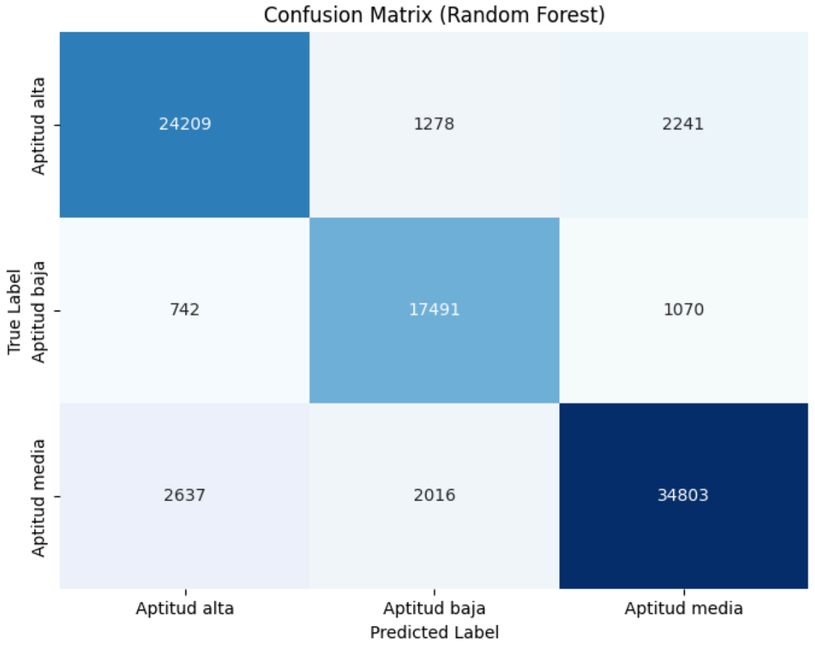
Talero-Sarmiento, L., Roa-Prada, S., Caicedo-Chacón, L., & Gavanzo-Cárdenas, O. (2025). A data-driven approach to improve cocoa crop establishment in Colombia: Insights and agricultural practice recommendations from an ensemble machine learning model. AgriEngineering, 7(6). https://doi.org/10.3390/agriengineering7010006

RESULTADOS


Confusion Matrix (Logistic Regression)
Aptitud alta
Aptitud baja
Aptitud media
True Label
12146
11289
4293
4471
12587
2245
13259
18271
7926
Aptitud alta
Aptitud baja
Aptitud media
Predicted Label


Confusion Matrix (MLP)
Aptitud alta
Aptitud baja
Aptitud media
True Label
18712
2051
6965
2862
11039
5402
7430
3604
28422
Aptitud alta
Aptitud baja
Aptitud media
Predicted Label

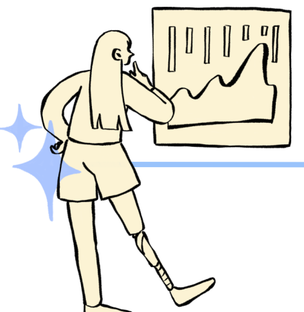

Matriz de Confusión (XGBoost)
Aptitud alta
Aptitud baja
Aptitud media
Etiqueta Verdadera
23487
1877
2364
944
17466
893
4041
3424
31991
Aptitud alta
Aptitud baja
Aptitud media
Etiqueta Predicha


Confusion Matrix (Random Forest)
Aptitud alta
Aptitud baja
Aptitud media
True Label
24209
1278
2241
742
17491
1070
2637
2016
34803
Aptitud alta
Aptitud baja
Aptitud media
Predicted Label

RESULTADOS

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.

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



DISCUSIÓN

Regresión Logística

Naturaleza líneal

Infrerepresentar interacción entre variables ambientales no líneales

(Talero-Sarmiento et al., 2025)

Random Forest

Relaciones no lineales

Interacción

Integrar información

(Talero-Sarmiento et al., 2025)

Ned Neuronal

Una sola capa oculta

Redes con mayor profundidad

Relaciones no líneales

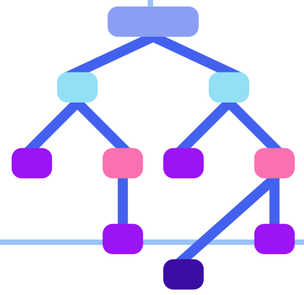
(Talero-Sarmiento et al., 2025)

XGBoost

Mejorar progresivamente predicciones

Boosting eficaces en contextos agrícolas

(Gawdiya et al., 2024)



CONCLUSIONES

MODELOS DE TIPO ENSEMBLE SON 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)

LA AUSENCIA DE VARIABLES EDÁFICAS - PRECISIÓN FUE LIGERAMENTE INFERIOR. DICHAS VARIABLES, AL REFLEJAR EL PERFIL DEL SUELO, SE CONSIDERAN DETERMINANTES PARA EL DESARROLLO DEL CACAO (ACHEAMPONG ET AL., 2019; NIETHER ET AL., 2020).

