BELO AFONSO MUETANENE

SELECTION INDICES AND SUPPORT VECTOR MACHINES IN THE SELECTION OF SUGARCANE FAMILIES

Thesis submitted to the Applied Statistics and Biometry Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

Adviser: Luiz Alexandre Peternelli

VIÇOSA - MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
M948s
2022

Muetanene, Belo Afonso, 1984-
Selection indices and support vector machines in the selection of sugarcane families / Belo Afonso Muetanene. – Viçosa, MG, 2022.
1 tese eletrônica (33 f.): il.

Texto em inglês.
Orientador: Luiz Alexandre Peternelli.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Estatística, 2022.
Referências bibliográficas: f. 29-33.
DOI: https://doi.org/10.47328/ufvbbt.2022.664
Modo de acesso: World Wide Web.

1. Cana-de-açúcar - Melhoramento genético - Métodos estatísticos. 2. Aprendizado do computador. I. Peternelli, Luiz Alexandre, 1966-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 631.520727

BELO AFONSO MUETANENE

SELECTION INDICES AND SUPPORT VECTOR MACHINES IN THE SELECTION OF SUGARCANE FAMILIES

Thesis submitted to the applied Statistics and Biometry Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

APPROVED: 27 October, 2022.

Assent:

Belo Afonso Muetanene
Author

Luiz Alexandre Peternelli
Adviser

To God, my mother and my kids

ACKNOWLEDGEMENTS

To God for everything.

To my parents, especially my mother Aida Florinda (*in memoriam*) for all She did while in life, my eternal Gratitude, my brothers and my kids Pércius Belo Afonso Muetanene and Eshley Belo Afonso Muetanene.

To the Federal University of Viçosa, a dream is coming true for the opportunity to attend the Doctorate course in Applied Statistics and Biometry.

My eternal Gratitude to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES). I would also like to extend my gratitude to UniLúrio for this great opportunity.

My Gratitude to my Adviser, Luiz Alexandre Peternelli, from the beginning to the end of this course, for all the support.

ABSTRACT

MUETANENE, Belo Afonso, D.Sc., Universidade Federal de Viçosa, October, 2022. **Selection indices and support vector machines in the selection of sugarcane families.** Adviser: Luiz Alexandre Peternelli.

The present study aimed to compare selection indices, namely: Smith and Hazel multiplicative, Mulamba and Mock's, and the support vector machines algorithm for sugarcane families selection. We used two datasets, from Moreira et al. (2021) and from Ferreira et al. (2022), both related to the sugarcane breeding program conducted at the Center for Sugar cane Research and Breeding at the Federal University of Viçosa, Oratórios, Minas Gerais. Both experiments were conducted in a randomized complete block design. We constructed the selection indices via mixed models approach. We adopted a selection percentage of 18% of the top families for the selection process. In both studies, we considered as explanatory traits: the number of stalks, stalks diameter and stalk height, and as the response trait the tons of stalks per hectare per family. In the dataset from Ferreira et al. (2022), the support vector machine was a better approach to select sugarcane families by learning from the data after multivariate simulation. Whereas in the dataset from Moreira et al. (2021), using similar methodology, lower performance for support vector machines was obtained.

Keywords: Synthetic data. Indirect selection. Yield prediction. Machine learning. BLUP

RESUMO

MUETANENE, Belo Afonso, D.Sc., Universidade Federal de Viçosa, outubro de 2022. **Índices de seleção e máquinas vector de suporte na seleção de famílias de cana-de-açúcar.** Orientador**:** Luiz Alexandre Peternelli.

O presente estudo teve o objetivo de comparar índices de seleção, nomeadamente: índices de Smith e Hazel, multiplicativo e Mulamba e Mock com máquinas vector de suporte para a seleção de famílias de cana-de-açúcar. Foram usadas duas bases de dados, a de Moreira et al. (2021) e a de Ferreira et al. (2022), ambas relacionadas ao programa de melhoramento genético da cana-de-açúcar conduzido no Centro para a Pesquisa e Melhoramento Genético da cana-de-açúcar na Universidade Federal de Viçosa, Oratórios, Minas Gerais. Ambos os estudos foram conduzidos em delineamento de blocos completos. Os índices de seleção foram construídos via modelos mistos, sendo adotada uma percentagem de seleção de 18% das melhores famílias para a seleção. Em ambos os estudos, foram consideradas como variáveis explicativas: número de colmos, diâmetro do colmo e altura do colmo e como variável resposta o rendimento do colmo por hectare (toneladas/hectare). Na base de dados de Ferreira et al. (2022), as máquinas vector de suporte foram melhor método para selecionar famílias de cana-de-açúcar a partir da aprendizagem dos dados após simulação multivariada. Para os dados de Moreira et al. (2021), usando metodologia similar, pior desempenho para as máquinas vector de suporte foi observado.

