SELECTION OF COVARIATES IN A LOGISTIC REGRESSION MODEL FOR THE PREDICTION OF RESISTANCE TO RICE BLAST

Momate Emate OSSIFO[1]
Marciel Lelis DUARTE[2]
Antônio Policarpo Souza CARNEIRO[2]
Vinicius Silva dos SANTOS[3]
Sebastião MARTINS FILHO[2]

- ABSTRACT: Rice (*Oryza sativa* L.) has been one of the most consumed foods on the planet, with economic and social importance. Diseases, mainly blast, caused by the fungus *Pyricularia oryzae*, are limiting factors for the production of rice. The present work aimed to select covariables that can influence resistance to rice blast, using the selection strategy proposed by Collett. Logistic regression models were adjusted to predict disease resistance, using the ROC curve to assess the predictive capacity. The data used were obtained from a population of 413 plants, with phenotypic information collected in 82 countries and classified into five subpopulations. The research found that, out of over fifteen variables embedded to assess the disease, only three revealed to be relevant for the final adjusted model, namely: width of flag leaf (V4), the mean number of primary panicle branches (V8) and the amount of amylose from ground grains (V15). The variable V4 presented the most significant influence on disease resistance. Additionally, for each unit increase in V4, V8 and V15, it is expected to obtain 279.3, 31.9 and 9.4% increases, respectively, in the probability of resistance to rice blast.

- KEYWORDS: ROC curve; Collet method; *Oryza sativa*; *Pyricularia oryzae*.

1 Introduction

Throughout history, rice (*Oryza sativa* L.) has been one of the most consumed foods on the planet, presenting fundamental economic and social importance. The cereal supplies at least half the energy calorie of the world population, mainly for the poverty-

[1] Escola Superior de Desenvolvimento Rural, Universidade Eduardo Mondlane, Av. Julius Nyerere, 3453, Maputo, Moçambique. E-mail: *momateemate@gmail.com*
[2] Universidade Federal de Viçosa, Departamento de Estatística, CEP: 36570-900, Viçosa, MG, Brasil, E-mail: *marciellelis@gmail.com, policarpo@ufv.br, martinsfilho@ufv.br*
[3] Universidade Federal do Acre, Centro de Ciências Exatas e Tecnológicas, Rodovia BR 364, Km 04, Distrito Industrial, CEP: 69920-900 - Rio Branco, AC, Brasil., E-mail: *2santosvinicius@gmail.com*

stricken populations of countries in tropical and subtropical regions, and the so-called emerging or developing countries. In Asia, where 90% of this cereal is grown, the average per capita consumption is high, 78kg, while in Latin American countries, the average per capita consumption is around 29kg, with emphasis in Brazil, considered a great consumer of rice, with an average consumption of 32 kg/person/year (SOSBAI, 2018).

At all stages of growth and development of this culture, biotic and abiotic factors have directly impacted the availability of this food. The factors limiting the productive potential of rice include diseases, mainly blast, which is caused by the fungus *Pyricularia oryzae*. It is the most significant cause of damage both in productivity and grain quality, compromising up to 100% of production, in terms of conditions susceptible to the disease (LAW *et al.*, 2017; SOSBAI, 2018).

The extent of the damage caused by blast depends on the degree of susceptibility of the cultivar, climatic conditions and cultural practices adopted. Despite all research efforts aimed at developing cultivars resistant to blast, the disease remains one of the main factors limiting rice productivity in Brazil and other countries (LAW *et al.*, 2017; SILVA-LOBO *et al.*, 2012).

The logistic regression technique is the statistical method most widely used to verify the relationship between a binary or dichotomous response variable and explanatory of interest. For example, when disease resistance is assessed in locations where the response variable is dichotomous, there are usually two possible responses: resistance ($y = 1$), and the susceptible complementary result ($y = 0$).

When a diagnostic test is developed, it is necessary to evaluate its ability to correctly classify individuals into two groups ($\hat{y} = 0 \vee \hat{y} = 1$), based on the concepts of sensitivity and statistical specificity, obtained from the construction of confusion matrices, generated by the model. According to Martinez *et al.* (2003), a widespread tool for assessing the predictive ability of a model with binary responses is the analysis of the ROC curve (Receiver Operating Characteristic). The ROC curve surged from the theory of signal detection in the early 1950s. Its initial application is dated from the early 1960s in the field of medicine. Since then, the logistic regression technique and ROC curve started to be used in several areas. Yu *et al.* (2014) utilized the ROC curve in the selection and classification of markers. Similarly, Kim (2019) utilized the ROC curve in the classification of Asian rice (*Oryza sativa*) accessions in two subpopulations: indica and japonica.

Currently, logistic regression and ROC curve techniques are widely used in medical biotechnology, but little used in agricultural sciences. Therefore, the present work aimed to select covariates that can influence the resistance to blast and build models to estimate and evaluate this resistance.

