Aplicaciones de la agricultura predictiva: comparación de modelos y procesos de Machine Learning.

Elisa Rendón Cadavid y María Camila Álvarez Bedoya.

Resumen

La agricultura predictiva combina datos, modelado estadístico y aprendizaje automático para anticipar eventos y decisiones en los sistemas agrícolas. Este trabajo evalúa experimentalmente distintos modelos clásicos de aprendizaje supervisado mediante dos aproximaciones complementarias: (1) clasificación de la aplicación de pesticidas en campos agrícolas reales, usando variables estructurales y ambientales, y (2) regresión sobre datos sintéticos generados controladamente para analizar el impacto de distintos procesos de preprocesamiento. Todos los experimentos se desarrollaron en Google Colab, priorizando la reproducibilidad y el análisis comparativo.

Parte 1. Clasificación de Campos Orgánicos en Cultivos Agrícolas.

Este estudio tuvo como propósito desarrollar un modelo de aprendizaje automático capaz de clasificar si un campo agrícola es orgánico o no, a partir de variables relacionadas con el uso de pesticidas, las condiciones del suelo y las características taxonómicas del cultivo. A partir de un conjunto de datos con 99,533 registros y 73 variables (Larsen, 2021), se realizó una limpieza, selección y reducción de características, seguida por el entrenamiento de cuatro modelos supervisados: Regresión Logística, Árbol de Decisión, Random Forest y SVM. Posteriormente, se aplicó una búsqueda de hiperparámetros (GridSearchCV) y un método de ensamble (Voting Classifier) para optimizar la precisión y robustez del sistema. El modelo Random Forest obtuvo el mejor desempeño, con una precisión del 95.6% y un F1-score de 0.95, demostrando ser el más adecuado para este tipo de clasificación.

Objetivo

El objetivo principal fue construir y evaluar modelos de clasificación que permitan predecir de manera precisa si un campo agrícola es orgánico o no, utilizando información sobre el uso de pesticidas, tipo de cultivo, calidad del suelo y superficie cultivada. El propósito final es identificar patrones que puedan apoyar la toma de decisiones sostenibles en la gestión agrícola.

Metodología

Preparación y exploración de datos:

Se trabajó con el conjunto de datos *"Identifying and characterizing pesticide use on 9,000 fields of organic agriculture"* (Larsen, 2021), con un total de 99,533 registros y 73 columnas. Estas variables incluyen información sobre uso de pesticidas, calidad del suelo, tamaño de las fincas, códigos de cultivos y clasificaciones taxonómicas. La mayoría de los datos son numéricos (enteros o tipo *float*), aunque algunas columnas son de tipo *object* (texto).
Se observó además una proporción significativa de valores nulos, especialmente en las variables relacionadas con el suelo, y un rango temporal comprendido entre los años 2013 y 2019. Durante la exploración, se eliminaron variables redundantes, identificadores (como *permitID* y *PermFam*) y columnas con más del 80% de valores faltantes (*soilX_hydro*, *soilX_temp*). Se descartaron también transformaciones logarítmicas y variantes correlacionadas, conservando solo las más representativas.

Selección de variables:

La variable objetivo (*Y*) fue *pur_cdfa_org* (1 = orgánico, 0 = no orgánico).

Las variables predictoras (*X*) finales fueron: *KgPestAIHa100, year, soil_quality, soilX_chem, comm_code2, CropFarmerOverlap, Area (ha), genus_code y family_code.*

Se seleccionó *KgPestAIHa100* como medida representativa del uso de pesticidas, al reflejar directamente la cantidad aplicada por unidad de área. *Area (ha)* se obtuvo al revertir las transformaciones logarítmicas presentes en el dataset, y las variables *genus* y *family* fueron codificadas numéricamente mediante *LabelEncoder* para preservar su valor informativo.

Transformación y reducción de dimensionalidad:

Las variables numéricas se estandarizaron con *StandardScaler*. Luego, se aplicó *Análisis de Componentes Principales (PCA)*, manteniendo el 95% de la varianza, lo que redujo las características a ocho componentes principales.

Modelos de clasificación:

Se entrenaron cuatro modelos de aprendizaje supervisado:

- *Regresión Logística*: como modelo base lineal.
- *Árbol de Decisión*: por su interpretabilidad y capacidad para capturar relaciones no lineales.
- *Random Forest*: para mejorar la generalización mediante el promedio de múltiples árboles.
- *SVM (RBF)*: para modelar fronteras complejas y no lineales entre clases.

La división de los datos fue del 80% para entrenamiento y 20% para prueba.

