

EVALUACIÓN DE MODELOS PARAMÉTRICOS Y DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN GENÓMICA DE VARIABLES AGRONÓMICAS EN HÍBRIDOS DE MAÍZ (*Zea mays* L.) EN LA ZONA NÚCLEO DE ARGENTINA: AVANCES PRELIMINARES

FASSIO N.^{1,2}; MISSIMA J.³; BARRETO C.³; FEDERICO M.L.^{1,4}; SCHUSTER I.³

¹ Universidad Nacional del Noroeste de la Provincia de Buenos Aires (UNNOBA), Doctorado en Mejoramiento Genético, Pergamino, Buenos Aires, Argentina.

² Longping High-Tech Argentina.

³ Longping High-Tech Brasil.

⁴ Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Estación Experimental Agropecuaria Pergamino, Instituto Nacional de Tecnología Agropecuaria (INTA), Buenos Aires, Argentina

Correo electrónico: nicolasfassio@outlook.com

EVALUATION OF PARAMETRIC AND MACHINE LEARNING MODELS FOR GENOMIC PREDICTION OF AGRONOMIC TRAITS IN MAIZE HYBRIDS (*Zea mays* L.) IN THE ARGENTINE CORE AGRICULTURAL ZONE: PRELIMINARY ADVANCES

Genomic selection (GS) has become a key tool to accelerate genetic gain by predicting breeding values of genotyped individuals prior to phenotypic evaluation. However, the relative efficacy of parametric and machine learning (ML) models for genomic prediction depends on the trait, population and environment considered. This study aims to compare three parametric models and four ML models for genomic prediction of seven agronomic traits in 1,000 maize single-cross hybrids evaluated across three environments of the core production area of Argentina. As a preliminary advance, phenotypic and genotypic data from two environments (Colón, early and late sowing dates) were used to train and validate four models (RRBLUP, RF - Random Forest, GBM - Gradient Boosting Machine and DNN - Deep Neural Network) for two traits (yield and harvest moisture) using a three-fold cross-validation scheme. RRBLUP showed the best overall performance, particularly for harvest moisture, whereas RRBLUP and RF exhibited similar predictive ability for yield. DNN showed the lowest performance for both traits. These preliminary results will be expanded to the remaining environments, campaigns, traits and methods to identify the best-performing model(s) for genomic prediction of maize hybrids in the core region of Argentina.

Palabras clave: predicción genómica; aprendizaje automático; híbridos de maíz; modelos paramétricos.

Keywords: genomic prediction; machine learning; maize hybrid; parametric models.

Introducción

El maíz (*Zea mays* L.) es un cultivo estratégico para la Argentina. En la campaña 2024/25 la producción nacional se situó en torno a 48,5 millones de toneladas, posicionando al país como el tercer mayor exportador mundial del cereal (Bolsa de

Cereales de Rosario, 2025; USDA, 2025). Sin embargo, la creciente variabilidad climática y el aumento en la frecuencia de eventos extremos representan un desafío crítico para sostener e incrementar la productividad del cultivo. Frente a este escenario, el mejoramiento genético constituye una herramienta clave para desarrollar híbridos más productivos, estables y adaptados a diferentes condiciones ambientales (Budhlakoti et al., 2022).

Para acelerar este proceso, la selección genómica (SG) permite estimar el valor genético de individuos genotipados, pero no fenotipados, a partir de marcadores de polimorfismo de nucleótido simple (SNP) distribuidos a lo largo del genoma (Meuwissen et al., 2001; Goddard, 2009). En comparación con la selección asistida por marcadores (Lande & Thompson, 1990), la SG incrementa la eficiencia al reducir costos e intervalos generacionales, al tiempo que aumenta la precisión y la intensidad de selección (Crossa et al., 2017; Yan, 2021). Sin embargo, la eficacia relativa de los distintos modelos de predicción genómica —tanto paramétricos clásicos como de aprendizaje automático (ML)— no es constante, sino que depende de la arquitectura genética del carácter (Holland, 2007), el tamaño de la población de entrenamiento y las condiciones ambientales evaluadas (Heslot et al., 2012; Azodi et al., 2019; Wang et al., 2018).

