MAÍZIFICANDO CONCIENCIA

XII CONGRESO NACIONAL DE MAÍZ

Eje

Genética y mejoramiento

8, 9 y 10 de Noviembre
Pergamino, BA
UNNOBA

AIANBA · UNNOBA · INTA · Secretaría de Agricultura, Ganadería y Pesca · Ministerio de Economía · Argentina

2022

MAÍZIFICANDO
CONCIENCIA
XII CONGRESO NACIONAL DE MAÍZ

IDENTIFICACIÓN DE GRUPOS HETERÓTICOS EN MAÍZ MEDIANTE MODELOS PREDICTIVOS

Hisse, I.R.[1,2]; D'Andrea, K.E.[1,2]; Otegui, M. E.[1,2,3]
[1]FAUBA; [2]CONICET; [3]EEA Pergamino, INTA. Av. San Martín 4453 (C1417DSE), CABA, Argentina.
hisse@agro.uba.ar

HETEROTIC GROUP CLASSIFICATION OF MAIZE THROUGH SUPERVISED MACHINE LEARNING

Abstract

For an efficient hybrid breeding program, it is desirable to organize the germplasm into heterotic groups (HGs). Supervised learning algorithms can be useful in HG assignment from phenotype information of the lines. This study aimed to evaluate two supervised machine learning models, decision tree (DT) and light gradient boosting (LGBM), to classify maize inbreds into HGs using their phenotypic information. Six inbred lines of different germplasm (flint and dent) were evaluated across diverse environments (seven years, three nitrogen levels, and two water supply conditions). The accuracy of both models was high (DT = 0.86, LGBM = 0.85) considering the lack of identification (i.e. name) of the lines, and the broad range of explored environments. Grain protein concentration was the most relevant trait to classify inbreds into HGs in both models, followed by both grain yield components (kernel number and weight). The area under the ROC curve, a parameter of the prediction level of the model, was higher in LGBM (0.84) than in DT (0.72). Therefore, the classification of lines into HGs from easily measured phenotypic variables can be possible by combining a simpler and more informative machine learning model (DT) with a complex but more predictive one (LGBM).

Palabras claves
Modelos predictivos, líneas endocriadas, Zea mays, rendimiento, proteína en grano

Keywords
Machine learning models, inbred lines, Zea mays, grain yield, grain protein concentration

Introducción

En los programas de mejoramiento de obtención de híbridos, resulta clave organizar el germoplasma en grupos heteróticos (GH) (Melchinger, 1999; Reif *et al.*, 2007), evaluándose posteriormente los materiales en cruzamientos de prueba con individuos del GH opuesto (Hallauer y Miranda, 1998). Las ventajas del uso de GHs opuestos radican en la explotación de la heterosis, y la baja relación varianza de dominancia vs. varianza aditiva (Falconer y Mackay, 1996; Melchinger, 1999). La identificación de GHs se puede llevar a cabo evaluando el germoplasma disponible en cruzamientos de prueba a campo. Sin embargo, la gran cantidad de líneas disponibles en un programa de mejoramiento hace que la evaluación de todos los cruces posibles sea inviable, sugiriéndose el uso de marcadores moleculares para organizar el germoplasma en subgrupos, seleccionando genotipos representativos de cada subgrupo para ser evaluados posteriormente en cruzamientos de prueba (Melchinger, 1999). 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.

El árbol de decisión es un modelo predictivo supervisado conformado por nodos, en cada uno de los cuales se toma una decisión binaria o multinomial con un nivel de precisión determinado (Moore, 1987). Los árboles de decisión constituyen algoritmos de clasificación robustos, informativos, y fáciles de interpretar; no obstante, en algunas ocasiones su poder predictivo y capacidad de generalización puede ser bajo, debido principalmente al riesgo de sobreajuste que presentan. Frente a estas desventajas, los algoritmos de ensamble como *random forest* (Ho, 1995) o *lightgbm* (https://github.com/microsoft/LightGBM) que combinan múltiples modelos de árboles en uno solo, poseen un alto poder predictivo, siendo generalizable a distintas situaciones. Modelos predictivos complejos como los mencionados se han utilizado previamente en selección genómica (Heslot et al., 2012), predicción de rendimiento y momento de floración en parcelas de cría (Adak et al., 2021), y en el uso de marcadores moleculares para asignar líneas a GHs ya establecidos (Ornella y Tapia, 2010). Sin embargo, no existen reportes acerca de la implementación de modelos de *machine learning* en la clasificación de las líneas en distintos GHs a partir de caracteres fenotípicos de fácil medición. El objetivo de este trabajo consistió en 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 (flint, dentada) y en un rango amplio de ambientes (combinaciones de distintos años, niveles de N, y oferta hídrica).

