Agronomy Research **22**(2), 887–897, 2024
https://doi.org/10.15159/AR.24.063

Genotype prediction in maize (*Zea mays* L.) progeny using different predictive models

A. Polyvanyi[1,*], A. Butenko[2], M. Mikulina[1], V. Zubko[1], S. Kharchenko[3], V. Dubovyk[4], O. Dubovyk[4] and B. Sarzhanov[1]

[1]Sumy National Agrarian University, Faculty of engineering and technology, Department of Agroengineering, H. Kondratieva Str. 160, UA40021 Sumy, Ukraine
[2]Sumy National Agrarian University, Faculty of agrotechnologies and natural resource management, Plant growing Department, H. Kondratieva Srt. 160, UA40021 Sumy, Ukraine
[3]Sumy National Agrarian University, Faculty of agrotechnologies and natural resource management, Department of Physical Education, H. Kondratieva Str. 160, UA40021 Sumy, Ukraine
[4]Sumy National Agrarian University, Faculty of agrotechnologies and natural resource management, Department of Biotechnology and chemistry, H. Kondratieva Str. 160, UA40021 Sumy, Ukraine
[*]Correspondence: polivanui1@gmail.com

Received: March 29th, 2024; Accepted: June 27th, 2024; Published: July 11th, 2024

Abstract. This study utilized two probabilistic methods, Gaussian Naive Bayes (GNB) and Logistic Regression (LR), to predict the genotypes of the offspring of two maize varieties: SC604 and KSC707, based on the phenotypic traits of the parent plant. The predictive performance of both models was evaluated by measuring their overall accuracy and calculating the area under receiver operating characteristic curve (AUC). The overall accuracy of both models ranged from 80% to 89%. The AUC values for the LR models were 0.88 or higher, while the GNB models had AUC values of 0.83 or higher. These results indicated that both models were successful in predicting the genetic makeup of the progeny. Furthermore, it was observed that both models were more accurate in predicting the SC604 genotype, which was found to be more consistent and predictable compared to the KSC707 genotype. A chi-square test was conducted to assess the similarity between the prediction results of the two models, revealing that both models had a similarly high likelihood of making accurate predictions in all scenarios.

Key words: Gaussian Naive Bayes (GNB), genotype prediction, Logistic Regression (LR), predictive models, *Zea mays* L.

INTRODUCTION

The implementation and development of learning algorithms have paved the way for novel approaches to data processing, information extraction, and decision-making. These algorithms have widespread applications across various fields of human activity, as evidenced by the abundance of literature on their development and use. Scientists

have devised an algorithm for the selection of genetic biomarkers and classification of subjects through the analysis of genome-wide single nucleotide polymorphism data. The algorithm demonstrated significantly improved classification accuracy in identifying biomarkers for Type 1 diabetes (Sambo et al., 2012).

Other researchers have used a naive Bayes classifier combined with small RNA deep sequencing statistics, as well as genomic signatures to recognize offspring microRNAs in different types of crops and herbs (Douglass et al., 2016). The classifier exhibited high accuracy in identifying microRNA for all four plants, as determined by the AUC's receiver operating characteristic curve. The values of the area under the receiver operating curve reported in the study range from 0.9750 to 0.9960, with the highest value being obtained for the identification of Arabidopsis microRNAs. These machine learning algorithms have been widely employed in the classification of various plant organisms as well. Typically, the identification of plants is based on leaf characteristics, such as their shape, color, and texture, which are unique to each plant species. Unlike flowers and fruits, plants maintain functional leaves throughout their lifespan (Siravenha & Carvalho, 2015).

Another group of scientists from China and the US created a probabilistic series of algorithms that endeavors to recognize underlying relationships in a set of data, employing image and data processing techniques to automatize herb classifying (Wu et al., 2007). This algorithm successfully classified 32 different herb species based on their visual characteristics with a greater than 90% accuracy rate. The authors found their algorithm to be both rapid and efficient in recognizing and classifying herbs.