El objetivo general de este trabajo es evaluar y comparar la eficacia de modelos paramétricos y de ML en la predicción genómica de siete variables agronómicas en híbridos simples de maíz, a partir de datos fenotípicos y genotípicos obtenidos en tres localidades de la región núcleo de Argentina. Se presenta aquí un avance preliminar de dicha evaluación contemplando dos variables, rendimiento (T/ha) y humedad de cosecha (%) en dos ambientes (Colón, Buenos Aires, en fecha de siembra temprana y tardía) y cuatro modelos de predicción, un modelo paramétrico clásico, *Ridge Regression* BLUP (RRBLUP) (Meuwissen et al., 2001; Meuwissen, 2009), y tres modelos de aprendizaje automático: *Random Forest* (RF) y *Gradient Boosting Machine* (GBM) (González-Recio & Forni, 2011; González-Recio et al., 2013), y una Red Neuronal Profunda (DNN) (Okut et al., 2011).

Materiales y Métodos

Se evalúan híbridos simples generados a partir de líneas endogámicas de distintos grupos heteróticos del banco de germoplasma de Longping High-Tech Argentina. Para este avance preliminar se utilizaron los datos fenotípicos y genotípicos correspondientes a dos ambientes ya evaluados de la campaña 25/26 (Colón, fechas de siembra temprana y tardía; 1.461 y 1.064 híbridos evaluados, respectivamente), sobre los cuales se aplicó previamente una corrección espacial de los valores fenotípicos mediante el mejor Estimador Lineal Insesgado (eBLUES).

Las líneas parentales fueron genotipadas con un panel de 5.076 marcadores SNP (plataforma Ion Gene Studio S5 Prime, Thermo Fisher Scientific), y el genotipo de cada híbrido se obtuvo por concatenación de los genotipos de sus dos líneas parentales, siguiendo el enfoque utilizado en estudios previos de predicción genómica en híbridos de maíz (Riedelsheimer et al., 2012).

El análisis se realizó en R 4.4.2 (R Core Team, 2024), evaluando en esta etapa cuatro modelos de predicción genómica sobre dos de las siete variables del estudio:

rendimiento (T/ha) y humedad de cosecha (%). Los modelos evaluados fueron un modelo paramétrico clásico, RRBLUP, y tres modelos de aprendizaje automático: RF, GBM y DNN. Los modelos restantes contemplados en el plan de tesis —GBLUP (De Los Campos et al., 2009), Bayesian Lasso Regression (Usai et al., 2009) y LightGBM— se incorporarán en las próximas etapas del trabajo.

Para cada modelo, los datos se dividieron aleatoriamente en poblaciones de entrenamiento (66,7%) y de validación (33,3%), repitiendo el procedimiento en un esquema de validación cruzada de tres particiones (*folds*). El desempeño de cada modelo se estimó mediante tres métricas calculadas en cada *fold*: (i) precisión predictiva, definida como la correlación de Pearson entre los valores genómicos predichos y los valores fenotípicos observados y (ii) la correlación de Spearman, que mide la asociación entre el orden (*ranking*) de los valores predichos y observados. Esta última métrica es especialmente relevante en mejoramiento genético porque la selección de genotipos superiores depende, en la práctica, de su posición relativa dentro de un *ranking*, más que del valor absoluto predicho para cada uno (Montesinos-López et al., 2022).

Resultados Preliminares

Las **Tablas 1 y 2** muestran las métricas de desempeño de los cuatro modelos evaluados para **rendimiento** y **humedad de cosecha**, respectivamente. Los datos corresponden a la campaña actual e incluyen ambos ambientes analizados (Colón, siembra temprana y tardía).

Tabla 1. Desempeño predictivo (precisión de la predicción y correlación de Spearman) de cuatro modelos de predicción genómica para rendimiento, en dos ambientes (Colón, siembra temprana y tardía) evaluados mediante validación cruzada de tres particiones.

Localidad	Fecha de siembra	Método	Precisión de la predicción	Correlación de Spearman
Colón	Temprana	RRBLUP	0.163	0.158
Colón	Temprana	RF	0.164	0.155
Colón	Temprana	GBM	0.107	0.117
Colón	Temprana	DNN	0.028	0.02
Colón	Tardía	RRBLUP	0.235	0.237
Colón	Tardía	RF	0.236	0.231
Colón	Tardía	GBM	0.231	0.221
Colón	Tardía	DNN	0.177	0.172

Tabla 2. Desempeño predictivo (precisión de la predicción y correlación de Spearman) de cuatro modelos de predicción genómica para humedad de cosecha, en los mismos dos ambientes (Colón, siembra temprana y tardía) evaluados mediante validación cruzada de tres particiones.

