ARPHA Preprints

PENSOFT

PREPRINT

Author-formatted, not peer-reviewed document posted on 11/02/2025

DOI: https://doi.org/10.3897/arphapreprints.e149212

imanr: An R Tool for the Identification of Mexican Native Maize Complexes

Arturo Sanchez-Porras, Aline Romero-Natale, Otilio Acevedo-Sandoval, Edlin Guerra-Castro

ARPHA Preprints

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

imanr: An R Tool for the Identification of Mexican Native Maize Complexes

Arturo Sanchez-Porras[‡], Aline Romero-Natale[§], Otilio Arturo Acevedo-Sandoval[§], Edlin Guerra-Castro[‡]

‡ Escuela Nacional de Estudios Superiores Unidad Mérida, Mérida, Mexico
§ Universidad Autónoma del Estado de Hidalgo, Pachuca Hidalgo, Mexico

Corresponding author: Arturo Sanchez-Porras (sp.arturo@gmail.com)

Abstract

The conservation of the genetic diversity of native maize in Mexico is a priority due to its cultural, agricultural, and environmental importance. This study presents the development and evaluation of the *imanr* package, a computational tool based on Boosted Ensembles designed to automate the classification of racial complexes of native maize. Using a national database, a model was implemented that leverages morphological and geographical variables to provide precise and rapid classifications. The methodology included the optimization of key parameters through cross-validation, achieving up to 90% in balanced accuracy and a Cohen's Kappa coefficient of 0.84. These results highlight the robustness of the model compared to traditional methods, which rely on subjective expert judgment and require extended evaluation times. The findings demonstrate that the package not only surpasses conventional methods in terms of efficiency but also offers an accessible tool for conserving and monitoring native maize diversity, aligning with the recommendations of the Global Maize Project (PGMN). Moreover, its usability was enhanced by developing a graphical user interface, allowing non-specialized users to fully utilize its potential. *imanr* represents a significant advancement in native maize conservation science, contributing to the modernization of identification processes and strengthening sustainable management strategies for this essential genetic resource. This model directly addresses the need for innovative tools to monitor and preserve maize diversity in Mexico and suggests a promising pathway for future applications in the agricultural sector.

Keywords

Agrobiodiversity management; Boosted Ensemble classification; Genetic diversity monitoring; Morphological and geographical data; Native maize biodiversity

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

Introduction

Maize (*Zea mays L.*) represents an invaluable biocultural resource, particularly in Mexico, where its genetic, historical, and cultural diversity is deeply intertwined with national identity and food security. The modern maize consumed today is the result of the domestication of its ancestor, known as *teosinte* (*Zea mays ssp. parviglumis*), a process initiated by Mesoamerican peoples between 9,000 and 5,000 BCE (Caballero-García et al. 2019, Gouttefanjat 2020). Over millennia, this co-evolution of domestication and migration has enabled farmers across diverse Mexican regions to develop unique maize landraces, adapted to specific local conditions and fulfilling nutritional, flavor, and resilience demands for various climates and soils (González González et al. 2023, Reza-Solis et al. 2024).

In 2011, the National Commission for Knowledge and Use of Biodiversity (CONABIO) published the results of the Global Maize Project (PGMN), whose primary objective was the identification and classification of maize landraces in Mexico, in compliance with the Biosafety Law for Genetically Modified Organisms (CONABIO 2011). This initiative identified 64 maize landraces in the country, of which 59 were classified as native, while additional undescribed landraces were also suggested. Based on genetic, morphological, and adaptive characteristics, these 59 native landraces were grouped into seven racial complexes, each with a defined geographic distribution and specific phenotypic traits (Sánchez-González 2011). The project provided recommendations for conserving this biodiversity, including flexible monitoring strategies and support for farmers through subsidies and certifications.

