UNIVERSIDAD DEL VALLE DE GUATEMALA

Proyecto 2

Andre Marroquin Tarot - 22266
Sergio Orellana - 221122
Rodrigo Mansilla - 22611
Carlos Valladares - 221164

Data Science

Link archivo de google:
https://docs.google.com/document/d/1J6Q1pFgZmEEzgL2px3BJNoFXPlDDDi4mqihAujE1PH0/edit?usp=sharing

Link repositorio:
https://github.com/mar22266/PY2-DS.git

Branch: Resultados Parciales y Visualizaciones Estáticas

https://github.com/mar22266/PY2-DS/tree/Resultados-Parciales-y-Visualizaciones-Est%C3%A1ticas

Objetivo
Desarrollar y comparar modelos de aprendizaje automático que estimen el porcentaje de daño EXTENT, 0–100 por cada combinación ID, DAMAGE del reto CGIAR "Eyes on the Ground", utilizando metadatos season, growth_stage, damage y rasgos derivados del nombre de archivo, con una validación que evite fuga de información entre observaciones del mismo ID.

Investigación de algoritmos
La literatura en detección y cuantificación de enfermedades en plantas muestra que los enfoques no lineales y de ensemble (Random Forest y métodos de gradiente) suelen rendir mejor en datos tabulares heterogéneos, mientras que los modelos lineales sirven como línea base interpretable. Además, cuando se dispone de imágenes, el transfer learning con CNN puede aportar mejoras al combinarse con metadatos. Para este trabajo se priorizó la capa tabular, con la opción de integrar visión por computadora en fases futuras.

Posibles algoritmos considerados

1. Modelos lineales regularizados (Ridge) como baseline interpretable.

2. Ensembles de árboles: Random Forest robusto a outliers y útil con muchas categóricas y métodos de gradiente HistGradientBoosting por su buen sesgo-varianza en tabular.

Modelos entrenados y justificación
Se entrenaron tres modelos principales: Ridge, Random Forest y HistGradientBoosting. La selección responde a disponer de un baseline lineal, probar un ensemble robusto (RF) y un booster eficiente (HGB). Todo se implementó con pipelines reproducibles preprocesamiento + modelo y búsqueda de hiperparámetros con GridSearchCV.

Evaluación y resultados
La evaluación utilizó GroupKFold por ID para evitar fuga entre filas de la misma imagen. La

métrica principal fue MAE en validación cruzada (CV). El ranking obtenido fue: Random Forest aprox 3.00 mejor, HistGradientBoosting aprox 3.05, Ridge aprox 4.52 (menor es mejor). Las visualizaciones estáticas incluyen, barras comparando MAE de CV por modelo, dispersión y_true vs y_pred con línea ideal y, para modelos de árbol, la importancia de características.

Conclusiones y selección final

Los modelos no lineales superan al baseline lineal, evidenciando interacciones entre damage, season, growth_stage y rasgos del filename. Random Forest fue seleccionado como modelo final por su mejor MAE de CV y robustez operativa. El modelo se reentrenó con todos los datos de entrenamiento usando los mejores hiperparámetros y se generó el archivo de predicciones para test.

Referencias

- Mohanty, S. P., Hughes, D. P., y Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419. https://doi.org/10.3389/fpls.2016.01419
- Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145, 311–318. https://doi.org/10.1016/j.compag.2018.01.009
- Shoaib, M., Shah, B., El-Sappagh, S., Ali, A., Ullah, A., Alenezi, F., Gechev, T., Hussain, T., y Ali, F. (2023). An advanced deep learning models-based plant disease detection: A review of recent research. Frontiers in Plant Science, 14, 1158933. https://doi.org/10.3389/fpls.2023.1158933
- Barbedo, J. G. A. (2018). Factors influencing the use of deep learning for plant disease recognition. Biosystems Engineering, 172, 84–91. https://doi.org/10.1016/j.biosystemseng.2018.05.013
- Hasan, R. I., Yusuf, S. M., y Alzubaidi, L. (2020). Review of the state of the art of deep learning for plant diseases. Intelligent Systems with Applications, 6, 200002. https://doi.org/10.1016/j.iswa.2020.200002
- CGIAR Eyes on the Ground (Descripción del reto/dataset). Zindi y documentación de Eyes on the Ground. https://zindi.africa/competitions/cgiar-eyes-on-the-ground-challenge