LA **REGRESIÓN LOGÍSTICA** PERMITIÓ ESTABLECER UNA LÍNEA BASE INTERPRETATIVA Y EVIDENCIAR LA **NATURALEZA NO LINEAL DEL FENÓMENO**. LA **RED NEURONA**L, **POTENCIAL DE MEJORA** CON ARQUITECTURAS MÁS PROFUNDAS .

02

EVALUACIÓN DE LA **SENSIBILIDAD** DE MODELOS DE CLASIFICACIÓN ANTE VARIACIONES EN LA ESTRUCTURA DE LOS **DATOS** MEDIANTE EXPERIMENTOS **SINTÉTICOS.**

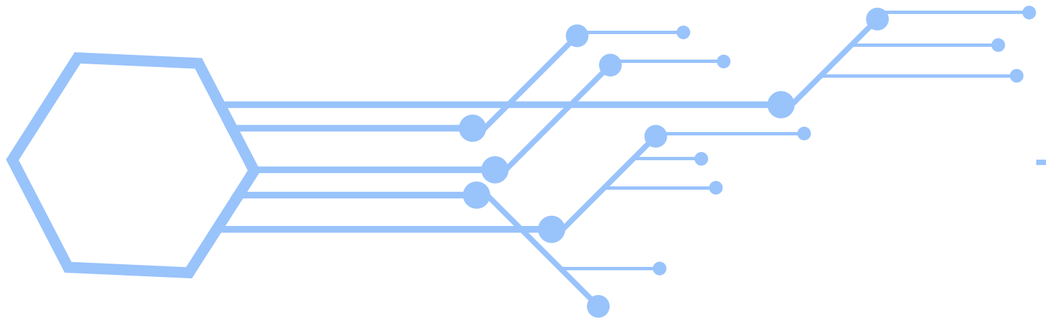
Análisis de datos sintéticos con modelos de Machine Learning

Objetivo: analizar la sensibilidad de los modelos de clasificación (SVM, KNN y RF) frente a variaciones en la estructura de los datos.

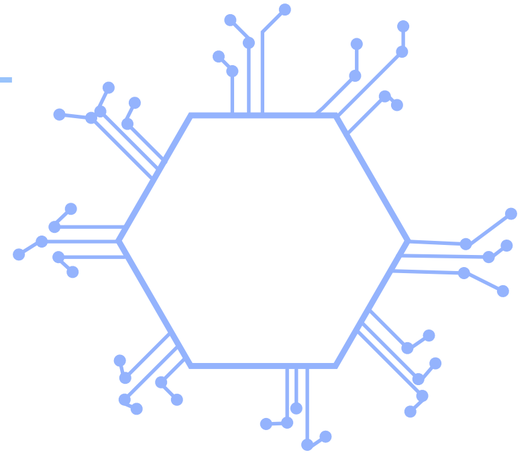
Se implementaron tres experimentos para observar el efecto de:

- Escala de las variables
- Presencia de valores atípicos (outliers)
- Desbalance de clases

Los datos fueron generados artificialmente mediante distribuciones normales multivariadas, con cuatro clases.



Experimento 1. Transformación de Datos



Objetivo: evaluar el efecto de la estandarización (StandardScaler).

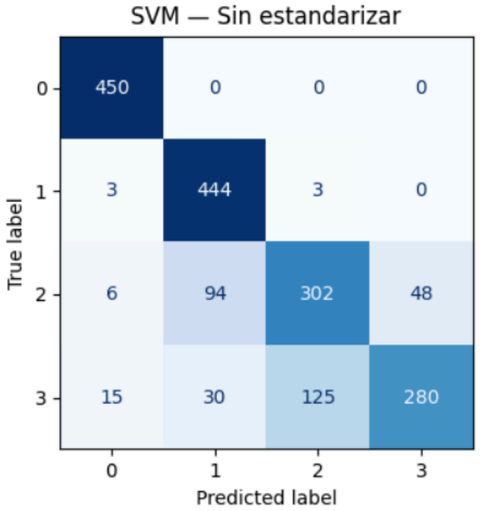
Hallazgos principales:

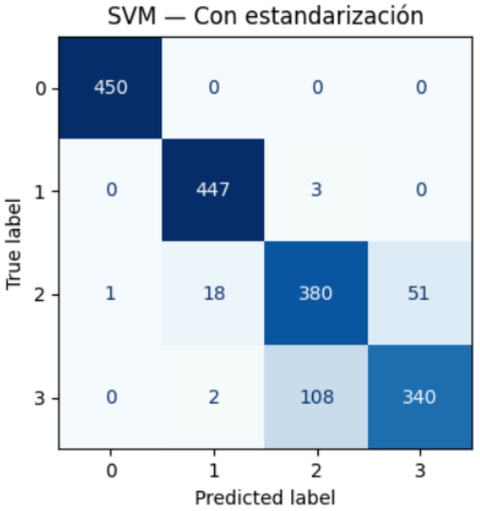
- SVM y KNN mejoran al estandarizar (F1).
- Random Forest se mantiene casi igual.
- Los modelos basados en distancia son sensibles a la escala de las variables (Hsu et al., 2016).
- Los modelos de árboles no requieren normalización (Pedregosa et al., 2011; Kuhn & Johnson, 2013).

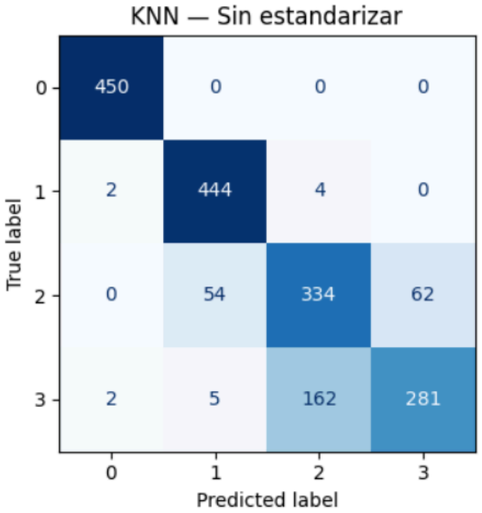
Conclusión:

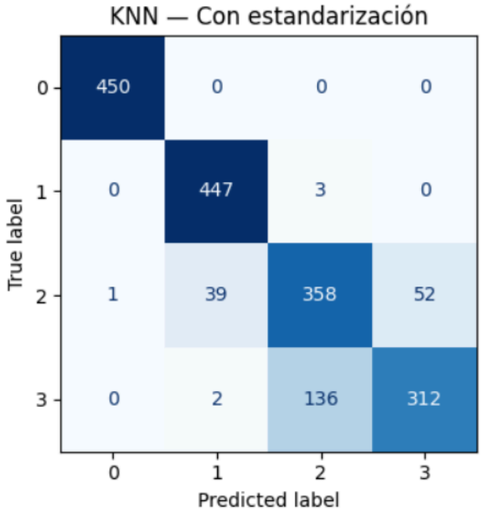
La estandarización mejora la precisión en modelos sensibles a magnitudes, pero no afecta a los basados en árboles.