Materiales y métodos

El material genético incluyó seis líneas endocriadas pertenecientes a dos grupos heteroticos: (i) GH_1 compuesto por las líneas flint LP2, LP561, LP611, LP662 y ZN6 (Olmos et al., 2014), desarrolladas por el programa de mejoramiento de maíz del INTA Pergamino; y (ii) GH_2, integrado por la línea B100 (Reid Yellow Dent; Hallauer et al., 1995). Las líneas se evaluaron en la Estación Experimental Pergamino del INTA, Argentina (33°56´ S, 60°34´ O), sobre un suelo Argiudol Típico, y durante siete años (2002/03, 2003/04, 2004/05, 2006/07, 2008/09, 2013/14, y 2014/15). En todos los años se incluyeron dos niveles de nitrógeno (N), un control sin aplicación de N, y una condición de alto N fertilizado con 400 (2002/03, 2003/04, y 2004/05) y 200 kg N ha^{-1} (2006/07, 2008/09, 2013/14, y

2014/15). En todos los años se aplicó riego, salvo en el 2006/07 y 2008/09 donde se incluyó además una condición de secano. Los experimentos se mantuvieron limpios de malezas, plagas y enfermedades.

Se empleó un diseño experimental de parcelas divididas con tres repeticiones, con la condición de N en la parcela principal (y el tratamiento de riego en 2006/07 y 2008/09), y las líneas endocriadas en las sub-parcelas (de aquí en adelante denominadas parcelas). Cada parcela tuvo tres hileras, las cuales estaban distanciadas entre sí a 0,7 m y con una longitud de 5,5 m. La densidad fue siempre de 7 plantas m^{-2}. En V_3 (Ritchie *et al.,* 1992) se marcaron siete plantas sucesivas en la hilera central de cada parcela. Las plantas marcadas se cosecharon en madurez fisiológica (R_6) para la obtención del rendimiento en grano por planta (RGP, en g $planta^{-1}$), número de granos por planta (NGP), y peso de los granos (PG, cociente entre RGP y NGP, en mg $grano^{-1}$) luego de secar el material en estufa a 70°C hasta peso constante. La concentración de proteína en grano (PROT, en %) a cosecha se obtuvo mediante un análisis de espectroscopia del infrarrojo cercano (NIR).

Además de las variables medidas, se calcularon nuevas variables para ser incluidas en los modelos: (i) RG_PROT, para expresar la proteína en grano en rendimiento de proteína (en g $planta^{-1}$); (ii) Ratio NGP/PG, a partir de cociente entre NGP y PG; y (iii) aquellas a las que se les añadió el "_N" y "_R" (e.g. NGP_N), las cuales se computaron por la mediana de cada nivel de N y condición de riego, respectivamente. También se incluyeron las variables categóricas año, nivel de N, y condición de riego, dando un total de 17 variables conformando la base de datos. La identificación de cada línea por su nombre se eliminó del set de datos. La variable objetivo a predecir es el GH, la cual es binaria ya que se sabe a priori que se cuenta con dos GHs. Para todo el preprocesamiento mencionado se utilizó *pandas y numpy packages* de python en colab (https://colab.research.google.com/).