2 Material and methods

The set of phenotypic data used was initially generated and analyzed by Zhao *et al.* (2011). It is composed of a population of 413 rice plants (*Oryza sativa* L.) from 82 countries. The data set was formed according to the following subpopulations: mixed (62), aromatic (14), aus (57), indica (87), tempered japonica (96), and tropical japonica (97). Japonica temperate and tropical japonica were combined in japonica, summarizing the variation of global plant genetics. The phenotypic evaluation of rice was carried out in

Stuttgart (Arkansas, USA) from May to October in 2006 and 2007. Two repetitions per year were grown in a randomized block design, in 5m plots, with 25cm spacing between plants and 0.50m between rows. Therefore, the evaluation of resistance of the rice blast involved a sample of sixteen quantitative and categorical variables related to the leaf, seed morphology, yield components, grain quality and level of susceptibility to rice blast disease.

The susceptibility to the disease was rated on a scale of "0" (without lesions of the disease) to "9" (total plant death), when the plants were between three and four weeks old, as described by Marchetti *et al.* (1987). The scale was converted as response (resistant or susceptible plant), according to the size and characteristics of the injuries, as presented by Mackill and Bonman (1992). Thus, plants belonging to classes 0, 1 and 2 were reclassified as resistant ($y = 1$) and plants from classes 3 to 9 were reclassified as susceptible ($y = 0$).

To obtain a parsimonious model (lesser number of covariates) that keeps accuracy in the prediction results, a plan must be developed for the selection of the initial covariates that will be tested and a method that assists in the selection and adequacy of these covariates (HOSMER JUNIOR *et al.*, 2013). Some methods commonly used for the selection of covariates are forward, backward and stepwise, whose algorithms are implemented in various computer programs. However, these methods present some disadvantages, as they tend to identify a particular set of covariates, instead of possible equally good sets, to explain the variable answer, making it impossible for the researcher to choose the most relevant covariates in his practice area (COLOSIMO and GIOLO, 2006).

In this study, the covariate selection strategy proposed by Collett (2003) was used and described by Colosimo and Giolo (2006), in which the information of the researcher can be included in the decision-making process. Thus, it involves more active participation of the statistician and the researcher at each step of the selection process. It can include, for example, covariables considered necessary in the study. According to Colosimo and Giolo (2006), when using this covariate selection procedure, one should avoid being very strict when testing the individual level of significance. The decision of whether a term should be included in the model must be followed by a reasonable (not too low) level of significance; usually, values close to 0.10 are recommended. Gouvêa *et al.* (2009) used this covariate selection strategy to study the variables that affect the time until death or transplantation in patients with chronic renal failure.

The binary response variable evaluated was resistance to rice blast disease, caused by the fungus *Pyricularia oryzae* ($y = 0$: susceptibility, $y = 1$: resistance). The following qualitative and quantitative information of the plants constituted the variables:

V1: Stalk habit: average stalk angle of plants at maturity;
V2: Pubescence of the leaf (0: Absence, 1: Presence);
V3: Length of the flag sheet (cm);
V4: Width of the flag sheet (cm);
V5: Average number of panicles (inflorescences) per plant;
V6: Height of the plant (cm);
V7: Panicle length (cm);
V8: Mean number of branches of the primary panicle;
V9: Average number of seeds per panicle;
V10: Seed length with shell (mm);
V11: Seed width with shell (mm);

V12: Volume of seed with husk;
V13: Surface area of the seed with husk;
V14: Seed length / width ratio;
V15: Amylose quantity present in the ground grains (%).

In the logistic model in which the response variable (y) is binary or dichotomous, its mean conditional must be greater than or equal to zero and less than or equal to one, that is $0 \leq E(Y_i|X) \leq 1$. However, assuming $Y_i \sim Bern(\pi_i)$, the probability of success (occurrence of the event of interest, that is, resistance to rice blast disease), according to logistic regression model, can be defined as:

$$\pi_i = \pi(x_i) = P(Y_i = 1|X = x_i) = \frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)}, \text{ i = 1, 2, ..., n} \quad (1)$$

Furthermore, we can obtain the probability of failure (susceptibility to the disease) by difference:

$$1 - \pi_i = P(Y_i = 0|X = x_i) = \frac{1}{1 + exp(\beta_0 + \beta_1 x_i)} \quad (2)$$

where, β_0 and β_1 are two unknown parameters.

A measure widely used in logistic regression is the *Odds ratio (OR)*, which can be obtained from the ratio between (Eq. 1) and (Eq. 2), given by:

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 x_i} \quad (3)$$

Since (Eq. 1) is non-linear, a transformation called logit is applied, defined by g(x), to make the model linear in its continuous parameters and make it assume values between $-\infty$ and $-\infty$, depending on the limit of the variable, as presented below:

$$g(x) = ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i \quad (4)$$

where $g(x) = ln(\pi_i/1 - \pi_i)$ is the canonical link function for the binomial model (HOSMER JUNIOR *et al.*, 2013).

For this reason, when interpreting the logistic regression coefficients, one opts for the interpretation of $exp(\beta)$ and not directly of β. Ayres *et al.* (2005) reported that the reason of chance is a measure of straightforward interpretation, with statistical properties fundamental for several studies.