In different years, many scientists were engaged in the development of an algorithm for the visual recognition of plants. One such research group, for example, created an algorithm that can differentiate weeds from two main crops - carrots and cabbage (Hemming & Rath, 2001). They initially used eight morphological features in combination with three color features, selecting the most relevant features to discriminate between the different plant species. The inclusion of color features improved classification accuracy, with the researchers reporting between 50% and 95% of crops being accurately classified. The average rate of successful classifying was 87% and 73% for cabbage and carrots, respectively.

Continuing research on this topic, another group of scientists from the University of Southern Denmark created a dataset representing visual features using distance data from images of seedlings of various plants (Giselsson et al., 2013). The researchers utilized a high-degree Legendre polynomial to fit the distance data and subsequently extracted the coefficients of the polynomial, which they referred to as the Legendre polynomial feature set. In addition, they collected another set of data, which they named the standard feature set. The data sets were then subjected to four classification algorithms. The researchers observed that the Legendre polynomial feature set exhibited a high degree of robustness, generating an accuracy rate of nearly 99% when used in conjunction with the classification algorithms. Conversely, the standard feature set yielded an accuracy rate of 87%. Despite these promising results, the researchers recommended further testing of the attribute data collection method to determine its true value.

A new method for data acquisition on leaf features has also been proposed previously (Siravenha & Carvalho, 2015; Änäkkälä et al., 2023; Esan et al., 2023). It includes contour-to-centroid distance and transforms the data using a fast Fourier transform. The authors of this method used a feature selection approach to reduce

dimensionality, which led to increased classification accuracy. Also, they employed various classification algorithms for plant identification, achieving classification accuracies ranging from 66% (with the C4.5 algorithm) to 98% (with the Pattern Net algorithm applied to main components). The field of phenotypic predictions for quantitative traits of sequenced whole genomes has seen significant advancement with the incorporation of learning algorithms and other statistical methods. Scientists from Fondazione Bruno Kessler, Italy, made a significant contribution to the development of this field with their research (Guzzetta et al., 2010). They outlined a learning process that uses a naive elastic network-based L1L2 regularization method to predict phenotypes. The effectiveness of this method was found to be highly accurate. Similarly, a group of Chinese scientists employed machine learning techniques to distinguish root traits that caused cultivar differentiation in a binomial environment (Zhao et al., 2016).

Lippert et al. (2017) and Dragov et al. (2023) developed models for phenotypic feature prediction and combined them into a singular machine learning model for genome re-identification. In all instances, various learning methods were utilized, each with satisfactory to excellent performance in prediction or classification. It should be noted, however, that no single machine-learning method is optimal for every circumstance, as each has its strengths and limitations (Kell et al., 2001; Hudzenko et al., 2023).

The exceptional proficiency of the algorithms in recognizing and grouping the subjects being examined led us to undertake a study that aims to forecast the genotypes of plants by examining the phenotypic traits of their parents. Furthermore, this study serves as a quantitative confirmation of the theory that GNB with continuous characteristics and LR are fundamentally identical, despite being generative and discriminative respectively (Ng & Jordan, 2001; Chen et al., 2019).

The present study incorporates a comprehensive analysis of the materials and methods used in its execution. This involved the procurement of all necessary materials, including instruments, equipment, and chemicals, from reputable sources. The methods employed in this study were carefully formulated and optimized to obtain the most accurate and reliable data possible. The techniques used were based on established protocols and were performed under standardized conditions to minimize experimental variability and ensure consistency. All procedures were documented in detail to facilitate reproducibility and ease of replication.

MATERIALS AND METHODS

Location of the Study

In May 2020, an experiment was conducted in the experimental field of Sumy National Agrarian University, located in the Sumy district of the Sumy region, which is part of the Forest-steppe zone of Ukraine. The soils on the experimental site were dark gray podzolized and fertilizers were added to the soil.

Plant Material

For this research, two genotypes of maize (*Zea mays* L.) were used: SC604 and KSC707. They differed in the color of their kernels, with SC604 having translucent, white kernels, and KSC707 having yellow kernels.