Se utilizaron dos modelos predictivos de clasificación, un árbol de decisión (*dtreeviz package*) y un ensamble de árboles (*lightgbm package*). Inicialmente, el set de datos completo (n = 428) se dividió aleatoriamente en dos subconjuntos, uno de entrenamiento, conformado por el 66% de la base y excluyendo la variable objetivo a predecir (i.e. GH), y uno de testeo conformado por el 34% restante con el fin de evaluar la capacidad predictiva de los modelos (*sklearn package*). Para evaluar la capacidad de predicción se usaron dos métricas, la efectividad (*accuracy*) y el área bajo la curva ROC (*receiver operating characteristic*). La efectividad indica la proporción de predicciones correctas respecto del total de observaciones predichas. La curva ROC se obtiene a partir de la respuesta de la tasa de los positivos verdaderos (TPV) a la tasa de los falsos positivos (TFP, o Error de Tipo 1), siendo el área bajo la curva ROC (AUC), su integral. Ambas métricas, i.e. efectividad y AUC, oscilan entre 0 (predicción nula) y 1 (predicción perfecta).

Resultados y discusión

Se exploró un amplio rango de valores para las variables evaluadas (Tabla 1) producto del uso de líneas endocriadas de origen diverso, y de la amplia gama de ambientes explorados (combinación de años, niveles de N, y oferta hídrica). El GH_2 (línea dentada) fue, en promedio, superior al GH_1 para RGP y componentes numéricos, pero lo inverso se observó para el porcentaje de proteína en grano (GH_1 > GH_2; Tabla 1), en concordancia con lo observado al comparar materiales flint vs. dentados (Tamagno et al., 2015; Hisse et al., 2021).

Variable	Promedio		Q_25	Q_75	Mín.-Máx.
	GH_1	GH_2			
NGP *(granos pl^{-1})*	246	273	177	318	15 - 536
PG *(mg grano^{-1})*	190	204	175	213	82 - 280
RGP *(g pl^{-1})*	47	56	33,6	62,6	2,6 - 114
PROT *(%)*	11,2	9,9	10,2	12,0	6,2 - 15,2
RG_PROT *(g pl^{-1})*	5,2	5,4	3,5	6,6	0,25 - 13,4
Ratio NGP/PG	1,31	1,37	0,94	1,6	0,13 - 3,8
NGP_N *(granos pl^{-1})*	0,99	1,09	0,73	1,2	0,06 - 2,1
PG_N *(mg grano^{-1})*	0,97	1,04	0,89	1,1	0,42 - 1,4
RGP_N *(g pl^{-1})*	0,97	1,14	0,69	1,3	0,04 - 2,3
PROT_N *(%)*	1,02	0,90	0,92	1,1	0,59 - 1,5
NGP_R *(granos pl^{-1})*	0,98	1,08	0,73	1,25	0,06 - 2,1
PG_R *(mg grano^{-1})*	0,97	1,05	0,90	1,1	0,45 - 1,5
RGP_R *(g pl^{-1})*	0,99	1,16	0,71	1,3	0,03 - 2,3
PROT_R *(%)*	1,01	0,90	0,92	1,1	0,56 - 1,4

Tabla 1. Valores promedio de ambos grupos heteróticos (GH_1, GH_2) y estadísticos descriptivos para las variables evaluadas. Q_25 y Q_75 representan el primer y tercer cuartil, respectivamente

La efectividad del árbol de decisión para predecir los GHs de las líneas fue de 0,86, lo que representa un valor elevado considerando la diversidad de ambientes en los que se evaluó el germoplasma, y la ausencia de identificación de cada una de las líneas (la variable "nombre" se eliminó del set de datos). El atributo PROT fue el más importante a la hora de dividir ambos GHs (Fig. 2), fijándose como punto de corte un valor de proteína en grano del 10,25% (nodo 0; Fig. 1). Dicho atributo fue seguido por PROT_N (nodo 6; Fig. 1), y ambos (PROT + PROT_N) explicaron alrededor del 70% de la variabilidad total explorada por el modelo (Fig. 2). Los componentes numéricos NGP y PG relativizados por la condición de riego (NGP_R) y N (PG_N) fueron más importantes que el RGP para predecir los GHs, siendo más relevante el uso de ambos componentes que el RGP en sí mismo a la hora de caracterizar a las líneas.