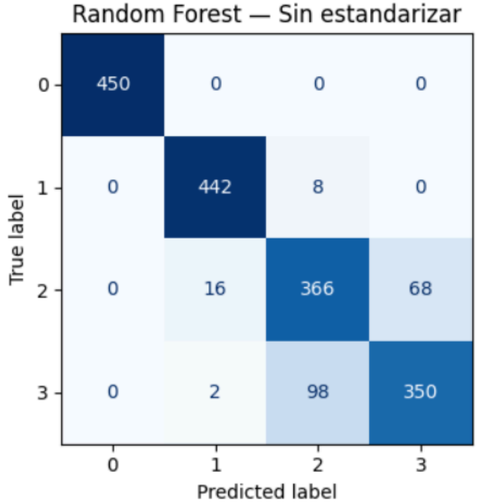
TRANSFORMACIÓN DE DATOS

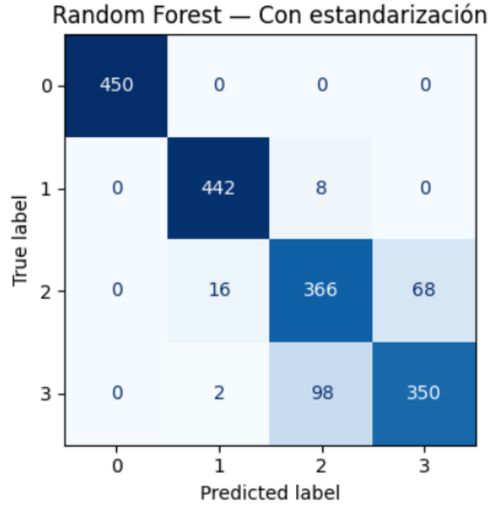




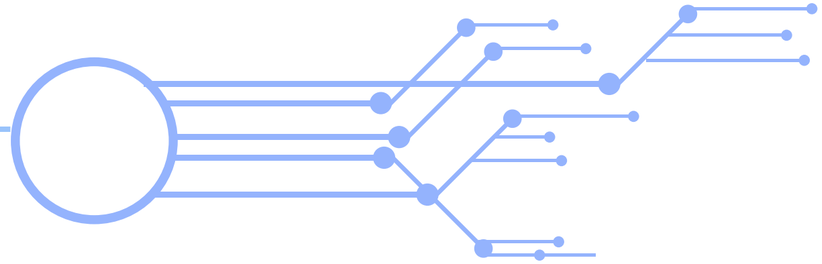








Experimento 2 – Efecto de outliers



Objetivo: analizar el impacto de los valores atípicos en el entrenamiento.

Condiciones:

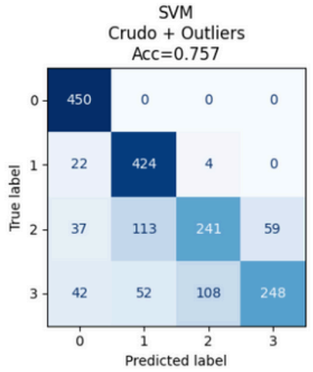
1. Datos crudos con outliers
2. Datos estandarizados limpios
3. Datos estandarizados contaminados

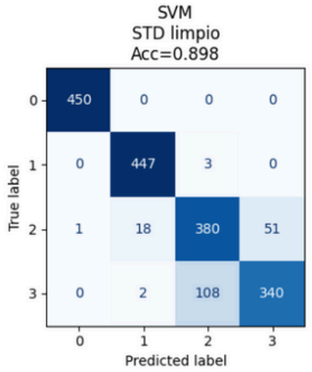
Resultados:

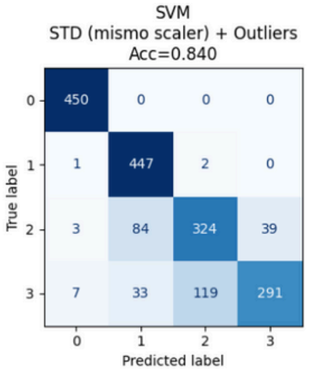
- SVM: cae la exactitud (0.757 → 0.898) por alta sensibilidad a outliers (Hodge & Austin, 2004).
- KNN: variaciones leves, pero también afectado por ruido local (Wilson & Martínez, 2000).
- RF: desempeño estable (~0.88), robusto ante ruido (Breiman, 2001).

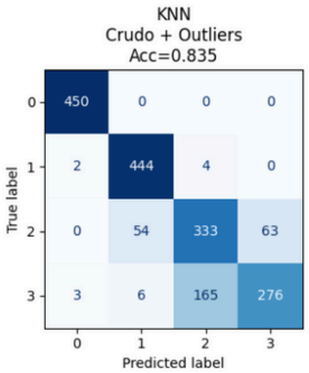
Conclusión:
SVM y KNN se ven afectados por datos atípicos, mientras que Random Forest mantiene estabilidad.