Localidad	Fecha de siembra	Método	Precisión de la predicción	Correlación de Spearman
Colón	Temprana	RRBLUP	0.837	0.823
Colón	Temprana	RF	0.824	0.815
Colón	Temprana	GBM	0.826	0.810
Colón	Temprana	DNN	0.790	0.785
Colón	Tardía	RRBLUP	0.843	0.837
Colón	Tardía	RF	0.831	0.821
Colón	Tardía	GBM	0.834	0.831
Colón	Tardía	DNN	0.823	0.813

Para humedad de cosecha, RRBLUP mostró el mejor desempeño en ambas métricas en los dos ambientes (precisión de la predicción = 0,837 y 0,843; correlación de Spearman = 0,823 y 0,837, en Colón temprano y tardío, respectivamente), mientras que la Red Neuronal Profunda (DNN) mostró el desempeño más bajo en ambos ambientes (0,790/0,785 y 0,823/0,813). Para rendimiento, RRBLUP y RF alcanzaron una precisión de la predicción similar en los dos ambientes (0,163 frente a 0,164 en Colón temprano; 0,235 frente a 0,236 en Colón tardío), aunque RRBLUP presentó una correlación de Spearman superior en ambos casos (0,158 frente a 0,155 en Colón temprano; 0,237 frente a 0,231 en Colón tardío), lo que indica un mejor ordenamiento relativo de los genotipos. DNN mostró nuevamente el desempeño más bajo, de forma particularmente marcada en Colón temprano (precisión de la predicción = 0,028). Los modelos de ML basados en árboles (RF y GBM) se ubicaron en una posición intermedia en ambas variables y ambientes.

Se observa además una diferencia marcada en la magnitud de la precisión predictiva entre variables: mientras que la humedad de cosecha alcanzó valores de precisión de entre 0,79 y 0,84 en los cuatro modelos y los dos ambientes, el rendimiento mostró valores considerablemente menores y más variables (0,03 a 0,24), consistente con la mayor complejidad de la arquitectura genética y la interacción ambiental que habitualmente presenta esta última variable (Holland, 2007) (Figura 1).