Accurate identification of native maize landraces in Mexico is crucial for conserving genetic diversity and understanding the racial complexes that comprise this vital agricultural resource (González-Santos et al. 2023). Native maize, adapted to diverse environmental conditions and management practices, represents an invaluable reservoir of genetic traits that can contribute to agricultural sustainability and resilience in the face of climate change (González-Martínez et al. 2020). However, current methods for identifying native maize landraces are limited by their reliance on manual morphological analyses, which are often subjective and labor-intensive, thereby constraining the precision, reproducibility, and efficiency of the identification process.

In this context, machine learning (ML) models have proven to be robust and efficient tools for classification and prediction in various areas of biology and ecology (Kuhn and Johnson 2013, Kuhn and Silge 2022, Pichler and Hartig 2023). It is within this framework that the R package *imanr* was developed as an innovative tool for the automated identification of native maize racial complexes using boosted ensemble algorithms. Developed using data from the PGMN, *imanr* employs machine learning techniques to combine the predictive power of multiple weak learners, enabling objective, replicable, and accurate classification. The model was trained using a combination of geographical, morphological, and quantitative descriptors of maize samples, which provide the

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

information such as latitude, longitude, kernel characteristics, cob shape, and row arrangment, allowing the algorithm to capture complex relationships among racial complexes. Boosted ensembles were selected for their ability to optimize model performance by iteratively reducing errors, improving both accuracy and reliability compared to traditional methods.

This article introduces the *imanr* package as an accessible and versatile tool for farmers, scientists, and stakeholders interested in studying and preserving native maize in Mexico. It describes the underlying methodology, model performance, and practical implementation in R, highlighting its ability to integrate qualitative and quantitative traits for identifying racial complexes. Beyond improving the precision and efficiency of native maize racial complex identification, this tool directly promotes strategies for conserving genetic diversity. Moreover, it reinforces the recognition of native maize as an invaluable biocultural heritage linked to traditional agricultural systems that sustain essential ecological and cultural practices for agroecosystem resilience.

Methodology

Dataset: Global Maize Project (PGMN) The development of the *imanr* package was based on data collected by the Global Maize Project (PGMN), managed by the National Institute for Forestry, Agriculture, and Livestock Research (INIFAP), and available through the website of the National Commission for Knowledge and Use of Biodiversity (CONABIO 2011). Two main datasets were used:

- **2010 Database:** This dataset includes 22,932 records with geographic information, producer data, agricultural practices, and detailed morphological measurements of ears and plants.
- **2017 Database:** This dataset comprises 25,861 records containing only georeferenced information, race, and racial complex classification.

Given the larger diversity of variables available, the 2010 database was selected as the primary source of training for this model.

Data Selection and ProcessingA panel of three experts independently evaluated the available variables to identify those most relevant for ensuring quality and consistency in the classification of racial complexes. The following criteria were established:

1. **Proportion of complete records:** Variables with a high percentage of missing data, such as the number of ears per plant or grain moisture content, were excluded.
2. **Direct relevance:** Variables related to descriptors or cultivation practices that did not provide inherent information about maize characteristics were discarded.

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

3. **Format consistency:** Variables such as planting and flowering dates were eliminated due to inconsistencies in data entry.

The variables considered for inclusion were those with at least 35% complete records, as shown in Fig. 1. The heatmap visually represents the proportion of missing data across all variables, grouped by racial complex. Variables with a high density of missing data (evidenced by the dominance of brighter tones) were systematically excluded, while those with more complete records (represented with darker tones) were retained for further analysis. This visualization underscores the heterogeneity in data completeness across both variables and racial complexes, providing a strong justification for the selection thresholds applied.

The final set of selected variables included a mix of geographic, qualitative, and quantitative data, as detailed in Table 1. These variables were chosen for their potential to provide robust and relevant information for classifying maize racial complexes.

To ensure compatibility with classification algorithms, qualitative variables were numerically encoded, while quantitative variables were normalized to prevent biases arising from differences in scale. Missing values were handled through mean imputation for quantitative variables and mode imputation for qualitative variables, ensuring consistency and minimizing the impact of incomplete data.