Selection Process

The SC604 genotype underwent a selection process spanning three cycles, to develop larger, wider leaves, an increased number of leaves, and taller plant structures. In contrast, the KSC707 genotype underwent random selection after each cycle.

Pollination Process

To obtain seeds from selected plants of each strain, pollination was carried out using bulk pollen harvested from male inflorescences of the same genotype. Controlled pollination was ensured by protecting the ear with a paper bag immediately after emergence.

Field Research Methods

The experiment involved planting seeds of fourth-generation parent plants and fifth-generation offspring plants of the two genotypes mentioned. It was conducted using a completely randomized design with two factors - genotype and generation. The experimental field was divided into two rows for each variety, with each row being 5 meters long, and separated by a distance of 0.9 meters. The plots were overseeded, and after three weeks of planting, the number of plants in each plot was thinned to 40,000 units per hectare. During silking, when almost half of the planted plants in the field developed silks, in order to obtain accurate data, we employed a series of measurements to determine various plant attributes for accurate data collection. We used a node-counting method to determine the number of leaves present (nl). Additionally, we measured the length and width of the ear leaf (el, ew), and the overall height of the experimental samples (hs), defined as the distance from the soil base to the apex of the m-inflorescence. To assess grain filling rate (gf), we used a linear coefficient derived from the orthogonal contrast of grain dry weight during the linear phase of the filling time, as part of a sequential selection process.

Prediction Methods

To predict the genotypes of the offspring, we utilized two classification techniques: the GNB and LR. The GNB algorithm was used to determine the probability of a progeny belonging to either the SC604 or KSC707 strain, based on attributes such as the number of leaves, length and width of the ear leaf and others. The probability model was simplified as follows:

$$\Pr(Class|P_1, P_2, \dots, P_5) = \Pr(Class) \times \prod_{i=1}^{5} \Pr(P_i|Class) \quad (1)$$

In this model, Class represents the class label (SC604 or KSC707) and P_1 is used to represent the attributes. The model calculates the posterior probability of a progeny's lineage by multiplying the prior probability of the class label with the conditional probabilities of the attributes associated with that class label. It is important to note that this model falls under the discriminative category for prediction or classification, unlike LR, which is generative.

To determine whether a progeny belongs to the KSC707 or SC604 genotypes, we used measurements from the parents of these genotypes to fit an LR model, following the methodology outlined by Tsangaratos & Ilia (2016). The model parameters were

estimated to compute the posterior probabilities of the progeny being classified into either the KSC707 or SC604 genotypes.

$$\Pr(Class|P_1, P_2, \dots, P_5) \quad (2)$$

The LR model assumes a parametric structure when the class variable is binomial, as shown by Ng & Jordan (2001):

$$\Pr(Class = "SC604"|P_1, P_2, \dots, P_5) = \frac{1}{1 + \exp(b_o + \sum_{i=1}^{5} b_i P_i)} \quad (3)$$

and

$$\Pr(Class = "KSC707"|P_1, P_2, \dots, P_5) = \frac{\exp(b_o + \sum_{i=1}^{5} b_i P_i)}{1 + \exp(b_o + \sum_{i=1}^{5} b_i P_i)} \quad (4)$$

Ng & Jordan (2001) demonstrated that the mathematical structure of Pr(Class|P_1, P_2,… P_5) employed by LR conforms exactly to the structure of a GNB classifier under the assumptions made. Furthermore, in the majority of cases, both methods yield comparable outcomes. It is pertinent to note that LR estimates the parameters of Pr(Class|P_1, P_2,… P_5) directly, whereas GNB estimates the parameters of Pr(Class) and Pr(P_1, P_2,… P_5|Class) directly.