In the presence of p independent variables denoted as a vector $\boldsymbol{X_i'} = (x_{0i}, x_{1i}, x_{2i}, \ldots, x_{pi})$ and vectors of unknown parameters denoted by $\boldsymbol{\beta'} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ categorical or continuous, the logistic model establishes a relationship between these p variables and the probability of successful occurrence of a dependent variable binary or dichotomous. So, we can rewrite (π_i) as follows:

$$\pi_i = P(Y = 1) = \frac{exp(\boldsymbol{X_i'\beta})}{1 + exp(\boldsymbol{X_i'\beta})} \quad (5)$$

In each step of the covariate selection process, the likelihood ratio test (LRT) was used, in which the statistics of the test is given by the difference between the *deviances* of the two models to be compared. This difference has a chi-square distribution with the number of degrees of freedom given by the difference between the number of parameters of the two models. LRT compares only models with the same hierarchical or nested structure (HOSMER JUNIOR *et al.*, 2013).

The following expression defines the deviance of a model:

$$LRT = -2\ ln\left(\frac{likelihood\ of\ the\ adjusted\ model}{likelihood\ of\ the\ saturated\ model}\right) \quad (6)$$

in which the adjusted model has unknown p parameters, p is the number of covariables included in the model, and the saturated model has n unknown parameters, where n is the number of observations. The LRT statistic has a chi-square distribution with $n - p$ degrees of freedom and plays the same role as the sum of squares of residues in linear regression (HOSMER JUNIOR *et al.*, 2013). After selecting the covariates using the proposal of Collett (2003), the quality of adjustment of the final model was verified through the Hosmer and Lemeshow test and the ROC curve, that is, if the model was efficient to describe the relationship between the predictor variables and the response variable.

The Hosmer and Lemeshow test corresponds to a chi-square test with $g - 2$ degrees of freedom and consists of dividing the number of observations into ten groups ($g = 10$) and then compare the predicted frequencies with those observed. The purpose of this test is to check if there are significant differences between the classifications performed by the model and the observed reality. For this test, the *generalhoslem* package (Jay, 2019) of the R *software* (R Core Team, 2020) is used.

Hence, to estimate and evaluate the resistance of plants, models that best described the relationship between the predictor variables and the response variable were selected based on the Hosmer and Lemeshow test, at the level of significance of 0.05.

Besides, the values of the Akaike information criterion (AIC) for the selected models were utilized based on the concept of information entropy. The AIC criteria offer a relative measure of information that is lost when a model is used to describe reality (Akaike, 1974), and is calculated from the expression:

$$AIC = -2\ log\ L(\hat{\beta}) + 2p \quad (7)$$

where $L(\hat{\beta})$ is the maximum value of the likelihood function and p is the number of parameters of the model.

According to Burnham and Anderson (2004), the AIC criterion provides an effective means for comparing models, and its operation is based on estimating the information lost by the model (the less information a model loses, the higher its quality). Thus, we must select the model that minimizes the amount of information lost, and the best model is the one with the smallest AIC.

In order to assess the predictive capacity of the selected models, a completed ROC curve was built, and the values of the area below the ROC curve – AUC (Area Under Curve) were obtained. For this purpose, the values of (1 – specificity) were plotted on the abscissa axis and sensitivity in the ordinate axis, obtained from the matrix of 2×2 confusion, generated by each model.

According to Hosmer *et al.*, 2013, the general rule for evaluating the result of the area under the ROC Curve is presented below:

- If $AUC = 0.5$: there is no discrimination;
- If $0.5 < AUC < 0.7$: poor discrimination;
- If $0.7 \leq AUC < 0.8$: reasonable discrimination;
- If $AUC \geq 0.8$: excellent discrimination.

The construction of the ROC curve was implemented through the ROCR package (SING *et al.*, 2005) of the R *software* (R Core Team, 2020).

3. Results and discussion

The strategy derived from Collett's proposal (2003) was utilized to select the covariates included in the final model, whose results are shown in Table 1.

Table 1 – Selection of covariables using logistic regression model and strategy derived from Collett's (2003) proposal, to predict rice blast resistance (*Oryza sativa* L.)

Steps	Model	-2log L(β)	Statistic of test LRT	P-value
Step 1	Null	314.7266	-	-
	V1	308.9986	5.7280	0.0167*
	V2	313.0265	1.7001	0.1923
	V3	313.7101	1.0166	0.3133
	V4	300.4386	14.288	0.0002*
	V5	314.2728	0.4539	0.5005
	V6	314.6075	0.1191	0.7300
	V7	310.6362	4.0904	0.0431*
	V8	298.4124	16.3143	0.0001*
	V9	310.4151	4.3116	0.0379*
	V10	308.8089	5.9177	0.0150*
	V11	307.0889	7.6377	0.0057*
	V12	313.2573	1.4694	0.2254
	V13	314.7111	0.0155	0.9009
	V14	306.4341	8.2926	0.0040*
	V15	306.3386	8.3880	0.0038*
Step 2	V1+V4+V7+V8+V9+V10+V11+V14+V15	265.9581	-	-
(Without V1)	V4+V7+V8+V9+V10+V11+V14+V15	266.6430	0.6849	0.4079