Classification Algorithms Tested

The R programming language, widely used in bioinformatics and data analysis, was chosen as the platform for testing and implementing the models. Seven supervised learning algorithms, each chosen for its distinct strengths in handling classification tasks and complex datasets, were compared with a comprehensive evaluation of their performance. These models include:

1. **Naive Bayes (NB):** Known for its simplicity and efficiency, this probabilistic classifier assumes independence among predictors, making it computationally fast and effective for datasets with well-separated classes.

2. **K-Nearest Neighbors (KNN):** Selected for its ability to identify patterns in datasets with distinct groups, and adapting well to non-linear relationships.

3. **Support Vector Machines (SVM):** Ideal for classifying high-dimensional data, SVM maximizes the separation margins between classes in complex scenarios.

4. **Random Forests (RF):** A robust ensemble that combines predictions from multiple decision trees, RF is highly resistance to overfitting, and handles multicollinear and high-dimensional data efficiently.

5. **Boosted Ensembles (BE):** This algorithm iteratively enhances model performance by focusing on correcting misclassified instances, achieving superior accuracy and generalization capabilities compared to other methods.

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

6. **Neural Networks (NN):** Designed to model complex, non-linear relationships, making them suitable for detecting highly sophisticated classification patterns within large datasets.

7. **Bagged Trees (BT):** As an ensemble method, bagged trees reduce variance by averaging predictions from multiple bootstrapped datasets, enhancing stability and robustness in the classification process.

The functionality of the *tidymodels* package (Kuhn and Wickham 2024) was utilized for training and comparing the performance of these models. All models were tuned within a structured workflow, applying consistent preprocessing steps to ensure comparability. Cross-validation was employed to identify the best parameters for each model, using metrics such as Balanced Accuracy, F1-score, Cohen's Kappa, and the Area Under the Precision-Recall Curve (AUPRC). These optimized parameters were then used to train the final models, providing a robust basis for comparing their suitability in the classification tasks.

Model Training and Validation

The dataset was randomly split into 75% for training and 25% for testing. No separate validation set was used, as five-fold cross-validation was performed on the training set to optimize hyperparameters before final model evaluation on the test set. This method ensures robust hyperparameter tuning while maximizing the amount of data available for final testing and performance evaluation.

Different performance metrics were prioritized during the tuning process to account for the characteristics of the dataset and the classification goals. Models like RF and NB were tuned using Balanced Accuracy to optimize the trade-off between sensitivity and specificity across classes. Conversely, models such as SVM, BE, and NN were tuned using the Area Under the Precision-Recall Curve (AUPRC) to enhance performance in detecting minority classes, given its importance in this study. Additionally, KNN was tuned using the F1-score, as it effectively combines precision and recall, making it well-suited for evaluating the model's handling of imbalanced data. Finally, BT was tuned using Cohen's Kappa, which measures the agreement between predicted and actual classes while accounting for overall performance and class-specific errors. This tailored approach ensured that each model was optimized for its specific strengths and the requirements of the dataset.

After training and testing the models, their performance was evaluated using metrics including Balanced Accuracy, F1-score, Cohen's Kappa, and AUPRC. Fig. 2 and Table 2 highlight the BE model as the best performer, achieving the highest Balanced Accuracy (0.903), F1-score (0.843), Cohen's Kappa (0.835), and AUPRC (0.912). In contrast, the NB model performed poorly across all metrics, with the lowest Balanced Accuracy (0.534) and F1-score (0.173), reflecting its limitations in handling high-dimensional and imbalanced data.

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

RF and BT showed a high performance, closely behind BE, with RF achieving a Balanced Accuracy of 0.893 and AUPRC of 0.916. KNN and SVM also showed strong performance, particularly in Balanced Accuracy (0.876 and 0.859, respectively). NN performed moderately well with an AUPRC of 0.799 but lagged in Cohen's Kappa (0.752). These results underscore the robustness across metrics of BE, making it the most reliable model for classifying native maize racial complexes.