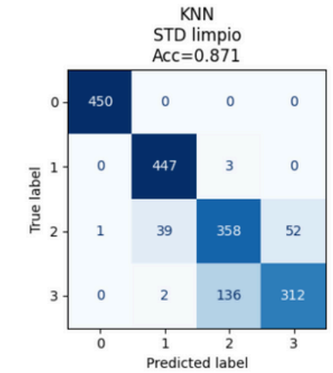
EXPERIMENTO 2 – EFECTO DE OUTLIERS

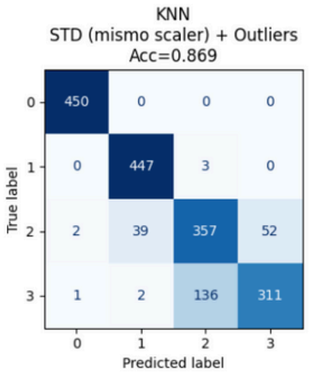


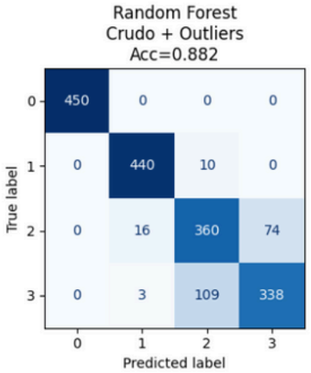


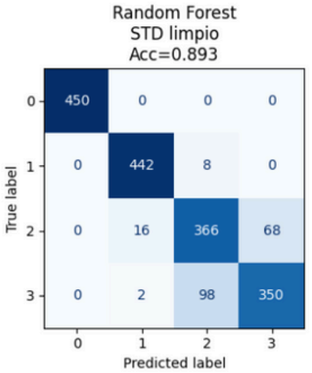


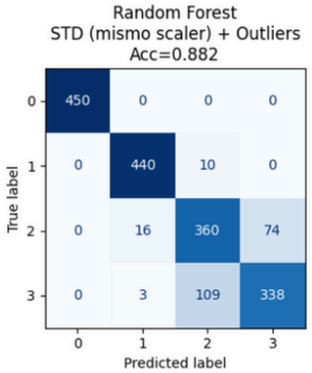




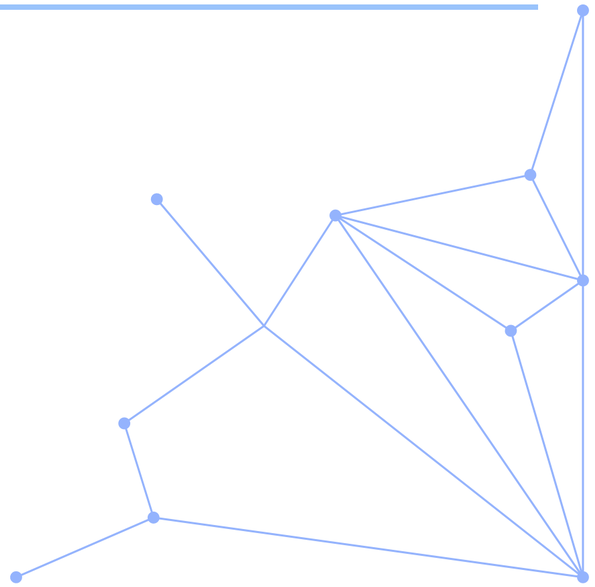








Experimento 3 – Desbalance de clases



Objetivo: evaluar cómo el desbalance influye en el rendimiento del modelo.
Escenarios:

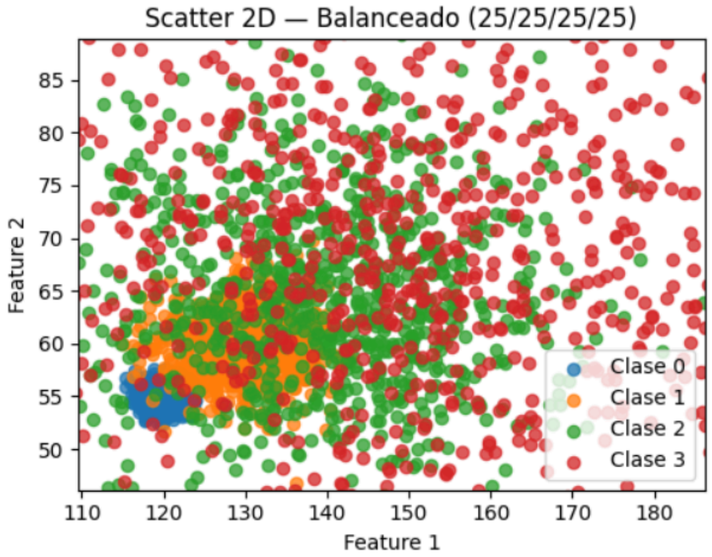
- Balanceado (25/25/25/25)
- Desbalance medio
- Desbalance extremo

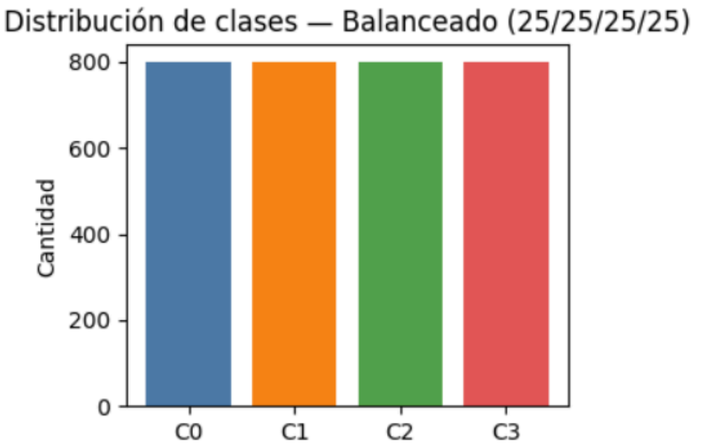
Resultados:

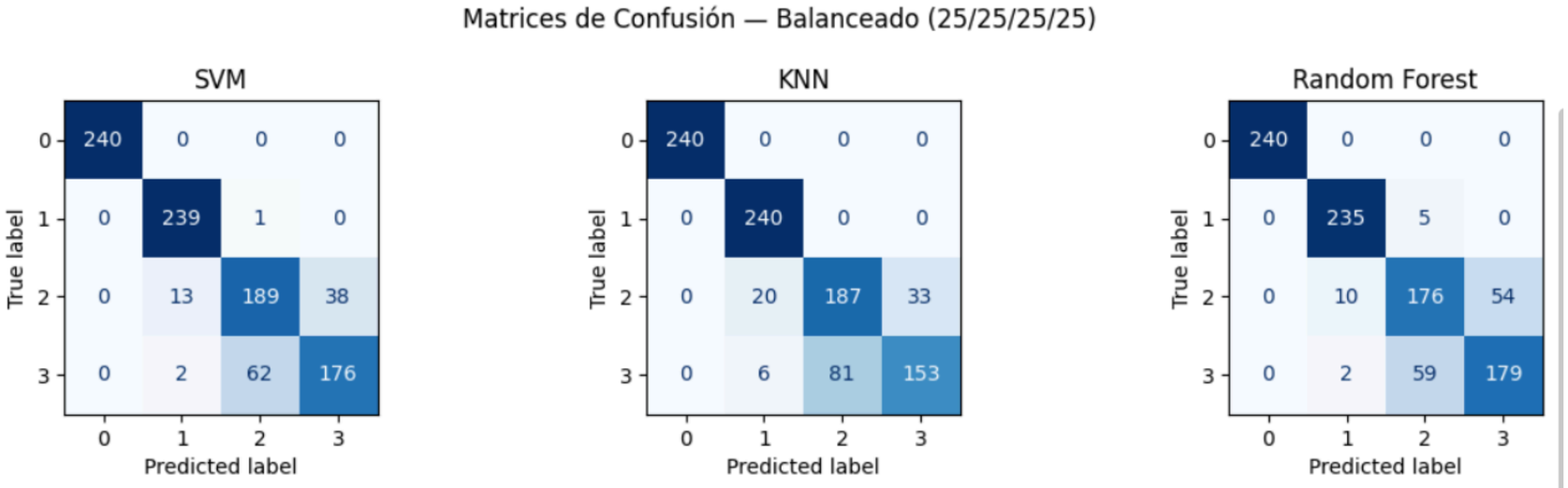
- Accuracy ↑ (0.86 → 0.97)
- F1 macro ↓, mostrando favorecimiento hacia clases mayoritarias (He & Garcia, 2009).
- F1 por clase:
  - Clase 1 (más datos) mantiene F1≈1.0
  - Clases 2–3 disminuyen progresivamente.
- RF se mantiene más estable (F1≈0.84), mientras SVM y KNN pierden equilibrio (Chen et al., 2004).

Conclusión:
El desbalance genera falsas mejoras en accuracy. Es necesario usar F1 macro y análisis por clase.