In order to evaluate the proposition, that GNB and LR models yield identical outcomes (Ng & Jordan, 2001), we conducted a series of experiments. First, we created several sub-samples from the original dataset by iteratively excluding some attributes and then computed their main components. Next, we used both models to predict the offspring based on the conserved attributes or the respective main components of each dataset. The main components, which are an orthogonal conversion of the initial observations that maintain the total variance, are expected to eliminate any bias associated with the assumption of independence between attributes, compared to using the original variables. This approach is intended to help identify any such bias in the prediction results. We evaluated the relative significance of the characteristics in distinguishing between the two categories. We calculated several metrics, including precision, sensitivity, the maximum AUC of receiver operating characteristics and others for both models. We assessed each model's efficacy using both the accuracy and the AUC value. In our experiments, the SC604 genotype was defined as the positive class. All analyses were conducted using the R statistical software package (R Core Team, 2017).

RESULTS AND DISCUSSION

Kuhn (2015) in his study utilized the variable significance ranking technique to identify the most pertinent traits among the properties of the data set in distinguishing between the two genotypes. The analysis determined that *ew*, *gf*, and *hs* were the most crucial attributes, with respective significance values of 0.86, 0.83, and 0.74 on a 1-point scale. This conclusion was validated through the LR model, which also identified *ew*, *gf*, and *hs* as the most significant predictor variables in the data set, with p-values of 0.0395, 0.0475, and 0.05, respectively. Means and standard deviations for each of the five traits were calculated and displayed (Table 1), with the *nl* trait exhibiting minimal variability across generations and genotypes and ranking last in its capability to distinguish one genotype from another with a significance value of 0.58.

Table 1. Mean and standard deviation for the five parent-offspring properties considered

Attribute	Parent		Offspring	
	SC604	KSC707	SC604	KSC707
ew (cm)	9.71(0.60)	8.41(1.13)	9.52(0.66)	8.99(0.79)
el (cm)	88.45(6.55)	83.97(7.44)	90.12(5.11)	84.92(6.04)
nl (cm)	12.22(1.57)	12.13(1.08)	12.94(1.03)	12.31(1.21)
hs (cm)	236.73(24.12)	217.32(24.86)	232.10(21.34)	217.06(21.48)
gf ($mg \times d^{-1}$)	5.52(0.31)	5.05(0.35)	5.62(0.29)	5.17(0.39)

By utilizing diverse subsets of the original data set and their main components, we were able to predict the offspring of two parents. In the following, we will present the performance of both models on specific subsets (Table 2 and Table 3).

Table 2. Values of the confusion matrix of the LR and GNB models with subsets of the initial information and their respective main components

Data	Attribute	Model	Confusion Matrix			
			PT	NF	NT	PF
Initial data	(1) *ew*+*el*+*nl*+*hs*+*gf*	LR	27	3	24	6
		GNB	28	2	25	5
	(2) *ew*+*gf*+*hs*+*nl*	LR	28	2	23	7
		GNB	28	2	23	7
	(3) *ew*+*gf*+*hs*	LR	28	2	23	7
		GNB	28	2	24	6
	(4) *gf*+*hs*	LR	27	3	27	3
		GNB	25	5	28	2
	(5) *ew*+*gf*	LR	28	2	24	6
		GNB	27	3	24	6
Main component	$mc_1+mc_2+mc_3+mc_4+mc_5$ from (1)	LR	27	3	24	6
		GNB	27	3	23	7
	$mc_1+mc_2+mc_3+mc_4$ from (2)	LR	28	2	23	7
		GNB	29	1	20	10
	$mc_1+mc_2+mc_3$ from (3)	LR	28	2	23	7
		GNB	28	2	23	7
	mc_1+mc_2 from (4)	LR	27	3	27	3
		GNB	24	6	26	4
	mc_1+mc_2 from (5)	LR	28	2	24	6
		GNB	27	3	23	7

mc_i(i=1, 2,…,5) – main components; PT – positive true; NF – negative false; NT – negative true; and PF – positive false.