Implementation in R and Package Design

The *imanr* package was developed using the Boosted Ensemble (BE) algorithm as the primary model, based on its superior performance in the comparative analysis of metrics, as well as its inherent ability to handle complex, high-dimensional datasets with multiple features. This approach is well-known for its ability to iteratively enhance model performance by focusing on misclassified instances, resulting in superior accuracy, robustness and generalization (Chen and Guestrin 2016). The BE methodology combines predictions from multiple weak learners, typically decision trees, to produce a strong predictive model, making it particularly suitable for classifying native maize racial complexes, where data imbalance and feature interactions play a critical role (Natekin and Knoll 2013). Model parameters, such as the number of iterations, learning rate, and maximum tree depth, were optimized through a five-fold cross-validation procedure to maximize accuracy while minimizing overfitting and model variance (Probst et al. 2019).

Regarding the package design, the *xgboost* library was used to implement the model, while the *tidymodels* framework was employed for preprocessing, metrics evaluation, training and hyperparameter tuning. The structure of *imanr* includes a main function that accepts morphological and geographical data as input and returns the classification of the racial complex. Additionally, the package incorporates an imputation function that can help the user in the case of missing data. This design makes imanr a versatile and accessible tool for users with different levels of programming and statistical analysis experience.

External Validation and Reproducibility

External validation was conducted using individual maize samples that were collected in a project to document the native maize from the Otomi-Tepehua region. The *imanr* package is publicly available on CRAN at https://cran.r-project.org/package=imanr, accompanied by detailed documentation and reproducible examples to ensure its accessibility and usability for the scientific community and other users interested in native maize classification. The developing version of the package is available on GitHub at https://github.com/rafa6174/imanr.

Development of an Interactive Interface

To enhance the accessibility of the *imanr* package for users without programming expertise, an interactive graphical interface was developed using ShinyApps. This platform, available at https://arturosp.shinyapps.io/imanrWeb/, allows users to upload morphological and geographical data, perform classifications, and intuitively visualize

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

results. The interface democratizes the use of the package by eliminating the need for advanced knowledge of R, making the tool more inclusive and efficient for researchers, students, and technicians in the agricultural field. Fig. 3 presents a screenshot of the Shiny interface, illustrating the user-friendly design and the classification output, which includes the estimated racial complex and reference images to facilitate interpretation.

Results

The BE model implemented with the *xgboost* package (Chen and Guestrin 2016) demonstrated the best performance among the evaluated methods for identifying native maize racial complexes. The detailed results of the model tuning are presented in Fig. 4 and Table 3, highlighting an Area Under the Precision-Recall Curve of 0.914 and a Balanced Accuracy of up to 90.4% with a *tree_depth* value of 10.

Table 3 summarizes model performance across varying values of key hyperparameters, including *tree_depth, learn_rate,* and *sample_size.* For *tree_depth*, AUPRC and Balanced Accuracy show an upward trend as the depth increases from 1 to 5. Beyond this point, the performance plateaus, suggesting that deeper trees provide diminishing returns due to potential overfitting. Similarly, Fig. 4 illustrates the relationships between AUPRC and the hyperparameters, revealing that increasing *learn_rate*, *sample_size,* and *tree_depth* up to values of 0.057, 0.516, and 10, respectively, are the optimal parameters for training the final model.

The relationship between the hyperparameters and the model's performance is further illustrated in Fig. 4. The graph reveals that increasing *learn_rate* upt to approximately 0.06 results in significant improvements, after which the performance stabilizes. Similarly, *sample_size* values around 0.5 upt to 0.75 achieve optimal AUPRC, with lower and higher values reducing performance. Lastly, the *tree_depth* parameter strongly impacts performance, with the most significant gains observed as depth increases to 5 and stabilizes around 10.