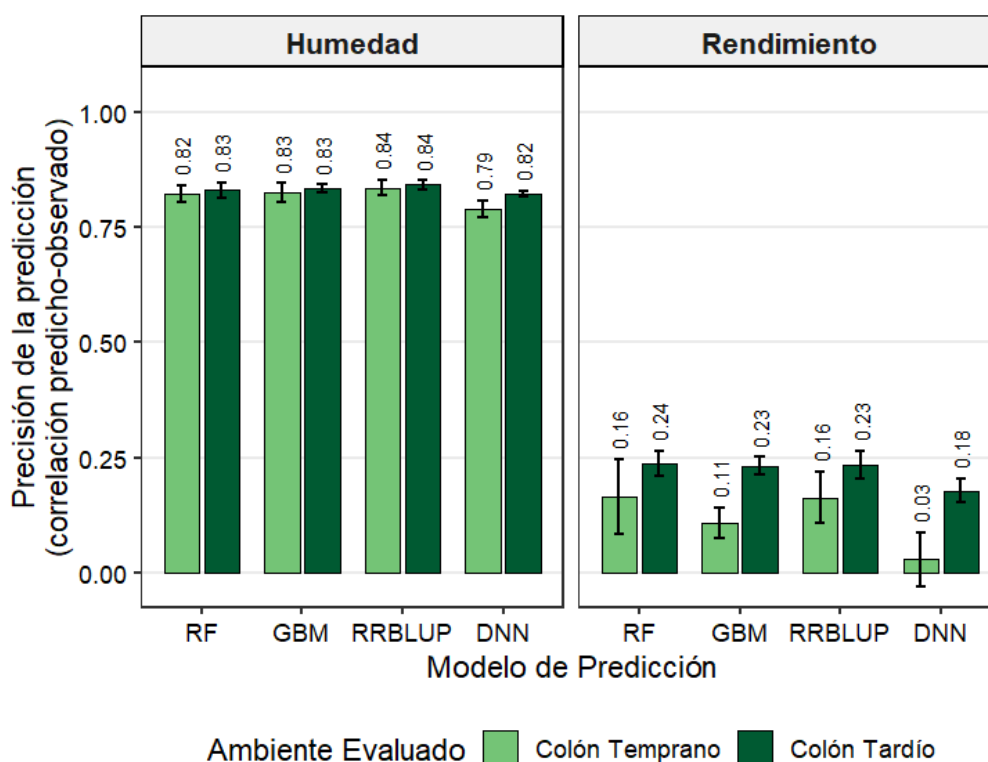


Figura 1. Precisión predictiva de cuatro modelos de predicción genómica (RF, GBM, RRBLUP y DNN) para humedad de cosecha y rendimiento en dos ambientes (Colón Temprano y Colón Tardío). Las barras representan la precisión promedio (correlación de Pearson) obtenida mediante validación cruzada de 3 particiones, con el valor numérico indicado sobre cada una. Las barras de error representan el desvío estándar.

Conclusiones y Perspectivas

Estos resultados preliminares, obtenidos con datos de dos ambientes, indican que el modelo paramétrico clásico (RRBLUP) y el modelo de aprendizaje automático *Random Forest* (RF) presentaron los desempeños más consistentes entre los modelos evaluados para la predicción genómica de rendimiento y humedad de cosecha en híbridos de maíz. Mientras que RRBLUP mostró el mejor desempeño para humedad de cosecha en ambos ambientes, para rendimiento ambos modelos exhibieron una capacidad predictiva muy similar, con diferencias mínimas según la métrica considerada. Este patrón fue consistente en ambos ambientes y coincide con antecedentes que indican que ningún algoritmo supera de manera sistemática a los demás en todos los escenarios evaluados (Azodi et al., 2019; Heslot et al., 2012; Montesinos-López et al., 2022).

Las próximas etapas del trabajo incorporarán más ambientes y las cinco variables agronómicas adicionales contempladas en el estudio (altura de planta, altura de espiga, quebrado, vuelco y enfermedades foliares), junto con los modelos aún no evaluados (GBLUP, *Bayesian Lasso Regression* y *LightGBM*) y el esquema completo de validación cruzada (seis repeticiones). Asimismo, se contempla explorar enfoques más recientes que integran datos ambientales y arquitecturas de aprendizaje profundo en la predicción genómica de híbridos (Montesinos-López et al., 2025; He et al., 2025; Zhang et al., 2025), en línea con antecedentes metodológicos sobre la comparación de modelos estadísticos

y de aprendizaje automático en ensayos multi-ambientales de maíz (Barreto et al., 2024), con el objetivo de determinar qué modelos ofrecen el mejor desempeño predictivo para diferentes variables agronómicas en híbridos de maíz y ambientes de la zona núcleo de Argentina.

Apoyo financiero

Este trabajo forma parte del programa de mejoramiento genético de Longping High-Tech Argentina, empresa que aportó el germoplasma, la infraestructura de ensayos a campo y el genotipado utilizados en este estudio.