(Without V4)	V1+V7+V8+V9+V10+V11+V14+V15	268.6786	2.7205	0.0991*
(Without V7)	V1+V4+V8+V9+V10+V11+V14+V15	267.4885	1.5304	0.2161
(Without V8)	V1+V4+V7+V9+V10+V11+V14+V15	275.4432	9.4851	0.0021*
(Without V9)	V1+V4+V7+V8+V10+V11+V14+V15	267.5873	1.6292	0.2018
(Without V10)	V1+V4+V7+V8+V9+V11+V14+V15	272.9615	7.0033	0.0081*
(Without V11)	V1+V4+V7+V8+V9+V10+V14+V15	273.3302	7.3721	0.0066*
(Without V14)	V1+V4+V7+V8+V9+V10+V11+V15	271.8139	5.8558	0.0155*
(Without V15)	V1+V4+V7+V8+V9+V10+V11+V14	270.6782	4.7201	0.0298*
Step 3	V4+V8+V10+V11+V14+V15	271.1824	-	-
	V4+V8+V10+V11+V14+V15+V1	269.2947	1.8877	0.1695
	V4+V8+V10+V11+V14+V15+V7	268.7549	2.4276	0.1192
	V4+V8+V10+V11+V14+V15+V9	268.6713	2.5111	0.1131
Step 4	V4+V8+V10+V11+V14+V15	271.1824	-	-
	V4+V8+V10+V11+V14+V15+V2	271.1223	0.0601	0.8063
	V4+V8+V10+V11+V14+V15+V3	265.1750	6.0074	0.0142*
	V4+V8+V10+V11+V14+V15+V5	269.4139	1.7685	0.1836
	V4+V8+V10+V11+V14+V15+V6	266.4102	4.7722	0.0289*
	V4+V8+V10+V11+V14+V15+V12	265.8614	5.3210	0.0211*
	V4+V8+V10+V11+V14+V15+V13	264.1504	7.0320	0.0080*
Step 5	V4+V8+V10+V11+V14+V15+V3+V6+V12+V13	259.1169	-	-
(Without V4)	V8+V10+V11+V14+V15+V3+V6+V12+V13	262.6172	3.5004	0.0614*
(Without V8)	V4+V10+V11+V14+V15+V3+V6+V12+V13	267.7432	8.6265	0.0033*
(Without V10)	V4+V8+V11+V14+V15+V3+V6+V12+V13	259.3160	0.1992	0.6554
(Without V11)	V4+V8+V10+V14+V15+V3+V6+V12+V13	266.7422	7.6255	0.0058*
(Without V14)	V4+V8+V10+V11+V15+V3+V6+V12+V13	260.9080	1.7912	0.1808
(Without V15)	V4+V8+V10+V11+V14+V3+V6+V12+V13	265.0047	5.8879	0.0153*
(Without V3)	V4+V8+V10+V11+V14+V15+V6+V12+V13	261.1983	2.0815	0.1491
(Without V6)	V4+V8+V10+V11+V14+V15+V3+V12+V13	260.1793	1.0626	0.3026
(Without V12)	V4+V8+V10+V11+V14+V15+V3+V6+V13	259.2079	0.0911	0.7627
(Without V13)	V4+V8+V10+V11+V14+V15+V3+V6+V12	260.3089	1.1922	0.2749
Step 6	V4+V8+V11+V15	281.2497	-	-
	V4+V8+V11+V15+V4×V8	281.2217	0.0280	0.8671
	V4+V8+V11+V15+V4×V11	278.7223	2.5274	0.1119
	V4+V8+V11+V15+V4×V15	278.4753	2.7744	0.0958*
	V4+V8+V11+V15+V8×V11	281.0476	0.2021	0.6531

	V4+V8+V11+V15+V8×V15	280.3919	0.8578	0.3544
	V4+V8+V11+V15+V11×V15	274.2641	6.9856	0.0082*
Stage	V4+V8+V11+V15+V4×V15+V11×V15	268.4995		
Final	V4+V8+V11+V15+V4×V15	278.4753		
	V4+V8+V11+V15+V11×V15	274.2641		

*P-value < 0.10.

First, in step 1, all models containing a single covariate. By testing the likelihood ratio, it was found that the covariables V1, V4, V7, V8, V9, V10, V11, V14 and V15 were significant at the level of 0.10, that is, it was demonstrated that they have some influence on rice blast resistance (response variable).

In step 2, the previously significant covariates (step 1) were adjusted together. According to Colosimo and Giolo (2006), in the presence of certain covariables, others may cease to be significant. Thus, reduced models were adjusted, excluding a single covariate at a time. It was found that only the V4 covariates, V8, V10, V11, V14 and V15 significantly increased the ratio statistic of likelihood. Thus, only those covariates continued in step 3.

In step 3, with the covariates that were significant in step 2, a new model and the covariables that were excluded in step 2 returned to the model to confirm that they were not statistically significant.