The BE model offers high accuracy and computational efficiency, delivering reliable and reproducible outcomes within seconds to minutes, depending on dataset size. This level of efficiency is particularly notable when compared to traditional identification methods, which rely on expert agronomists, require extended evaluation periods, and often involve significant consultancy fees. By providing a scalable and accessible solution, the BE model implemented in the *imanr* package represents a transformative approach to classifying native maize racial complexes.

Discussion

The results obtained highlight the effectiveness of the proposed model, which represents an innovative tool for conserving the genetic diversity of native maize. The performance of the BE model, with an AUPRC exceeding 0.9, positions it as a robust solution for the

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

identification of racial complexes, significantly surpassing the limitations of traditional methods, such as the subjectivity and variability inherent in manual morphometric analyses (Vega-Álvarez et al. 2022).

Previous studies have demonstrated that BE methods, such as Gradient Boosting Machines (GBM) and XGBoost, are powerful classification techniques in conservation biology due to their ability to handle complex, high-dimensional datasets and their iterative improvement mechanism. For instance, (Ghafarian et al. 2022) discuss the superior classification accuracy of BE methods in ecological modeling, highlighting their adaptability to imbalanced data. Similarly, Ren et al. (2020)emphasize the effectiveness of BE in pollution distribution modeling, where the iterative nature of the algorithm is particularly beneficial for preventing overfitting and creating an efficient routine. In another study, Wieland et al. (2021) explore the application of BE methods in habitat suitability modeling, demonstrating their ability to integrate diverse ecological variables while maintaining robust predictive performance. These findings align with the results presented in this study, showcasing BE as a highly suitable approach for complex biological datasets, such as the classification of native maize racial complexes.

Despite the positive results, it is important to consider the model's limitations. One of the primary restrictions is its scope, which is currently limited to identifying racial complexes. This leaves open the possibility of expanding its capabilities to classify primary races. Such an enhancement would have a significant impact on census and genetic diversity monitoring programs, aligning with global initiatives for the preservation of native crops (González-Martínez et al. 2020).

The accessibility of the *imanr* package has been enhanced through the development of an interactive graphical interface using ShinyApps. This platform, available at https://arturosp.shinyapps.io/imanrWeb/, enables researchers and professionals without advanced programming knowledge to use the tool effectively. The interface allows users to upload morphological and geographical data, perform classifications, and visualize results intuitively. By providing a user-friendly and accessible solution, the ShinyApp democratizes the use of the *imanr* package, fostering its adoption by a broader audience, including researchers, students, and technicians in the agricultural field. This development aligns with studies that emphasize the importance of intuitive design in maximizing the impact of technological innovations (Menegidio et al. 2019, Nishizawa et al. 2020).

Finally, this project represents a unique advancement as the first to automate the identification of native maize racial complexes in Mexico using an efficient and consistent computational approach. This achievement is a step toward modernizing tools for conserving native crops, contributing to agricultural sustainability and food security in Mexico and other maize-producing regions.

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

Conclusion

This study introduces *imanr*, a BE-based package designed for the automatic classification of native maize racial complexes in Mexico. With a balanced accuracy of 90% and a Cohen's Kappa coefficient of 0.83, *imanr* stands out for its reliability, efficiency, and ability to overcome the limitations of traditional methods, offering an innovative and reproducible approach to the morphological and geographical identification of native maize.

The tool represents a significant advancement in modernizing the management and conservation of maize biodiversity, providing an accessible technological solution for researchers and specialists. While *imanr* is a robust model, future enhancements, such as integrating genomic data and optimizing its graphical interface, could expand its scope and utility.

Web location (URIs) and repository

The *imanr* package is publicly available through the following repositories and platforms:

- **CRAN Repository:** https://cran.r-project.org/package=imanr
    - Official stable version of the *imanr* package for installation and use.
- **GitHub Repository (Development Version):** https://github.com/rafa6174/imanr
    - Latest development version, including experimental features and updates.
- **Shiny Web Application:** https://arturosp.shinyapps.io/imanrWeb/
    - Interactive interface for classifying native maize racial complexes using *imanr*.

