IDENTIFICACIÓN DE GRUPOS HETERÓTICOS EN MAÍZ MEDIANTE MODELOS PREDICTIVOS

HETEROTIC GROUP CLASSIFICATION OF MAIZE THROUGH SUPERVISED MACHINE LEARNING

Hisse, I. R.[1,2]; D'Andrea, K. E.[1,2]; Otegui, M. E.[1,2,3]

[1]Dep. de Producción Vegetal, FAUBA. [2]CONICET. [3]EEA Pergamino INTA. Av. San Martín 4453 (C1417DSE), Buenos Aires, Argentina.
hisse@agro.uba.ar

Introducción

En los programas de mejoramiento de obtención de híbridos, resulta clave organizar el germoplasma en **grupos heteróticos (GH)**(1), evaluándose posteriormente los materiales en cruzamientos de prueba con individuos del GH opuesto(2). En este sentido, será de gran utilidad el uso de modelos predictivos que permitan identificar a priori distintos GHs a partir de características fenotípicas fácilmente medibles en las líneas. No obstante, no existen reportes sobre modelos de *machine learning* usados en la clasificación de líneas en GHs a partir de su fenotipo.

El objetivo de este trabajo fue evaluar dos modelos predictivos supervisados (árbol de decisión y ensamble de árboles) para identificar GHs a partir de atributos de rendimiento y composición del grano en un conjunto de líneas endocriadas y en un rango amplio de ambientes (distintos años, niveles de nitrógeno, y oferta hídrica).

Materiales y Métodos

GERMOPLASMA. Seis líneas endocriadas de dos GHs: GH_1 compuesto por líneas flint LP2, LP561, LP611, LP662, ZN6; y GH_2 compuesto por B100 (dentada).

EXPERIMENTOS. Las líneas se evaluaron en la EEA INTA Pergamino, Argentina (33°56´ S, 60°34´ O) durante 7 años (2002-03, 2003-04, 2004-05, 2006-07, 2008-09, 2013-14, 2014-15), dos niveles de nitrógeno (no fertilizado; fertilizado con 200 ó 400 kg N ha^{-1}), dos niveles de riego en dos años (con o sin riego) siendo en total 18 ambientes .

CARACTERES. Se estimó a madurez fisiológica el rendimiento en grano por planta (RGP), número de granos por planta (NGP), peso de los granos (PG, cociente entre RGP y NGP), porcentaje de proteína en grano (PROT), rendimiento de proteína en grano (RG_PROT), y cociente entre NGP y PG (ratio NGP/PG). Además, se incluyeron las variables RGP, NGP, PG, y PROT seguidas de "_N" y "_R" (e.g. RGP_N), las cuales se computaron por la mediana de cada nivel de N y condición de riego, respectivamente. Se incluyeron las variables categóricas año, nivel de nitrógeno, y riego, dando un total de 17 variables evaluadas.

MODELOS PREDICTIVOS. Se utilizó un **árbol de decisión**(3) y un **ensamble de árboles**(4). El set de datos completo (n = 428) se dividió aleatoriamente en un *set* de entrenamiento (66% de la base) y excluyendo la variable objetivo a predecir (**GH**), y uno de testeo (34% restante de la base) para evaluar la capacidad predictiva de los modelos(5). Para evaluar la capacidad de predicción se computó la efectividad (*accuracy*) y el área bajo la curva (AUC) ROC (*receiver operating characteristic*). La efectividad y AUC oscilan entre 0 (predicción nula) y 1 (predicción perfecta).

Resultados

Tabla 1. Valores promedio y rango de ambos grupos heteróticos (GH_1, GH_2)

VARIABLE	GH_1	GH_2	Mín.-Máx.
NGP *(granos pl^{-1})*	246	**273**	15 - 536
PG *(mg $grano^{-1}$)*	190	**204**	82 - 280
RGP *(g pl^{-1})*	47	**56**	2,6 - 114
PROT *(%)*	**11,2**	9,9	6,2 - 15,2
RG_PROT *(g pl^{-1})*	5,2	**5,4**	0,25 - 13,4
Ratio NGP/PG	1,31	**1,37**	0,13 - 3,8
NGP_N *(granos pl^{-1})*	0,99	**1,09**	0,06 - 2,1
PG_N *(mg $grano^{-1}$)*	0,97	**1,04**	0,42 - 1,4
RGP_N *(g pl^{-1})*	0,97	**1,14**	0,04 - 2,3
PROT_N *(%)*	**1,02**	0,90	0,59 - 1,5
NGP_R *(granos pl^{-1})*	0,98	**1,08**	0,06 - 2,1
PG_R *(mg $grano^{-1}$)*	0,97	**1,05**	0,45 - 1,5
RGP_R *(g pl^{-1})*	0,99	**1,16**	0,03 - 2,3
PROT_R *(%)*	**1,01**	0,90	0,56 - 1,4