In step 4, since the covariates included in step 3 (V1, V7, V9), one at a time, showed no significance, the reference model from step 3 was maintained and returned with the covariables excluded in step 1, one at a time, to confirm that they were not statistically significant.

Then, in step 5, a model included the covariables (V3, V6, V12, V13) significant in step 4, and it was tested whether any covariate could be removed from the model. It was observed that the covariables (V10, V14, V3, V6, V12, V13) presented no statistical significance and were removed from the model.

Finally, in step 6, a model was adjusted with the covariables that were significant in step 5. Furthermore, to complete modelling, the possible inclusion of terms of double interaction between covariates already included in the model required previous analysis.

It was observed that the interactions V4×V15 and V11×V15 were significant at the level of 0.10. Thus, in the final stage, three models were selected to estimate the probability occurrence of resistance to rice blast (*Oryza sativa* L.), as follows: Model 1: V4+V8+V11+V15+V4×V15+V11×V15; Model 2: V4+V8+V11+V15+V4×V15 and Model 3: V4+V8+V11+V15+V11×V15.

Once the three models were selected, the Hosmer and Lemeshow test was applied to verify the quality of the fit, the Area Under the ROC Curve (AUC) was used to assess the predictive capacity of the models, and the AIC values were used for the comparison of the models (Table 2).

Table 2 – Hosmer and Lemeshow test and model performance measures: Model 1 (with two interactions between the covariables V4×V15 and V11×V15), Model 2 (with the interaction between V4×V15 covariates) and Model 3 (with the interaction between covariates V11×V15)

	Chi-Quad.	GL	P-value	AIC	AUC
Model 1	14.58	8	0.068	282.4995	0.7553
Model 2	14.97	8	0.060	290.4753	0.7259
Model 3	16.27	8	0.039*	286.2641	0.7391

* Significant at 0.05
AIC - Akaike information criteria; AUC - Area under the curve.

Table 2 shows that there were no differences between the predicted and observed values for models 1 and 2, by the Hosmer and Lemeshow test, which indicates that these models were able to produce reliable ratings for resistance to rice blast. According to the area under the ROC curve, the three models presented good discrimination power ($0.7 \leq AUC \leq 0.8$), according to Hosmer Junior *et al.* (2013). However, model 1 was the best model; since it presented a lower AIC value and higher AUC value. In model 3, a significant difference was observed between the predicted and observed values by the Hosmer and Lemeshow test. This indicates that the model is not fitted to the data. Therefore, it was excluded from further analysis.

Once the models capable of producing reliable classifications were known, the percentages of the predictive capabilities of the models were obtained, as shown in Table 3.

Table 3 – Measures of predictive capacity (%) of the models adjusted with four covariates (V4, V8, V11 and V15), with Model 1 (with two interactions: V4×V15 and V11×V15) and Model 2 (with only one interaction: V4×V15)

	Sensitivity	Specificity	Accuracy
Model 1	93.7	80.3	74.6
Model 2	96.1	85.9	75.0

Sensitivity and Specificity: percentage of correct classifications or predictions of resistant and susceptible plants to rice blast, respectively. Accuracy: total percentage correct classifications.

Table 3 showed that the percentage of correct responses for sensitivity and specificity were higher in model 2 than in model 1. Regarding the percentage accuracy, both models presented almost equal values, around 75%.

Therefore, the ROC Curve of the two models was built according to the sensitivity and 1-Specificity (Figure 1).