These resources provide access to the package's source code, documentation, and interactive functionality.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Caballero-García M, Córdova-Téllez L, López- Herrera A (2019) Validación empírica de la teoría multicéntrica del origen y diversidad del Maíz en México. Revista fitotecnia mexicana 42 (4): 357-366. [In ISSN:0187-7380].

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI: https://doi.org/10.3897/arphapreprints.e149212

- Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [ISBN 978-1-4503-4232-2]. https://doi.org/10.1145/2939672.2939785
- CONABIO (2011) Recopilación, generación, actualización y análisis de información acerca de la diversidad genética de maíces y sus parientes y sus parientes silvestres en México. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. Proyecto Global de Maíces Nativos.
- Ghafarian F, Wieland R, Lüttschwager D, Nendel C (2022) Application of extreme gradient boosting and Shapley Additive explanations to predict temperature regimes inside forests from standard open-field meteorological data. Environmental Modelling & Software 156 (105466): 1-11. https://doi.org/10.1016/j.envsoft.2022.105466
- González González M, Medellin S, Heiras Rodríguez C, Lazcarro I (2023) Mundos de maíz en México. Primera Edición, 1. Ediciones Navarra, México, 494 pp. [In Sp]. URL: https://patrimoniobiocultural.com/producto/mundos-de-maiz-en-mexico/ [ISBN 978-607-8789-84-9]
- González-Martínez J, Rocandio-Rodríguez M, Contreras-Toledo A, Joaquín-Cancino S, Vanoye-Eligio V, Chacón-Hernández J, Hernández-Bautista A (2020) Diversidad morfológica y agronómica de maíces nativos del altiplano de Tamaulipas, México. Revista fitotecnia mexicana 43 (4): 361-370. https://doi.org/10.35196/rfm.2020.4.361
- González-Santos R, Hernández-Sandoval L, Hernández-Puente KN, Ortega-Paczka R, González-Santos R, Hernández-Sandoval L, Hernández-Puente KN, Ortega-Paczka R (2023) Distribución y caracterización ecogeográfica de maíces nativos de Querétaro, México. Revista fitotecnia mexicana 46 (4): 341-348. https://doi.org/10.35196/rfm.2023.4.341
- Gouttefanjat F (2020) El maíz como fuerza productiva civilizatoria: ecología y comunidad en Mesoamérica / Corn as a civilizing productive force: ecology and community in Mesoamerica. Pacha. Revista de Estudios Contemporáneos del Sur Global 1 (3): 51-63. https://doi.org/10.46652/pacha.v1i3.43
- Kuhn M, Johnson K (2013) Applied Predictive Modeling. 1. Springer New York, New York, NY, XIII, 600 pp. [In Sp]. [ISBN 978-1-4614-6848-6] https://doi.org/10.1007/978-1-4614-6849-3
- Kuhn M, Silge J (2022) Tidy Modeling with R: A Framework for Modeling in the Tidyverse. 1. O'Reilly Media, United States of America, 367 pp. [In Sp]. URL: https://www.tmwr.org/ [ISBN 978-1-4920-9648-1]
- Kuhn M, Wickham H (2024) tidymodels: Easily Install and Load the 'Tidymodels' Packages. posit Copyriht holder, funder. e MIT + file LICENSE. https://doi.org/10.32614/CRAN.package.tidymodels
- Menegidio FB, Aciole Barbosa D, Gonçalves RDS, Nishime MM, Jabes DL, Costa de Oliveira R, Nunes LR (2019) Bioportainer Workbench: a versatile and user-friendly system that integrates implementation, management, and use of bioinformatics resources in Docker environments. GigaScience 8 (4): 1-9. https://doi.org/10.1093/gigascience/giz041
- Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Frontiers in Neurorobotics 7 (21): 1-21. https://doi.org/10.3389/fnbot.2013.00021
- Nishizawa G, Liu J, Diaz Y, Dmello A, Zhong W, Zanibbi R (2020) MathSeer: A Math-Aware Search Interface with Intuitive Formula Editing, Reuse, and Lookup. Advances in