To compare the confusion matrices generated by the models on each subset, we conducted a Chi-square test (Agresti, 2007). The p-values resulting from the test of homogeneity of the confusion matrices ranged from 0.7812 to 1, indicating that the prediction performances of both models were very similar. GNB and LR are both models utilized to compute the conditional probability that an offspring belongs to a particular genotype. GNB updates prior knowledge from parents with current evidence of offspring, while LR estimates the parameters of the model with the parents' phenotypic values. Previously, scientists have already provided a comprehensive and refined demonstration

that, under certain conditions of a binary response variable (Y) with parameter π=Pr(Y="PositiveClass"), Gaussian distributed attributes (P_i) that are conditionally independent concerning Y, and $Pr(P_i|Y=y_k)\sim N(\mu_{ik},\sigma_i)$, the conditional probabilities under the GNB model can be expressed in parametric forms (Ng & Jordan, 2001).

$$\Pr(Class = "SC604"|P_1, P_2, \dots, P_5) = \frac{1}{1 + \exp(w_o + \sum_{i=1}^{5} w_i P_i)} \quad (5)$$

and

$$\Pr(Class = "KSC707"|P_1, P_2, \dots, P_5) = \frac{\exp(w_o + \sum_{i=1}^{5} w_i P_i)}{1 + \exp(w_o + \sum_{i=1}^{5} w_i P_i)} \quad (6)$$

where

$$w_i = \frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} \quad (7)$$

and

$$w_0 = \ln\frac{1-\pi}{\pi} + \sum_i \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2} \quad (8)$$

When the response variable has binary outcomes and the predictor variables are distributed according to a Gaussian distribution, the probabilities provided by LR remain identical. The numerical evidence obtained from the experiment supports the findings of other studies (Ng & Jordan, 2001; Bhowmik, 2015; Vasilaki et al., 2023). It is important to note that LR estimates the parameters for Pr(Class|Pi) directly, while GNB estimates the parameters for Pr(Class) and Pr(Pi|Class) directly.

Table 3. The predictive attributes of the LR and GNB models with subsets of the initial information and their respective main components

Data	Attribute	Model	*p*-value	Accuracy	AUC
Initial data	(1) *ew*+*el*+*nl*+*hs*+*gf*	LR	0.9710	0.84	0.88
		GNB		0.86	0.90
	(2) *ew*+*gf*+*hs*+*nl*	LR	1.000	0.83	0.91
		GNB		0.84	0.88
	(3) *ew*+*gf*+*hs*	LR	0.9886	0.85	0.90
		GNB		0.86	0.89
	(4) *gf*+*hs*	LR	0.8904	0.89	0.91
		GNB		0.88	0.90
	(5) *ew*+*gf*	LR	0.9822	0.85	0.89
		GNB		0.84	0.86
Main component	mc_1+mc_2+mc_3+mc_4+mc_5 from (1)	LR	0.9886	0.82	0.89
		GNB		0.83	0.83
	mc_1+mc_2+mc_3+mc_4 from (2)	LR	0.8196	0.85	0.91
		GNB		0.80	0.85
	mc_1+mc_2+mc_3 from (3)	LR	1.000	0.84	0.91
		GNB		0.84	0.88
	mc_1+mc_2 from (4)	LR	0.7812	0.89	0.90
		GNB		0.82	0.88
	mc_1+mc_2 from (5)	LR	0.9696	0.84	0.89
		GNB		0.82	0.88

mc_i(i=1, 2,…,5) – main components.

The findings indicate that logistic regression produced identical prediction outcomes, regardless of whether subsets of the initial data or their main components were utilized (Table 2 and Table 3). This consistency can be attributed to the logistic equation's right-hand side remaining consistent for both the original data and main components,

$$\frac{1}{1+\exp(\hat{\beta}_{obs}^{t}A)} = \frac{1}{1+\exp(\hat{\beta}_{pc}^{t}Z)}$$

which leads to:

$$\hat{\beta}_{obs}^{t}A = \hat{\beta}_{pc}^{t}Z \tag{9}$$

In this case, based on the initial data, $\hat{\beta}_{obs}^{t}$ is the vector of ERC (estimated regression coefficient) while A denotes the matrix composed of the subset of the original variables. On the other hand, the vector of ERC based on the main traits is $\hat{\beta}_{pc}^{t}$, and Z is the matrix of main traits that is derived from the subset of the primary data.