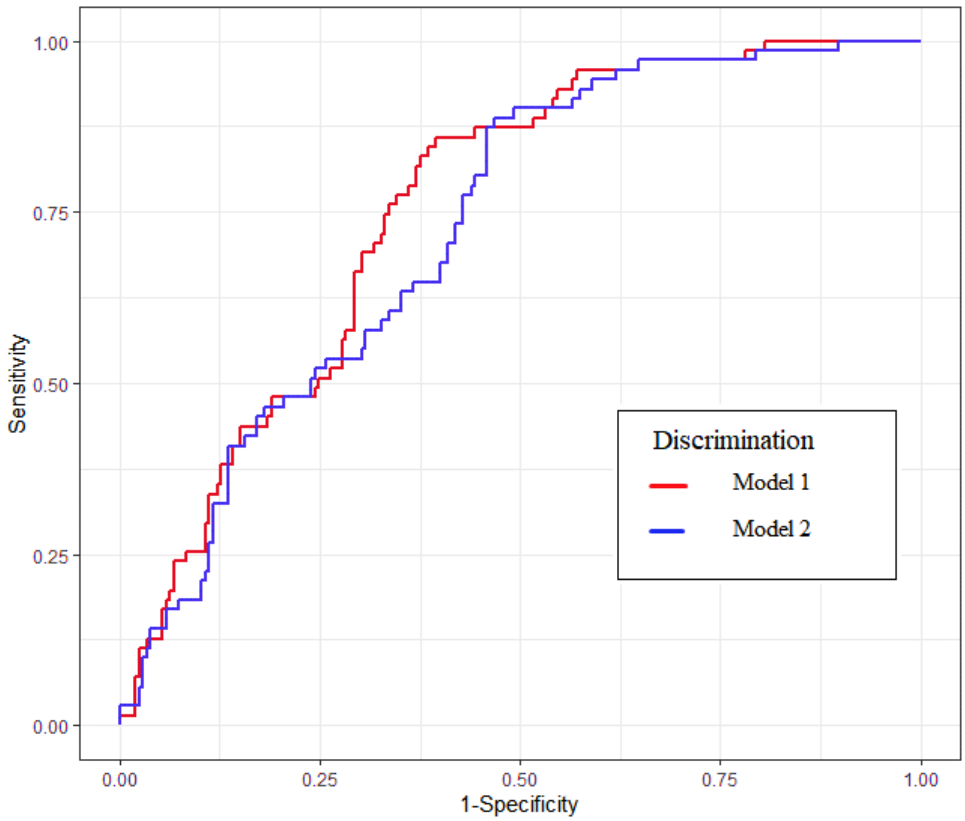

Figure 1 – ROC curve of the models fitted with four covariates (V4, V8, V11 and V15), so that Model 1 (with two interactions: V4✕V15 and V11✕V15) and Model 2 (with only one interaction: V4✕V15).

Although the curves are close (Figure 1), there were different performance models (1 and 2) at specific points of (1-Specificity) and sensitivity, which can be close between the points (0.25; 0.50) and (0.50; 0.875). According to Bozdogan (1987), if two or more models are well adjusted and have an adequate predictive capacity, one should prefer the model that involves the smaller number of parameters to be estimated, which explains well the behaviour of the response variable. Thus, for the prediction of rice blast resistance, it was selected model 2, for being the most parsimonious model.

Thus, estimates of model 2 parameters were obtained, in order to identify the significant effects. Table 4 shows the parameter estimates, their standard errors and the odds ratio values.

Table 4 – Estimates of the parameters of model 2 of logistic regression selected from 16 variables to predict resistance to rice blast (*Oryza sativa* L.), containing four covariables: width of the flag leaf (V4), average number of the branching of the primary panicle (V8), width of the seed with shell (V11), quantity of amylose presents in the ground grains (V15) and the (V4×V15) interaction

Variables	Coefficient	Standard error	Z	P-value	OR	95% CI
Intercept	-10.796	4.426	-2.439	0.015	-	-
V4	6.403	3.001	2.127	0.033	603.916	(4.274 - 85334.878)
V8	0.257	0.094	2.743	0.006	1.293	(1.108 - 1.509)
V11	-0.725	0.438	-1.658	0.097	0.484	(0.236 - 0.994)
V15	0.358	0.186	1.925	0.054	1.430	(1.053 - 1.941)
V4 × V15	-0.233	0.140	-1.663	0.096	0.792	(0.629 - 0.997)

OR - Odds ratio; 95% CI - 95% confidence interval for OR.

Table 4 demonstrates that, at 5% significance level, the variables included in the model, V4 and V8. The variable V15 presented significance ($p = 0.054$) and was also included in the model, which indicates that these variables influence the disease.

However, the variable V11 and the V4×V15 interaction presented no statistical significance ($p > 0.05$), which indicates that these can be removed from the model. According to the results obtained in Table 4, a new model was adjusted, including only the three covariates (V4, V8 and V15). Once the model was adjusted, the quality of the fit was tested, by the Hosmer and Lemeshow test, and no significant differences ($p > 0.05$) were found between the predicted and the observed frequencies in the model. It indicates that the model was able to produce reliable ratings. The values of parameter estimates, their standard error, and the values of the model´s odds ratios, are presented in Table 5.

Table 5 – Estimates of the parameters of logistic regression model selected for the prediction of resistance to rice blast (*Oryza sativa* L.) with three variables: flag leaf width (V4), mean number of primary panicle branch (V8) and amount of amylose present in the ground grains (V15)

Variables	Coefficient	Standard error	Z	P-value	OR	95% CI
Intercept	-7.370	1.289	-5.717	0.000	-	-
V4	1.333	0.653	2.042	0.041	3.793	(1.296 – 11.097)
V8	0.277	0.092	3.000	0.003	1.319	(1.133 - 1.535)
V15	0.090	0.033	2.760	0.006	1.094	(1.037 - 1.155)

OR - Odds ratio; 95% CI - 95% confidence interval for OR.

It was observed that the covariates width of the flag leaf (V4), average number of the branching of the primary panicle (V8) and amount of amylose present in the ground grains (V15) presented statistical significance at the level of 5%, which indicates that these variables influence blast resistance. According to estimates for odds ratio (OR), a one cm increase in V4 causes 279.3% of the expected increase in the probability of blast resistance. For the addition of a V8, a 31.9% increase in the probability of resistance is expected, on average, while for each percentage unit of increase in V15, a 9.4% increase is expected in the probability of blast resistance.