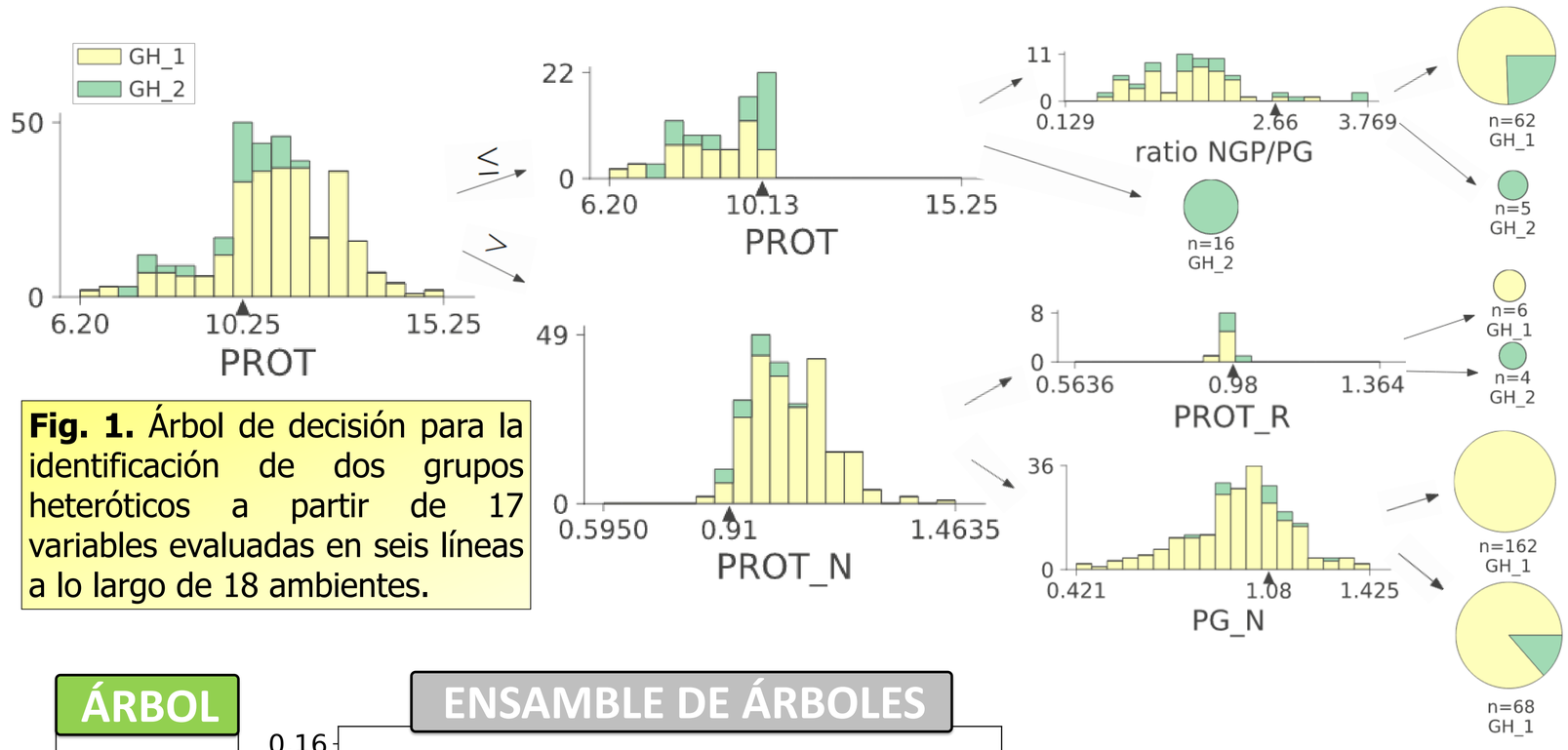

Fig. 1. Árbol de decisión para la identificación de dos grupos heteróticos a partir de 17 variables evaluadas en seis líneas a lo largo de 18 ambientes.

- El GH_2 (línea dentada) fue superior al GH_1 para RGP, NGP y PG, pero lo inverso (GH_1 > GH_2) se observó para el %PROT en grano (**Tabla 1**).
- La importancia relativa de los atributos evaluados fue mucho mas contrastante en el árbol que en el ensamble de árboles (**Fig. 2**).
- Los atributos de %PROT en grano fueron los mas importantes para clasificar a las líneas en los GHs (**Figs. 1 y 2**).
- Los componentes numéricos NGP y PG tanto absolutos como relativizados a la condición de riego y nitrógeno fueron más importantes que el RGP para clasificar a las líneas (**Fig. 2**).
- Ambos modelos tuvieron similar efectividad (0,85), pero el área bajo la curva ROC (AUC) del del ensamble fue mayor que la del árbol de decisión (**Fig. 3**).