So,

$$\Pr(Class = "SC604"|P, \beta_{obs}) = \Pr(Class = "SC604"|Z, \beta_{pc})$$

and

$$\Pr(Class = "KSC707"|P, \beta_{obs}) = \Pr\big(Class = "KSC707"\big|Z, \beta_{pc}\big).$$

When utilizing GNB, it was observed that the confusion matrices derived from subsets of the original data and their main components were not consistently identical. However, the disparities did not prove to be of significant value, as all *p*-values were found to be greater than or equal to 0.7812.

The prediction performances of both models were highly commendable. When all available data were tested, the overall accuracy of predictions ranged from 80% to 89% (Table 3), with AUC values between 0.83 and 0.91. The models were notably more accurate in predicting SC604 genotypes than KSC707 genotypes, with lower specificity but higher sensitivity. This observation may be attributed to the structures of the SC604 and KSC707 populations and their respective development procedures. The SC604 population was developed using a selection process aimed at promoting wider and longer leaves, with increased leaf numbers and taller plants. The three cycles of selection likely contributed to the partial realization of that objective. The SC604 genotypes have progressively formed a more homogenous group that exhibits the distinctive traits selected for, except for *nl*. As a result, the accuracy of prediction in these models was higher.

To date, scientists have provided a lot of evidence to support the influence of selection on a hereditary trait (García-Ruiz et al., 2016; Kolesnikov et al., 2023). They have demonstrated that selection, whether natural or artificial, influences the expression of the gene(s) responsible for the development of a modified phenotype. A group of Chinese researchers has discovered that the regulation of maize leaf width is governed by dominant genes that are not linked to the gene(s) responsible for the control of maize leaf length (Wang et al., 2018). This discovery suggests that selection can be employed to modify the width of maize leaves without any direct effect on other canopy traits.

The present study confirms the hypothesis that the increase in maize leaf width and plant height of the SC604 genotype can be attributed to the direct impact of selection for wider leaves and taller plants. Consequently, these two traits have emerged as the most significant discriminating features between the two populations. The discrepancy in grain filling rate between the SC604 and KSC707 genotypes can be attributed to an indirect selection for canopy size. Conversely, the KSC707 genotypes were selected at random, resulting in larger trait variability and less homogeneity among individuals. The increased dispersion of the KSC707 strains led to reduced model specificity. Although some individuals of the KSC707 strains shared traits with the SC604 genotypes, the majority maintained their distinct traits and were accurately predicted by the models.

The tests assessing the genotypes of offspring primarily focused on parent-offspring resemblance. In models trained with the phenotypic values of parents, similar traits were identified in their progeny. The receiver operating characteristic's AUC from these tests, which utilized a single predictor variable, may serve as a reliable predictor of heritability. Wray et al. (2010) developed an equation correlating the maximum receiver operating characteristic's AUC with heritability and the prevalence of a given trait. Dreyfuss et al. (2012) noted in their study that the precision of a study reflects the heredity of the trait being evaluated. They added that the high heritability of a phenotypic trait resulted in greater prediction accuracy, while traits with low heritability had lower accuracy predictions, being more susceptible to environmental factors than genetic factors. This study considered numerous phenotypic traits in determining offspring genotypes. However, it was unclear whether the AUC value was a suitable estimator of heritability. In this study, we may consider the AUC value as an adequate indicator for measuring the similarity between parents and their offspring. As the AUC value increases, so does the level of resemblance between the two parties.

CONCLUSIONS

Using field data, we employed two different predictive models, LR and GNB, to determine the genotype of two distinct maize strains. The accuracy of our predictions ranged from 80% to 89%, with the AUC values of the receiver operating characteristic falling between 0.83 and 0.91. Our tests showed a high level of sensitivity, indicating a correct identification of the SC604 genotype, with a modal value of 0.91 for both GNB and LR. Conversely, specificity, defined as the correct identification of the KSC707 genotype, had a modal value of 0.74 for both GNB and LR. This discrepancy in sensitivity and specificity can be attributed to the varying structures of the two populations, rather than the quality of the tests themselves. Specifically, the SC604 population was more homogenous and easier to identify, while the KSC707 population was more diverse, with a large proportion of its progeny being misclassified as SC604 genotypes.