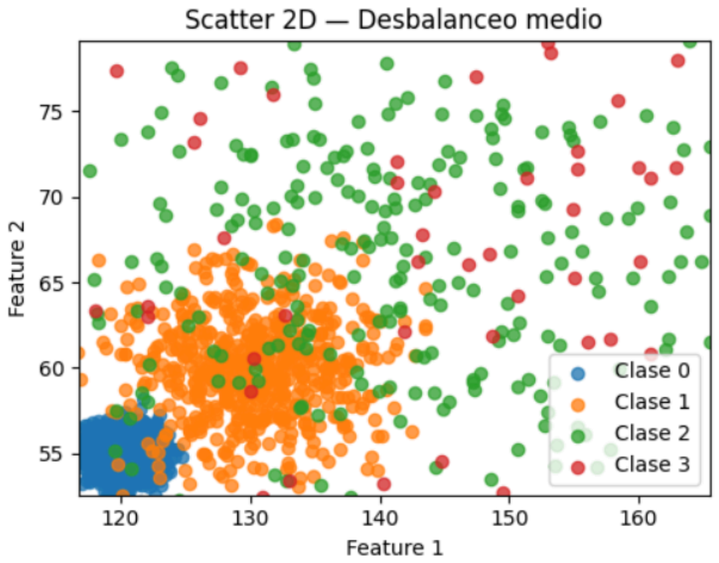
DESBALANCE DE CLASES

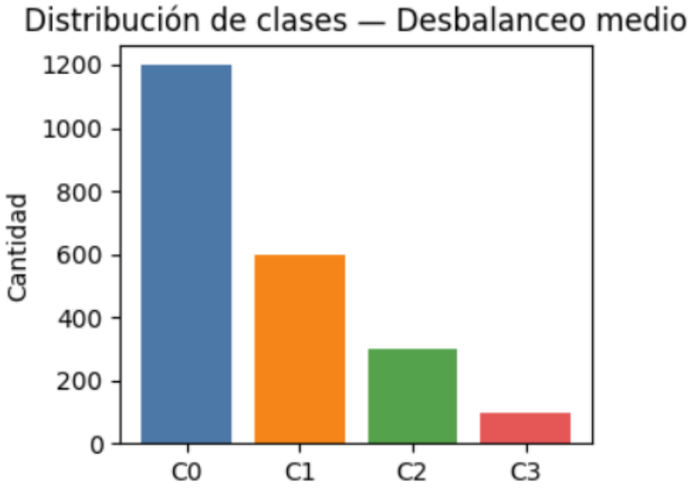




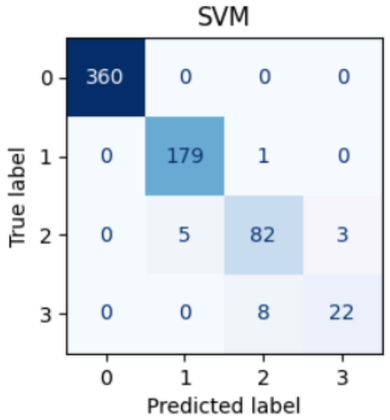


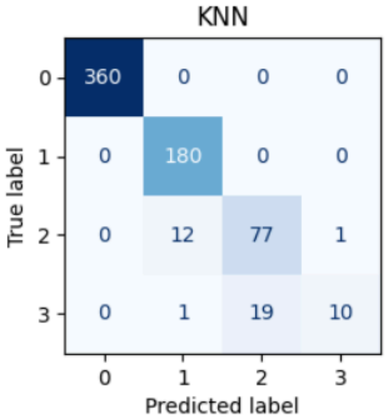
DESBALANCE DE CLASES

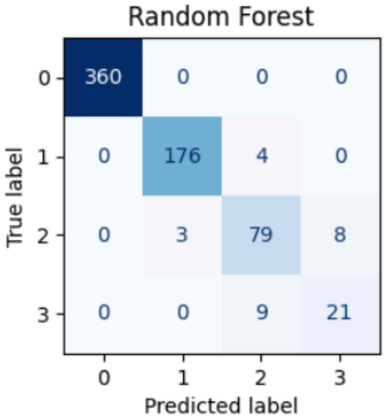




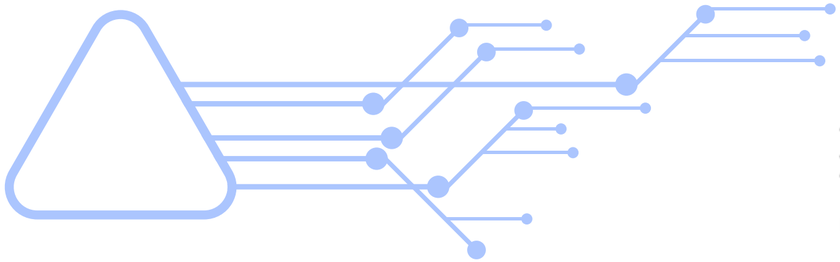
Matrices de Confusión — Desbalanceo medio