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI: https://doi.org/10.3897/arphapreprints.e149212

Information Retrieval. [ISBN 978-3-030-45442-5]. https://doi.org/10.1007/978-3-030-45442-5_60

- Pichler M, Hartig F (2023) Machine learning and deep learning—A review for ecologists. Methods in Ecology and Evolution 14 (4): 994-1016. https://doi.org/10.1111/2041-210X.14061
- Probst P, Wright M, Boulesteix A (2019) Hyperparameters and tuning strategies for random forest. WIREs Data Mining and Knowledge Discovery 9 (3): 1-19. https://doi.org/10.1002/widm.1301
- Ren X, Mi Z, Georgopoulos P (2020) Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. Environment International 142 (105827): 1-13. https://doi.org/10.1016/j.envint.2020.105827
- Reza-Solis IJ, Romero-Rosales T, Galeno CdÁH, Lagarda JLV, Lobato VJ (2024) Saberes tradicionales en el cultivo de maíces nativos. Revista Biológico Agropecuaria Tuxpan 12 (1): 167-178. https://doi.org/10.47808/revistabioagro.v12i1.551
- Sánchez-González JdJ (2011) Diversidad del Maíz y el Teocintle. Informe preparado para el proyecto: "Recopilación, generación, actualización y análisis de información acerca de la diversidad genética de maíces y sus parientes silvestres en México. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. Proyecto Global de Maíces Nativos.
- Vega-Álvarez I, Flores-Sánchez D, Escalona-Maurice MJ, Castillo-González F, Jiménez-Velázquez MA (2022) Tlaxcala, investigación en maíz nativo y mejorado: problemática, campos del conocimiento y nuevos retos. Revista mexicana de ciencias agrícolas 13 (3): 539-551. https://doi.org/10.29312/remexca.v13i3.2888
- Wieland R, Kuhls K, Lentz HK, Conraths F, Kampen H, Werner D (2021) Combined climate and regional mosquito habitat model based on machine learning. Ecological Modelling 452 (109594): 1-9. https://doi.org/10.1016/j.ecolmodel.2021.109594

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

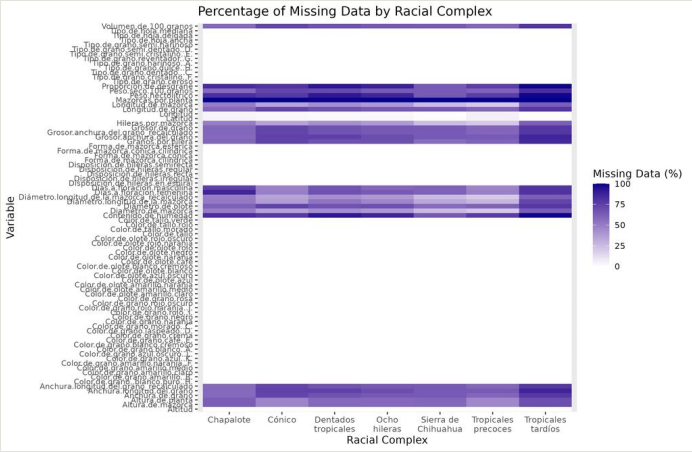

Figure 1.

Proportion of missing data by variable in the 2010 PGMN database.