To conduct the genotype-prediction tests, various subsets of the data along with their corresponding main components were utilized. A Chi-square test was performed to compare the prediction results of the LR and GNB models for each data set. The outcome indicated that both models produced similar prediction performances. The results of our field research are consistent with the conclusions of the theoretical works of other scientists (Ng & Jordan, 2001; Bhowmik, 2015). When using LR, there was no difference in prediction results whether the subset of the data or its corresponding main

components were used. The main components preserved the overall fluctuation of the information, and the outcome of the matrix of data multiplied by the ERC's vector from a subset remained consistent with the product of the matrix of primary constituents and the vector of estimated coefficients from those components. When employing the GNB model, the predictive results with a subset of the data or the corresponding primary constituents did not invariably match. However, any differences found were not significant.

REFERENCES

Agresti, A. 2007. *An Introduction to Categorical Data Analysis.* (2nd ed.). John Wiley & Sons, 38 pp.

Änäkkälä, M.A., Lehtilä, A., Mäkelä, P. & Lajunen, A. 2023. Application of UAV multispectral imaging for determining the characteristics of maize vegetation. *Agronomy Research* **21**(2), 644–653.

Bhowmik, T.K. 2015. Naive Bayes vs Logistic Regression: theory, implementation and experimental validation. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial* **18**(56), 14–30. doi: 10.4114/intartif.vol18iss56pp14–30

Chen, W., Yan, X., Zhao, Z., Hong, H., Bui, D.T. & Pradhan, B. 2019. Spatial prediction of landslide susceptibility using data mining–based kernel logistic regression, naive Bayes and RBFNetwork models for the Long County area (China). *Bulletin of Engineering Geology and the Environment* **78**, 247–266. doi: 10.1007/s10064–018–1256–z

Douglass, S., Hsu, S.W., Cokus, S., Goldberg, R.B., Harada, J.J. & Pellegrini, M. 2016. A naïve Bayesian classifier for identifying plant micro RNAs. *The Plant Journal* **86**(6), 481–492. doi: doi.org/10.1111/tpj.13180

Dragov, R., Taneva, K. & Bozhanova, V. 2023. Parametric and nonparametric stability of grain yield and grain protein content in durum wheat genotypes with various origins. *Agronomy Research* **21**(2), 693–710.

Dreyfuss, J.M., Levner, D., Galagan, J.E., Church, G.M. & Ramoni, M.F. 2012. How accurate can genetic predictions be?. *BMC Genomics* **13**(1), 1–8. doi: 10.1186/1471-2164-13-340

Esan, V.I., Sangoyomi, T.E., Ajayi, O.A., Christensen, M. & Ogunwole, J.O. 2023. Performance evaluation and variability analysis for morpho-physiological traits of orange fleshed tomato varieties introduced in Nigeria climatic conditions. *Agronomy Research* **21**(2), 711–727.

García-Ruiz, A., Cole, J.B., VanRaden, P.M., Wiggans, G.R., Ruiz-López, F.J. & Van Tassell, C.P. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proceedings of the National Academy of Sciences* **113**(28), E3995–E4004. doi: 10.1073/pnas.1519061113

Giselsson, T.M., Midtiby, H.S. & Jørgensen, RN. 2013. Seedling discrimination with shape features derived from a distance transform. *Sensors* **13**(5), 5585–5602. doi: 10.3390/s130505585

Guzzetta, G., Jurman, G. & Furlanello, C. 2010. A machine learning pipeline for quantitative phenotype prediction from genotype data. *BMC Bioinformatics* **11**(8), 1–9. doi: 10.1186/1471-2105-11-S8-S3

Hemming, J. & Rath, T. 2001. PA—Precision agriculture: Computer-vision-based weed identification under field conditions using controlled lighting. *Journal of Agricultural Engineering Research* **78**(3), 233–243. doi: 10.1006/jaer.2000.0639

Hudzenko, V.M., Lysenko, A.A., Tsentylo, L.V., Demydov, O.A., Polishchuk, T.P., Khudolii, L.V., ... & Kozelets, H.M. 2023. Genotype by yield x trait (GYT) biplot analysis for the identification of the superior winter and facultative barley breeding lines. *Agronomy Research* **21**(2), 739–757.