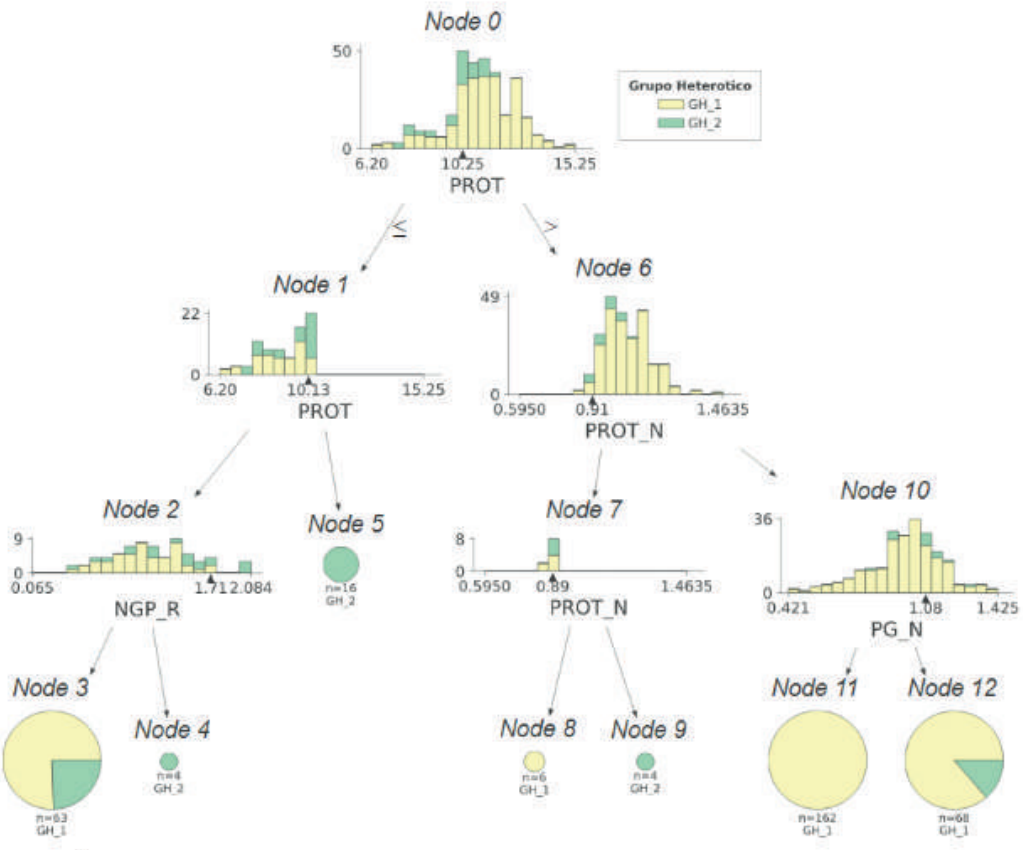

Fig. 1. Árbol de decisión para la identificación de dos grupos heteróticos a partir de 17 variables evaluadas en seis líneas endocriadas a lo largo de diversos ambientes (combinaciones incompletas de siete años, tres niveles de nitrógeno y dos condiciones de oferta hídrica; n = 424). El GH_1 corresponde a las líneas flint LP2, LP561, LP611, LP662, ZN6, y el GH_2 a la línea dentada B100

Cuando se evaluó el modelo de ensamble de árboles, su efectividad fue similar a la del árbol de decisión (0,85); no obstante, su área bajo la curva ROC (AUC) fue considerablemente más alta (Fig. 3), resultando un modelo más preciso en su predicción de los GHs de las líneas. La importancia relativa de los atributos del modelo no fue tan contrastante como la observada para el árbol de decisión (Fig. 2), como era de esperarse. Esto se debe a la naturaleza de los modelos de ensamble, los cuales incluyen muchos árboles débiles que aprenden conjuntamente, obteniéndose resultados más robustos (Ke et al., 2017). La variable PROT_N fue la más relevante en términos de contribución al modelo (Fig. 2), seguida por el PG (PG_R y PG). Nuevamente, el RGP no fue relevante para predecir el GH de las líneas.