The permanence of the covariates width of the flag leaf (V4) and average number of the branching of the primary panicle (V8), in the final logistic model, can be explained by the fact that blast attacks mainly the leaves and panicles, causing losses in the grain yield and quality. Blast in the leaves causes indirect damage to grain production, by reducing the rate of photosynthesis and respiration, while blast in panicles directly affects grain formation and weight (SILVA-LOBO *et al.*, 2012).

In their study on the associations between agronomic variables of rice genotypes, Castro *et al.* (2019) found high negative correlations between blast severity and rice grain yield (r = -0.96), which reveals that the more significant the incidence of the blast, the lower the productivity and growth of the plant.

The importance of the variable width of the flag leaf in rice productivity, and consequently, in resistance to diseases such as blast, was verified by Aditya and Bhartiya (2013). Dalchiavon *et al.* (2012) reported a significant correlation between the number of panicles and the yield of rice grains.

Amylose content is considered the most important characteristic related to the quality of rice grains. Amylose is one of the two factions that make up starch (the other is amylopectin). It generally varies between 3% and 33% in rice, and the varieties with intermediate amylose content (20% to 25%) are the most preferred by consumers all over the world, due to their dry, loose and soft grains (JAMALODDIN *et al.*, 2020). According to Ong and Blanshard (1995), grains with higher amylose content have a firmer texture after cooking.

Unlike the findings of Zhang *et al.* (2006), who detected no significant association between amylose content and blast resistance, this study demonstrated, through the logistic regression model, that the amylose content favors plant resistance to blast, which proves the effect of the disease on the quality of rice grains.

4. Conclusions

Out of the fifteen variables initially used to assess the disease, only three: flag leaf width (V4), the mean number of primary panicle branches (V8), and amount of amylose present in the ground grains (V15), proved to be important in the final adjusted model. The influence of these covariates showed that the more significant the increase in the value of these covariates, the greater the resistance to blast disease and, consequently, the greater the productivity of rice cultivars. It was found that the variable (V4) has the most significant effect on blast resistance, with a 279.3% probability of resistance to blast for each unit of cm increase of the flag leaf width.

Acknowledgments

The authors are thankful to the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Coimbra Group of Brazilian Universities (GCUB), for the Training Program for Higher Education Professors from African Countries (ProAfri) and the scholarship granting (Finance Code 001).

OSSIFO, M. E., DUARTE, M. L., CARNEIRO, A. P. S., SANTOS, V. S., MARTINS FILHO, S. Seleção de variáveis em modelo de regressão logística para predição da resistência à brusone do arroz. *Braz. J. Biom*. Lavras, v.40, n.2, p.166-180, 2022.

- *RESUMO: O arroz (Oryza sativa L.) tem sido um dos alimentos mais consumidos no planeta, com importância econômica e social. Doenças, principalmente a brusone, causadas pelo fungo Pyricularia oryzae, são fatores limitantes para a produção de arroz. O presente trabalho teve como objetivo selecionar covariáveis que possam influenciar a resistência do arroz à brusone, utilizando o método de seleção proposto por Collett. Modelos de regressão logística foram ajustados para prever a resistência à doença, usando a curva ROC para avaliar a capacidade preditiva. Os dados utilizados foram obtidos de uma população de 413 plantas, com informações fenotípicas coletadas em 82 países e classificadas em cinco subpopulações. A pesquisa constatou que, das mais de quinze variáveis incorporadas para avaliar a doença, apenas três se mostraram relevantes para o modelo final ajustado, sendo: largura da folha bandeira (V4), o número médio de ramos primários da panícula (V8) e a quantidade de amilose de grãos moídos (V15). A variável V4 apresentou uma maior influência significativa na resistência à doença. Sendo que, para cada aumento unitário em V4, V8 e V15, espera-se obter aumentos de 279,3, 31,9 e 9,4%, respectivamente, na probabilidade de resistência à brusone*.

- *PALAVRAS-CHAVE: Curva ROC; Método de Collet; Oryza sativa*; *Pyricularia oryzae*.

References

ADITYA, J. P.; BHARTIYA, A. Genetic variability, correlation and path analysis for quantitative characters in rainfed upland rice of Uttarakhand Hills. *Journal of Rice Research*, v.6, n.12, p.24–34, 2013.

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v.19, n.6, p.716–723, 1974.

AYRES, M.; AYRES JUNIOR, M.; AYRES, D. L.; SANTOS, A. S. *BioEstat 4.0: Statistical applications in the areas of biological and medical sciences.* Belém: Society Civil Mamirauá; Brasília: CNPq, 2005.

BOZDOGAN, H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, v.52, n.3, p.345–370, 1987.

BURNHAM, K. P.; ANDERSON, D. R. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, v.33, n.2, p.261–304, 2004.

CASTRO, D. G.; FERNANDES, M. C. N.; FÉLIX, M. R.; CAZASSA, R. S.; TOMÉ, L. M.; BOTELHO, F. B. S. Estimates of association between agronomic characters in the selection of upland rice genotypes. *Magistra*, v.30, p.359-367, 2019.