Kell, D.B., Darby, R.M. & Draper, J. 2001. Genomic computing. Explanatory analysis of plant expression profiling data using machine learning. *Plant Physiology* **126**(3), 943–951. doi: 10.1104/pp.126.3.943

Kolesnikov, M., Gerasko, T., Paschenko, Y., Pokoptseva, L., Onyschenko, O. & Kolesnikova, A. 2023. Effect of water deficit on maize seeds (Zea mays L.) during germination. *Agronomy Research* **21**(1), 156–174.

Kuhn, M. 2015. Caret: classification and regression training. *Astrophysics Source Code Library*, ascl–1505.

Lippert, C., Sabatini, R., Maher, M.C., Kang, E.Y., Lee, S., Arikan, O., Harley, A., Bernal, A., Garst, P., Lavrenko, V., Yocum, K., Wong, T., Zhu, M., Yang, W.-Y., Chang, C., Lu, T., Lee, C.W.H., Hicks, B., Ramakrishnan, S., …, & Venter, J.C. 2017. Identification of individuals by trait prediction using whole-genome sequencing data. *Proceedings of the National Academy of Sciences* **114**(38), 10166–10171. doi: 10.1073/pnas.1711125114

Ng, A. & Jordan, M. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in neural information processing systems* **14**, 841–848.

R Core Team. 2021. R: A language and environment for statistical computing. *R Foundation for Statistical Computing* **14**, 841–848.

Sambo, F., Trifoglio, E., Di Camillo, B., Toffolo, G.M. & Cobelli, C. 2012. Bag of Naïve Bayes: biomarker selection and classification from genome-wide SNP data. *BMC Bioinformatics* **13**(14), 1–10. doi: 10.1186/1471–2105–13–S14–S2

Siravenha, A.C.Q. & Carvalho, S.R. 2015. Exploring the use of leaf shape frequencies for plant classification. In *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 297–304. IEEE. doi: 10.1109/SIBGRAPI.2015.36

Tsangaratos, P. & Ilia, I. 2016. Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. *Catena* **145**, 164–179. doi: 10.1016/j.catena.2016.06.004

Vasilaki, C., Katsileros, A., Doulfi, D., Karamanos, A. & Economou, G. 2023. Evaluation of seven barley genotypes under water stress conditions. *Agronomy Research* **21**(1), 222–238.

Wang, B., Zhu, Y., Zhu, J., Liu, Z., Liu, H., Dong, X., Guo, J., Li, W., Chen, J., Gao, C., Zheng, X., Lizhu, E., Lai, J., Zhao, H. & Song, W. 2018. Identification and fine–mapping of a major maize leaf width QTL in a re-sequenced large recombinant inbred lines population. *Frontiers in Plant Science* **9**, 101. doi: 10.3389/fpls.2018.00101

Wray, N.R., Yang, J., Goddard, M.E. & Visscher, P.M. 2010. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS genetics* **6**(2), e1000864. doi: 10.1371/journal.pgen.1000864

Wu, S.G., Bao, F.S., Xu, E.Y., Wang, Y.X., Chang, Y.F. & Xiang, Q.L. 2007. A leaf recognition algorithm for plant classification using probabilistic neural network. In *2007 IEEE international symposium on signal processing and information technology*, pp. 11–16. IEEE. doi: 10.1109/ISSPIT.2007.4458016

Zhao, J., Bodner, G. & Rewald, B. 2016. Phenotyping: using machine learning for improved pairwise genotype classification based on root traits. *Frontiers in plant science* **7**, 1864. doi: 10.3389/fpls.2016.01864