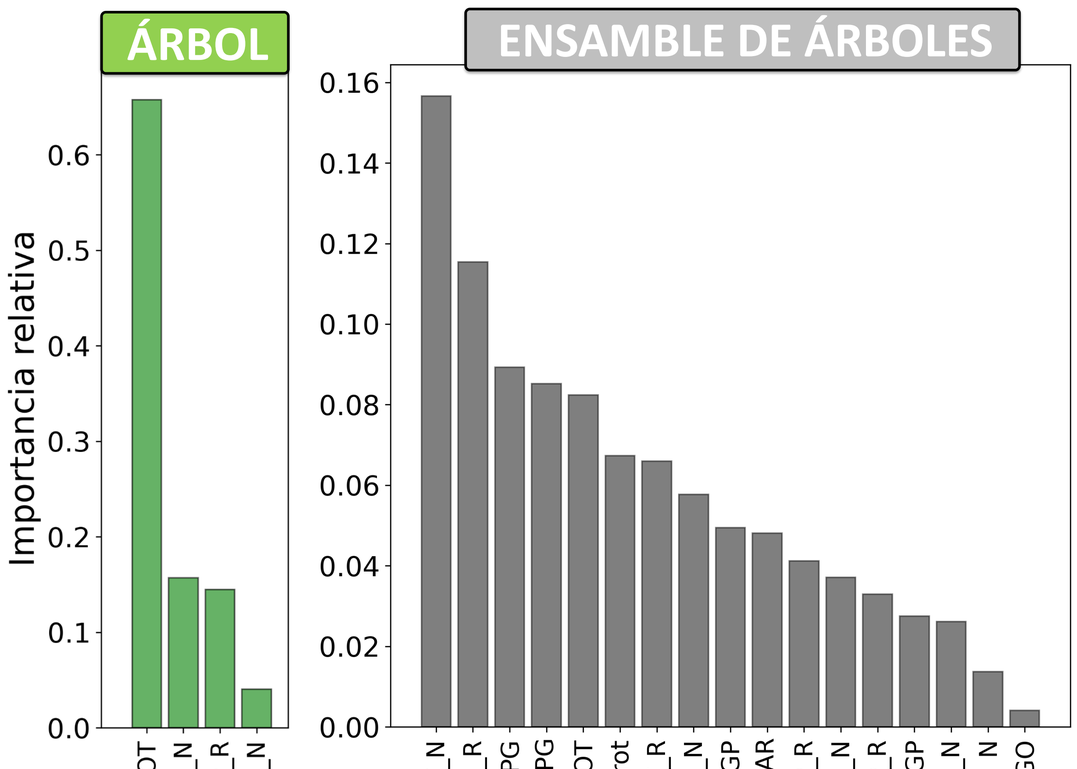

Fig. 2. Importancia relativa de las variables en la predicción de los grupos heteróticos para el modelo árbol de decisión (izquierda), y ensamble de árboles (derecha).

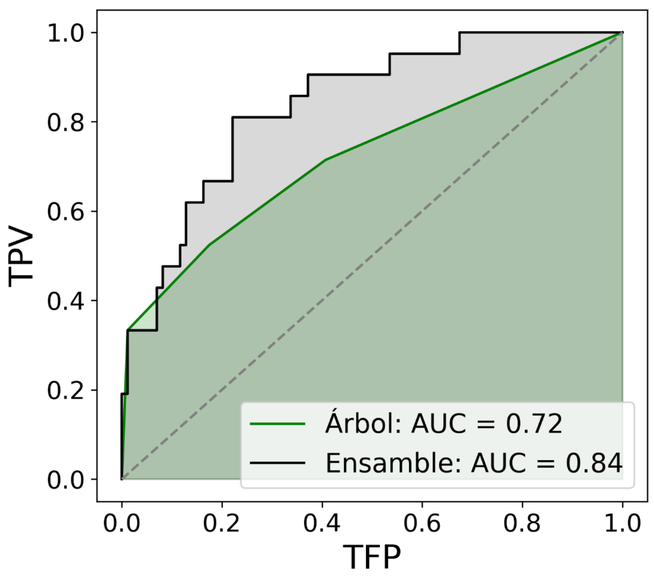

Fig. 3. Área bajo la curva (AUC) ROC a partir de la relación entre la tasa de positivos verdaderos (TPV) y de falsos positivos (TFP) para el árbol de decisión, y ensamble de árboles. La línea punteada indica la relación 1:1.

Conclusiones

- Ambos modelos lograron clasificar a las líneas endocriadas en los GHs con un nivel de precisión aceptable-alto (> 80%), sobre todo considerando el amplio rango de ambientes en el cual las líneas fueron evaluadas.
- El **%PROT en grano** fue la variable que más contribuyó a la hora de clasificar a las líneas, seguida de los componentes del rendimiento (NGP y PG).
- El **ensamble de árboles** fue más preciso en su predicción de los GHs de las líneas (mayor AUC). Sin embargo, al ser un modelo más complejo, su capacidad informativa es menor ("*black-box model*") que la de un **árbol de decisión**, por lo tanto, ambos modelos no serían mutuamente excluyentes sino complementarios en su implementación.

Este trabajo fue financiado por la ANPCyT (PICTs 1454 and 2671), la UBA (UBACYT 00493), y el INTA (PNCYO-1127042).

Referencias:
1- Melchinger 1999. Genetic diversity and heterosis. pp. 99–118. ASA, CSSA, SSSA
2- Hallauer et al. 1988. Quantitative genetics in maize breeding. Iowa State Univ. Press.
3- Moore 1987. Classification and regression trees. *Cytometry*, 8, 534-535.
4- https://github.com/microsoft/LightGBM