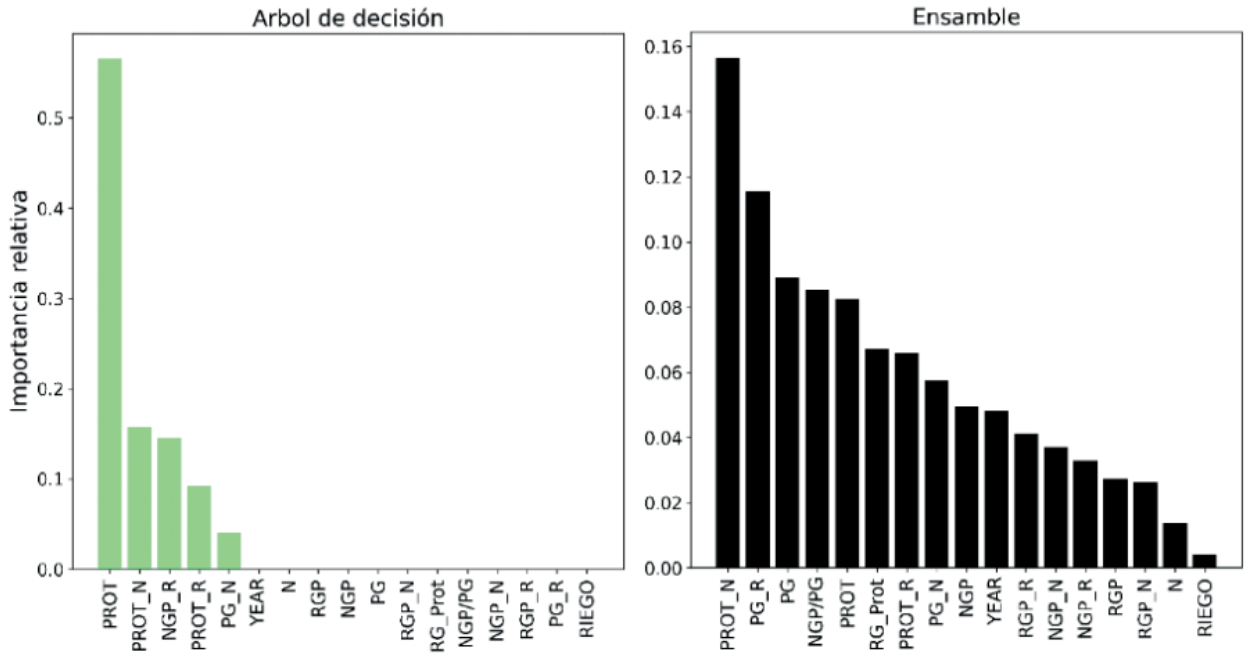

Fig. 2. Importancia relativa de las variables en la predicción de los grupos heteróticos de seis líneas endocriadas para los modelos de árbol de decisión (izquierda) y ensamble de árboles (derecha).

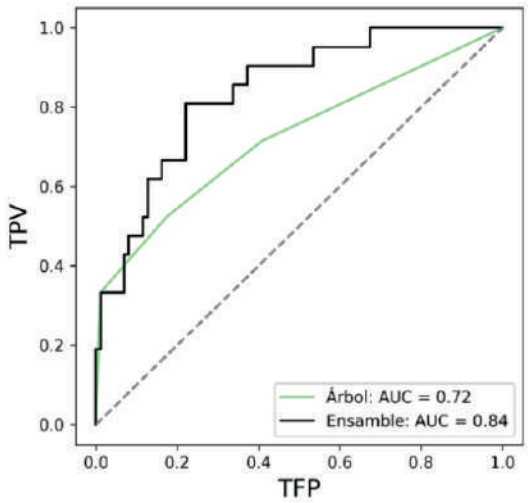

Fig. 3. Curva ROC y área bajo la curva (AUC) a partir de la relación entre la tasa de positivos verdaderos (TPV) y de falsos positivos (TFP) de los modelos de predicción árbol de decisión, y ensamble de árboles. La línea punteada indica la relación 1:1.