COLLETT, D. *Modeling survival data in medical research.* London: Chapman and Hall, 2nd ed., 2003. 410p.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. São Paulo: Edgar Blücher, 2006. 392p.

DALCHIAVON, F. C.; CARVALHO, M. P.; COLETTI, A. J.; CAIONE, G.; SILVA, A. F.; ANDREOTTI, M. Correlação linear entre componentes da produção e produtividade do arroz de terras altas em sistema plantio direto. *Semina: Ciências Agrárias*, v.33, n.5, p.1629–1642, 2012.

GOUVÊA, G. D. R.; OLIVEIRA, F. L. P.; VIVANCO, M. J. F. Event analysis factors: an application to hemodialysis data in the city of Lavras-MG, *Rev. Bras. Biom*., v.27, n.3, p.491–500, 2009.

HOSMER JUNIOR, D. W.; LEMESHOW, S.; STURDIVANT, *R. X. Applied logistic regression*, New York: John Wiley & Sons, 3rd ed., 2013. 528p.

JAMALODDIN, M.; DURGA RANI, C. V.; SWATHI, G.; ANURADHA, C.; VANISRI, S.; RAJAN, C. P. D.; *et al.* Marker assisted gene pyramiding (MAGP) for bacterial blight and blast resistance into mega rice variety "Tellahamsa", *PLOS ONE*, v.15, n.6, 2020.

JAY, M. *Generalhoslem: goodness of fit tests for logistic regression models*, 2019. Available in: https://cran.r-project.org/web/packages/generalhoslem/index.html.

KIM, B. Classifying Oryza sativa accessions into Indica and Japonica using a logistic regression model with phenotypic data. *PeerJ*, v.2019, n.11, 2019.

LAW, J. W. F.; SER, H. L.; KHAN, T. M.; CHUAH, L. H.; PUSPARAJAH, P.; CHAN, K. G.; GOH, B. H.; LEE, L. H. The potential of Streptomyces as biocontrol agents against the rice blast fungus, *Magnaporthe oryzae (Pyricularia oryzae). Frontiers in microbiology*, v.8, n.3. 2017.

MACKILL, A. O.; BONMAN, J. M. Inheritance of blast resistance in near-isogenic lines of rice. *Phytopathology*, v.82, p.746–749, 1992.

MARCHETTI, M. A.; LAI, X. H.; BOLLICH, C. N. Inheritance of resistance to *Pyricularia oryzae* in rice cultivars grown in the United States. *Phytopathology*, v.77, n.6, p.799-804, 1987.

MARTINEZ, E. Z.; LOUZADA-NETO, F.; PEREIRA, B. B. A curva ROC para testes diagnósticos. *Cadernos Saúde Coletiva*, v.11, n.1, p.7–31, 2003.

ONG, M. H.; BLANSHARD, J. M. V. Texture determinants in cooked, parboiled rice. I: Rice starch amylose and the fine structure of amylopectin. *Journal of Cereal Science*, v.21, n.3, p.251–260, 1995.

R CORE TEAM. *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. Available at: https://www.R-project.org/

SILVA-LOBO, V. L.; FILIPPI, M. C. C.; SILVA, G. B.; VENANCIO, W. L.; PRABHU, A. S. Relação entre o teor de clorofila nas folhas e a severidade de brusone nas panículas em arroz de terras altas. *Tropical Plant Pathology*, v.37, n.1, p.83-87, 2012.

SING, T.; SANDER, O.; BEERENWINKEL, N.; LENGAUER, T. ROCR: Visualizing classifier performance in R. *Bioinformatics*, v.21, n.20, p.78-81, 2005. Available at: http://rocr.bioinf.mpi-sb.mpg.de/.

SOSBAI - Sociedade Sul-Brasileira de Arroz Irrigado. *Arroz Irrigado: recomendações técnicas da pesquisa para o Sul do Brasil*. 32 Reunião da Cultura do Arroz Irrigado. Farroupilha, RS. Cachoeirinha: Sociedade Sul-Brasileira de Arroz Irrigado, 2018. 205p.

YU, W. B.; CHANG, Y. C. I.; PARK, E. A modified area under the ROC curve and its application to marker selection and classification. *Journal of the Korean Statistical Society*, v.43, n.2, p.161–175, 2014.

ZHANG, S.; LIU, B.; ZHU, X.; YANG, J.; WU, S.; HEI, L. Relationship between blast resistance and amylose content in a RIL population derived from rice crossed SHZ-2xLTH. *Acta Agronomica Sinica*, v.32, n.2, p.159–163, 2006.

ZHAO, K.; TUNG, C. W.; EIZENGA, G. C.; WRIGHT, M. H.; ALI, M. L.; PRICE, A. H.; NORTON, G. J.; ISLAM, M. R.; REYNOLDS, A.; MEZEY, J.; MCCLUNG, A. M.; BUSTAMANTE, C. D.; MCCOUCH, S. R. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa. Nature Communications*, v.2, n.1, p.1–10, 2011. Available at: https://doi.org/10.1038/ncomms1467.

Recebido em 14.06.2021

Aprovado após revisão em 10.09.2021