ARPHA Preprints
Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

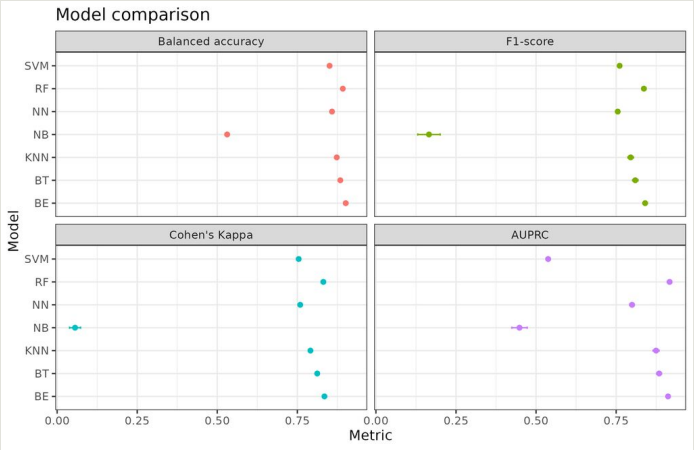

Figure 2.

Comparative of collected metrics for different models tested for the classification of native maize racial complexes.

ARPHA Preprints
Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

Figure 3.

Screenshot of the *imanr* Shiny application interface for native maize racial complex identification.

ARPHA Preprints
Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

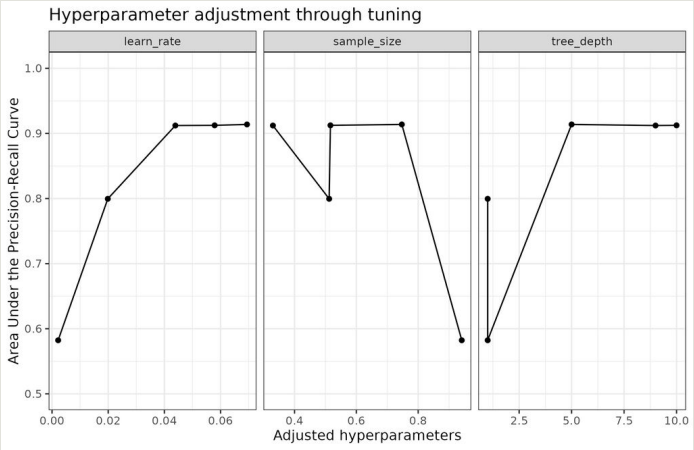

Figure 4.

Impact of hyperparameter tuning on model performance for the Boosted Ensemble model.

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

Table 1.
Selected variables for model training and application.

Category	**Variable**
Identification	ID, primary race, racial complex
Geographic	Latitude, longitude, altitude
Qualitative	Grain color, cob color, stem color, row arrangement, ear shape, grain type
Quantitative	Ear length, ear diameter, ear diameter-to-length ratio, rows per ear

Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

Table 2.

Performance comparison of classification models based on Balanced Accuracy, F1-score, Cohen's Kappa, and AUPRC.

Model	Balanced accuracy	F1-score	Cohen's Kappa	AUPRC
BE	0.904	0.846	0.838	0.914
RF	0.893	0.837	0.832	0.916
BT	0.887	0.813	0.813	0.887
KNN	0.863	0.792	0.770	0.847
SVM	0.858	0.771	0.693	0.558
NN	0.853	0.747	0.752	0.799
NB	0.535	0.173	0.062	0.458

ARPHA Preprints
Author-formatted, not peer-reviewed document posted on 11/02/2025. DOI:
https://doi.org/10.3897/arphapreprints.e149212

Table 3.

Hyperparameter tuning results for the Boosted Ensemble model.

tree_depth	learn_rate	sample_size	Balanced accuracy	F1-score	Cohen's Kappa	AUPRC
1	0.0021	0.9412	0.6749	0.5828	0.5748	0.5827
1	0.0198	0.5119	0.8073	0.6817	0.7243	0.8005
5	0.0694	0.7472	0.9043	0.8456	0.8378	0.9137
9	0.0439	0.3298	0.9024	0.8434	0.8368	0.9126
10	0.0578	0.5160	0.9036	0.8451	0.8369	0.9138