Conclusión

Ambos modelos de aprendizaje supervisado lograron clasificar a las líneas endocriadas en los GHs con un nivel de efectividad aceptable-alto (> 80%), sobre todo considerando que la identificación de las líneas por su nombre fue eliminada como variable informativa, y que las líneas se evaluaron en un rango amplio de ambientes. Esto resalta la importancia en el uso de modelos predictivos a la hora de identificar a las líneas parentales en los distintos GHs a partir de atributos fenotípicos de fácil estimación. La concentración de proteína 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 modelo de ensamble de árboles fue más preciso en su predicción de los GHs de las líneas (mayor AUC), siendo un modelo más adecuado para ser implementado en la clasificación de las líneas en los GHs. Sin embargo, al ser un modelo más complejo, su capacidad informativa es menor ("*black box model*") que la de un árbol de decisión ("*white box*"), de lo que se desprende que ambos modelos no serían mutuamente excluyentes sino complementarios en su implementación.

Apoyo financiero

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

Referencias bibliográficas

Adak, A.; Murray, S.C.; Božinović, S.; Lindsey, R.; Nakasagga, S.; Chatterjee, S.; Anderson, S.L., II; Wilde, S. 2021. Temporal Vegetation Indices and Plant Height from Remotely Sensed Imagery Can Predict Grain Yield and Flowering Time Breeding Value in Maize via Machine Learning Regression. Remote Sens., 13, 2141. https://doi.org/10.3390/rs13112141

Falconer, D.S.; Mackay, T.F.C. 1996. Introduction to quantitative genetics. Longman, London.

Hallauer, A.R.; Miranda, J.B. 1988. Quantitative genetics in maize breeding. Iowa State Univ. Press, Ames.

Hallauer, A.P.; Lamkey, K.R.; Russell, W.A.; White, P.R. 1995. Registration of B99 and B100 inbred lines of maize. Crop Science, 35, 1714–1715.

Heslot, N.; Yang, H.P.; Sorrells, M.E.; Jannink, J.L. 2012. Genomic Selection in Plant Breeding: A Comparison of Models. Crop Science, 52, 146–160. https://doi.org/10.2135/cropsci2011.06.0297

Hisse, I.R.; D'Andrea, K.E.; Otegui, M.E. 2021. Kernel weight responses to the photothermal environment in maize dent × flint and flint × flint hybrids. Crop Science, 61, 1996–2011.

Ho, T.K. 1995. Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC. pp. 278–282.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems, 30.

Melchinger, A.E. 1999. Genetic diversity and heterosis. pp. 99–118. In J.G. Coors and S. Pandey (ed.) The genetics and exploitation of heterosis in crops. ASA, CSSA, and SSSA, Madison, WI.

Moore, D.H. 1987. Classification and regression trees, by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Cytometry, 8, 534-535. https://doi.org/10.1002/cyto.990080516

Ornella, L.; Tapia, E. 2010. Supervised machine learning and heterotic classification of maize (Zea mays L.) using molecular marker data. Computers and Electronics in Agriculture, 74, 250–257. https://doi.org/10.1016/j.compag.2010.08.013

Olmos, S.E.; Delucchi, C.; Ravera, M.; Negri, M.E.; Mandolino, C.; Eyherabide, G.H. 2014. Genetic relatedness and population structure within the public Argentinean collection of maize inbred lines. Maydica 40, 16–31.

Reif, J.C.; Gumpert, F.M.; Fischer, S.; Melchinger, A.E. 2007. Impact of interpopulation divergence on additive and dominance variance in hybrid populations. Genetics 176, 1931–1934.

Ritchie, S.W.; Hanway, J.J.; Benson, G.O. 1992. How a plant crop develops. Iowa State University of Science and Technology, Coop. Ext. Serv., Ames, Iowa, USA.

Tamagno, S.; Greco, I.A.; Almeida, H.; Borrás, L. 2015. Physiological differences in yield related traits between flint and dent Argentinean commercial maize genotypes. Eur. J. Agron., 68, 50–56.