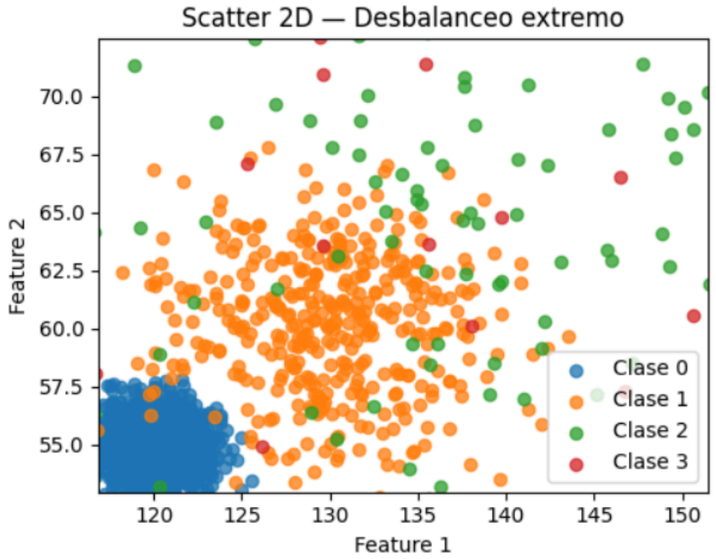


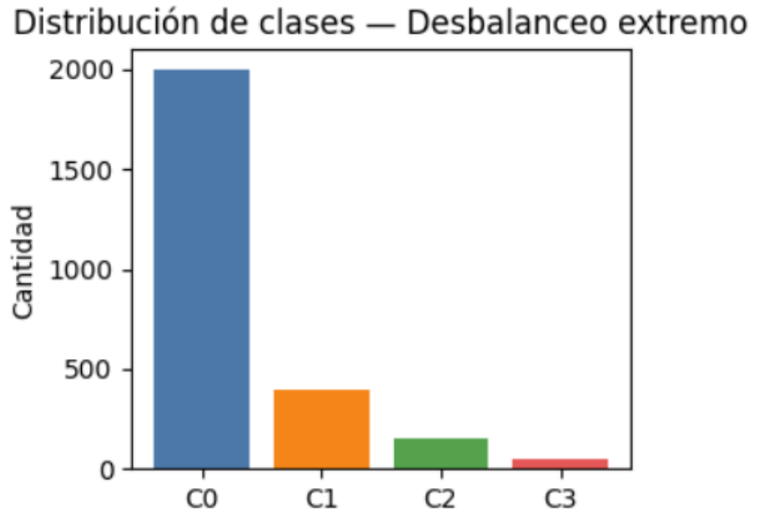




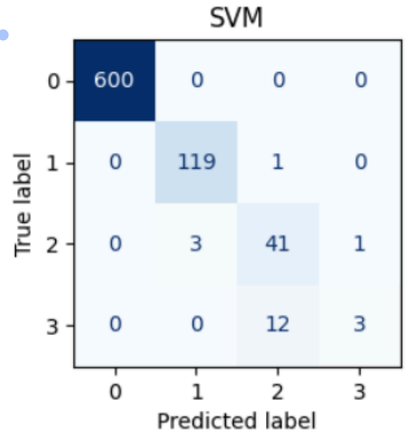
DESBALANCE DE CLASES

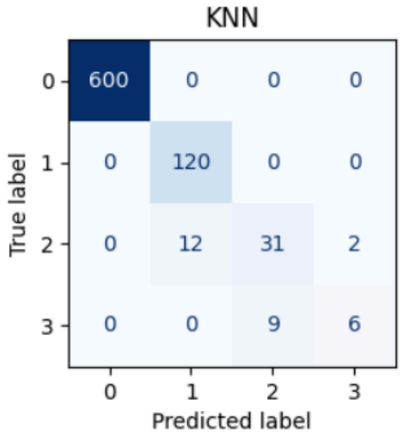


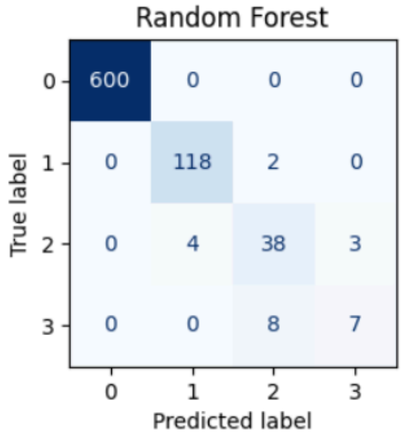




Matrices de Confusión — Desbalanceo extremo







DESBALANCE DE CLASES

Accuracy	F1_macro	Precision_macro	Recall_macro
866	864	866	866
976	807	865	791
963	877	914	865

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

Escenario	Modelo	Accuracy	F1_macro	F1_clase_0	F1_clase_1	F1_clase_2	F1_clase_3
Balanceado	SVM	879	878	10	968	768	775
Balanceado	KNN	854	851	10	949	736	718
Balanceado	RF	865	864	10	965	733	757
Desbalanceo medio	SVM	974	922	10	984	906	8
Desbalanceo medio	KNN	95	82	10	965	828	488
Desbalanceo medio	RaF	964	89	10	981	868	712
Desbalanceo extremo	SVM	978	782	10	983	828	316
Desbalanceo extremo	KNN	971	801	10	952	729	522
Desbalanceo extremo	RF	978	838	10	975	817	56

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.

CONCLUSIONES

- SVM Y KNN: SENSIBLES A ESCALA, RUIDO Y DESBALANCE.
- RANDOM FOREST: ESTABLE Y ROBUSTO ANTE VARIACIONES.
- EL F1 MACRO Y F1 POR CLASE PERMITEN EVALUAR EQUIDAD ENTRE CATEGORÍAS.
- LOS DATOS SINTÉTICOS DEMOSTRARON SER UNA HERRAMIENTA ÚTIL PARA ESTUDIAR LA SENSIBILIDAD ESTRUCTURAL DE LOS MODELOS SIN DEPENDER DE BASES REALES.

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
- 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

GRACIAS