Ajuste de hiperparámetros:
Mediante *GridSearchCV* con validación cruzada, se optimizaron los parámetros de cada modelo:

- Regresión Logística: *C=10, solver='liblinear'*.
- Árbol de Decisión: *max_depth=10, min_samples_split=5*.
- Random Forest: *n_estimators=200, max_depth=None*.
- SVM: *C=1, gamma='auto'*.

Este proceso permitió ajustar la complejidad y regularización de los modelos, reduciendo el riesgo de sobreajuste y mejorando la capacidad de generalización.

Método de ensamble:

Finalmente, se aplicó un *Voting Classifier* con votación “hard”, combinando los cuatro modelos para aprovechar las fortalezas de cada uno y mejorar la estabilidad general del sistema.

Resultados

Tabla 1. Resultados obtenidos tras el ajuste de hiperparámetros en los modelos evaluados.

Modelo	**Accuracy**	**F1-score**
Random Forest	**0.956**	**0.955**
Árbol de Decisión	0.944	0.944
SVM (RBF)	0.94	0.939
Regresión Logística	0.927	0.923

El método de ensamble (*Voting Classifier*) alcanzó una precisión del **94.2%** y un F1 ponderado de **0.94**, confirmando la coherencia de los modelos y la complementariedad de sus predicciones.

Discusión

El modelo Random Forest obtuvo el mejor rendimiento debido a su capacidad para manejar relaciones no lineales, variables correlacionadas y ruido en los datos. Al promediar los resultados de múltiples árboles, logró un equilibrio entre sesgo y varianza, lo que se tradujo en mayor precisión y estabilidad.

El Árbol de Decisión tuvo buen desempeño, aunque ligeramente inferior, evidenciando que un único árbol puede ajustarse bien a los datos.

El SVM, aunque eficiente en separación de clases no lineales, fue más sensible al desbalance de clases.

La Regresión Logística fue la menos precisa, lo que confirma que las relaciones entre variables predictoras y la condición orgánica no son estrictamente lineales.

El ajuste de hiperparámetros mejoró de forma notable la precisión en todos los modelos, mostrando la importancia de controlar la regularización y la profundidad de los árboles.

El método de ensamble permitió validar la estabilidad de los resultados al combinar enfoques con diferentes sesgos (lineales, basados en árboles y de márgenes). Aunque no superó al Random Forest individual, sí confirmó la consistencia del sistema y redujo la varianza global.

Conclusiones

El modelo Random Forest se consolidó como el más eficiente para la clasificación de campos agrícolas orgánicos, alcanzando una precisión del 95.6%. Su desempeño superior se debe a su capacidad para manejar relaciones no lineales, ruido y variables correlacionadas, lo que permitió capturar mejor la complejidad del fenómeno agrícola analizado. El ajuste de hiperparámetros mediante *GridSearchCV* mejoró notablemente el rendimiento de todos los algoritmos, evidenciando la importancia de optimizar la profundidad, regularización y número de estimadores. Además, el método de ensamble (*Voting Classifier*) confirmó la estabilidad y coherencia de los resultados al combinar diferentes enfoques de aprendizaje supervisado.

A futuro, se propone ampliar el análisis incorporando variables ambientales (como precipitación, temperatura y tipo de suelo) y datos satelitales o de drones que representen condiciones reales del terreno. También sería valioso aplicar técnicas avanzadas de aprendizaje profundo (*Deep Learning*) y modelos de interpretación para entender con mayor detalle la influencia de cada variable en la clasificación. Estos pasos fortalecerían la capacidad predictiva del modelo y su utilidad como herramienta de apoyo en la toma de decisiones para la agricultura sostenible.

Parte 2. Uso de datos sintéticos para comparación de procesos.

La generación de datos sintéticos constituye una estrategia clave en experimentos de aprendizaje automático cuando se busca evaluar modelos bajo condiciones controladas, especialmente en contextos agrícolas donde la obtención de grandes volúmenes de datos reales es costosa o limitada. En este trabajo se creó un conjunto de datos artificial con características numéricas simuladas a partir de distribuciones continuas, representando variables agroclimáticas y edáficas.

Objetivo.

Predecir una variable continúa asociada al rendimiento. Por medio de la exploración sistemática del impacto de distintos tipos de preprocesamiento y modelado sobre métricas de desempeño.

Metodología.

Se compararon tres modelos de aprendizaje supervisado: LassoCV, K-Nearest Neighbors (KNN) y una red neuronal multicapa (MLP), aplicados bajo diversas combinaciones de transformaciones (StandardScaler, MinMaxScaler, MaxAbsScaler, Normalizer) y métodos de selección de características (sin selección, ANOVA y RFE). Para cada configuración se implementaron tres variantes metodológicas: entrenamiento base, validación cruzada y búsqueda de hiperparámetros mediante GridSearch. Las métricas de evaluación con los datos de prueba incluyeron el coeficiente de determinación (R^2), el error absoluto medio (MAE) y la raíz del error cuadrático medio (RMSE).