Referencias bibliográficas

- Azodi, C.B.; Bolger, E.; McCarren, A.; Roantree, M.; de los Campos, G.; Shiu, S.H. (2019). Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3: Genes, Genomes, Genetics*, 9(11), 3691–3702.
- Barreto, C.A.V.; Dias, K.O.G.; de Sousa, I.C.; Azevedo, C.F.; Nascimento, A.C.C.; Guimarães, L.J.M.; Guimarães, C.T.; Pastina, M.M.; Nascimento, M. (2024). Genomic prediction in multi-environment trials in maize using statistical and machine learning methods. *Scientific Reports*, 14, 1062.
- Bolsa de Cereales de Rosario. (2025). Informe especial 185. Bolsa de Cereales de Rosario.
- Budhlakoti, N.; Kushwaha, A.K.; Rai, A.; Chaturvedi, K.K.; Kumar, A.; Pradhan, A.K.; Kumar, U.; Kumar, R.R.; Juliana, P.; Mishra, D.C.; Kumar, S. (2022). Genomic selection: a tool for accelerating the efficiency of molecular breeding for development of climate-resilient crops. *Frontiers in Genetics*, 13, 832153.
- Crossa, J.; Pérez-Rodríguez, P.; Cuevas, J.; Montesinos-López, O.; Jarquín, D.; de los Campos, G.; Burgueño, J.; González-Camacho, J.M.; Pérez-Elizalde, S.; Beyene, Y.; Dreisigacker, S.; Singh, R.; Zhang, X.; Gowda, M.; Roorkiwal, M.; Rutkoski, J.; Varshney, R.K. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science*, 22(11), 961–975.
- De Los Campos, G.; Naya, H.; Gianola, D.; Crossa, J.; Legarra, A.; Manfredi, E.; Weigel, K.; Cotes, J.M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182(1), 375–385.
- Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136(2), 245–257.
- González-Recio, O.; Forni, S. (2011). Genome-wide prediction of discrete traits using bayesian regressions and machine learning. *Genetics Selection Evolution*, 43(1), 7.
- González-Recio, O.; Jiménez-Montero, J.A.; Alenda, R. (2013). The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. *Journal of Dairy Science*, 96(1), 614–624.
- He, K.; Yu, T.; Gao, S.; Chen, S.; Li, L.; Zhang, X.; Huang, C.; Xu, Y.; Wang, J.; Prasanna, B.M.; Hearne, S.; Li, X.; Li, H. (2025). Leveraging automated machine learning for environmental data-driven genetic analysis and genomic prediction in maize hybrids. *Advanced Science*, 12(17), 2412423.
- Heslot, N.; Yang, H.P.; Sorrells, M.E.; Jannink, J.L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Science*, 52(1), 146–160.
- Holland, J.B. (2007). Genetic architecture of complex traits in plants. *Current Opinion in Plant Biology*, 10(2), 156–161.
- Lande, R.; Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124(3), 743–756.
- Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4), 1819–1829.
- Meuwissen, T. (2009). Accuracy of breeding values of “unrelated” individuals predicted by dense SNP genotyping. *Genetics Selection Evolution*, 41(1), 35.
- Montesinos-López, A.; Montesinos-López, O.A.; Ramos-Pulido, S.; Mosqueda-González, B.A.; Guerrero-Arroyo, E.A.; Crossa, J.; Ortiz, R. (2025). Artificial intelligence meets genomic selection: comparing deep learning and GBLUP across diverse plant datasets. *Frontiers in Genetics*, 16.

- Montesinos-López, O.A.; Montesinos-López, A.; Cano-Paez, B.; Hernández-Suárez, C.M.; Santana-Mancilla, P.C.; Crossa, J. (2022). A comparison of three machine learning methods for multivariate genomic prediction using the sparse kernels method (SKM) library. *Genes*, 13(8), 1494.
- Montesinos López, O. A., López, A. M., & Crossa, J. (2022). Multivariate Statistical Machine Learning Methods for Genomic Prediction. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-89010-0>
- Okut, H.; Gianola, D.; Rosa, G.J.M.; Weigel, K.A. (2011). Prediction of body mass index in mice using dense molecular markers and a regularized neural network. *Genetics Research*, 93(3), 189–201.
- R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Riedelsheimer, C.; Czedik-Eysenberg, A.; Grieder, C.; Lisec, J.; Technow, F.; Sulpice, R.; Altmann, T.; Stitt, M.; Willmitzer, L.; Melchinger, A.E. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genetics*, 44(2), 217–220.
- Usai, M.G.; Goddard, M.E.; Hayes, B.J. (2009). LASSO with cross-validation for genomic selection. *Genetics Research*, 91(6), 427–436.
- USDA. (2025). Commodity explorer version 4.0.
- Wang, X.; Xu, Y.; Hu, Z.; Xu, C. (2018). Genomic selection methods for crop improvement: current status and prospects. *The Crop Journal*, 6(4), 330–340.
- Yan, W. (2021). A systematic narration of some key concepts and procedures in plant breeding. *Frontiers in Plant Science*, 12, 724517.
- Zhang, D.; Yang, F.; Li, J.; Liu, Z.; Han, Y. et al. (2025). Progress and perspectives on genomic selection models for crop breeding. *Technology in Agronomy*, 5, e006.