Resultados generales

Los resultados muestran una consistencia notable en las métricas, especialmente en los modelos LassoCV y MLP con la mayoría de las transformaciones, con valores de R^2 superiores a 0.98 y errores bajos (MAE entre 10 y 12). En contraste, el modelo KNN exhibió un desempeño considerablemente inferior (R^2 entre 0.67 y 0.76), reflejando su alta sensibilidad a la escala y distribución de los datos.

Las configuraciones que combinaron LassoCV con cualquier método de selección y escalado lineal (StandardScaler o MinMaxScaler) alcanzaron los mejores resultados. Esto se explica porque el Lasso utiliza regularización L1, que penaliza los coeficientes de las variables menos relevantes. Dado que los datos sintéticos se generaron de forma lineal y sin ruido estructural, este modelo fue capaz de capturar casi toda la varianza de la variable objetivo.

La similitud entre los valores obtenidos en las métricas (particularmente en las repeticiones de MAE, RMSE y R^2 para las diferentes configuraciones) se debe a la estructura altamente controlada del dataset sintético. Al no existir ruido aleatorio ni variaciones temporales, y dado que la separación entre entrenamiento y prueba se realizó sobre distribuciones idénticas, los modelos convergieron hacia soluciones equivalentes en cada réplica del experimento.

En la tabla de resultados se observó además un patrón de repetición en las métricas: para cada tipo de transformación y método de mejora de hiperparámetros, cuando los selectores fueron ANOVA y RFE, los valores de R^2, MAE y RMSE no cambiaron. Este comportamiento indica que, bajo datos generados de forma perfectamente lineal, los procesos de selección de características no tuvieron impacto real en la capacidad predictiva del modelo. Asimismo, no se observaron diferencias entre los

métodos base y cross-validation, lo que refuerza la idea de que la homogeneidad del dataset impidió la detección de mejoras o ajustes por validación cruzada.

Asimismo, el uso de cross-validation y GridSearch no produjo variaciones significativas respecto al entrenamiento base, ya que la uniformidad de los datos impidió que el modelo encontrara combinaciones de hiperparámetros que mejoraran sustancialmente el ajuste. En contextos con datos reales, estas técnicas suelen generar diferencias más notorias al mitigar el sobreajuste o ajustar la complejidad del modelo.

Desempeño de cada modelo

1. **LassoCV (Regresión lineal regularizada):**

   Presentó el mejor rendimiento global ($R^2 \approx 0.992$) debido a que los datos se generaron mediante relaciones lineales, condición bajo la cual Lasso alcanza una representación óptima. La regularización L1 eliminó coeficientes redundantes y evitó el sobreajuste, lo que explica la consistencia en todas las métricas y escaladores.

2. **K-Nearest Neighbors (KNN):**

   Mostró bajo desempeño relativo ($R^2 < 0.78$), reflejando su sensibilidad a la escala y a la dimensionalidad de los datos. Aunque las transformaciones (particularmente la normalización) mejoraron ligeramente los resultados, la naturaleza no paramétrica de KNN lo hace dependiente de la densidad local y sensible a distribuciones uniformes como las del dataset sintético.

3. **Multilayer Perceptron (MLP):**

   Alcanzó un buen desempeño ($R^2 \approx 0.99$), aunque ligeramente inferior al de LassoCV en escenarios lineales. Dado que el MLP aprende representaciones no lineales, su ventaja no se manifiesta plenamente cuando las relaciones entre variables son estrictamente lineales y los datos carecen de mucho ruido. No obstante, el modelo demostró estabilidad ante distintos métodos de preprocesamiento, confirmando su capacidad de generalización.

Conclusiones

Los resultados evidencian que el uso de datos sintéticos es una herramienta eficaz para evaluar el comportamiento de los modelos en condiciones ideales y reproducibles, permitiendo aislar el efecto de cada etapa del flujo de aprendizaje automático. En este caso, el LassoCV se consolidó como el modelo más eficiente y estable, confirmando que las relaciones lineales entre variables son capturadas con alta precisión. La homogeneidad de los datos explica la repetición de métricas y la mínima ganancia entre métodos de validación. Como futura experimentación, se propone introducir más ruido o estructuras no lineales para evaluar la robustez de los modelos frente a condiciones más realistas.

Referencias

Larsen, A., McComb, S., & Powers, C. (2021). Analysis Data for «Identifying and characterizing pesticide use on 9,000 fields of organic agriculture» [Conjunto de datos]. En *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.25349/d9q02t