Palavras-chave: Dados sintéticos. Seleção indireta. Predição de rendimento. Aprendizagem de máquina. BLUP

LIST OF TABLES

SUMMARY

1. INTRODUCTION

Brazil is worldwide known as being the major producer of sugarcane (*Saccharum* sp) (BORDONAL et al., 2018), the first estimate for the 2021/2022 harvest season indicates the production of approximately 628 10^6 tons of sugarcane in a cultivated area of nearly 8.42 million hectares (CONAB, 2021). Sugarcane breeding plays a critical role, as it allows the breeders to develop varieties with better agronomic and industrial traits, such as superior yield, more resistance to pests and diseases (PEDROZO et al., 2009), and better adapted to specific regions (BARBOSA and PINTO, 1998). The process of obtaining new sugarcane varieties is lengthy and costly, generally, new varieties are launched after ten years of careful evaluations in numerous stages (FERREIRA et al., 2022). It´s known that among the sugarcane breeding phases, the initial phase is the main challenge facing breeders due to the enormous number of genotypes that need to be evaluated at the beginning of a selection cycle (MOREIRA et al., 2021).

Sugarcane breeders have been focused on improving some traits such as yield, stalk height, stalk diameter, sugar content, and disease resistance. It is known that some of these traits are positive or negatively correlated with yield and the selection performed for one or few traits may result in superior genotypes for only one or few traits (BÁRBARO et al., 2007; PEDROZO et al., 2009; VASCONCELOS et al., 2010).

The most import bioproduct derived from sugarcane crop is the sugar, sugar is positive and highly correlated with the tons of stalks per hectare. In sugarcane breeding programs, tons of stalks per hectare is the main trait of interest, however, having to weight the stalks from the plots is a laborious and expensive task. Thus, to reduce the harvest costs, sugarcane breeders have used a process known as indirect selection, where the traits such as number of stalks, stalks diameter and stalks are used indirectly to select sugarcane genotypes for the trait tons of stalks per hectare.

The indirect selection is performed by using selection indices, the selection indices combine multiple traits (ENTRINGER et al., 2016; COUTINHO et al., 2019), the important things to construct a selection index are determining the economic weights of the traits so that the selection can be more representative and accurate (SINGH and CHAUDHARY, 2007). In the selection indices approach, only those individuals predicted to have progeny of superior economic value are selected and then continued further in the breeding program (QUINTON and MCMILLAN, 1995).

However, determining appropriate economic weights for different traits can be challenging (CERÓN-ROJAS et al., 2006). Several studies have been conducted related to selection indices in various crops, such as potato (BARBOSA; PINTO, 1998), rice (SMIRDELE et al., 2019; VENMUHIL et al., 2020), bean (MENDES et al., 2009; MARINHO et al., 2014), sugarcane (PEDROZO et al., 2009; ALMEIDA et al., 2014) and soybean (GESTEIRA et al., 2018; FREIRIA et al., 2019).

Another way of performing sugarcane genotypes selection is by using machine learning models, such models learn from the data. Machine learning models, such as decision trees, artificial neural networks, and support vector machines, have been used to select sugarcane families (PETERNELLI et al., 2017; PETERNELLI et al., 2018; GUTIÉRREZ et al*.,* 2015; MOREIRA et al., 2021). Grapevine varieties were selected using support vector machines and artificial neural network*s* (GUTIÉRREZ et al., 2015;).

For generalization purposes, the support vector machine deserves a special attention, as it doesn´t depend on all the training data, but only on the support vectors which are also a dataset subset, the number of support vectors is very reduced when compared to the training dataset (QIN et al., 2014). The support vector machines has also the advantage of not requiring any data distribution assumptions or homogeneity of covariance matrices, facilitating the classification process.

Studies comparing the use of support vector machines and selection indices in sugarcane breeding programs in Brazil are scarce.

The present study aims to evaluate selection indices, namely multiplicative, Smith and Hazel, and Mulamba and Mock's indices, and support vector machine for sugarcane families selection in two datasets, from Moreira et al. (2021) and Ferreira et al. (2022).

2. FIRST CASE STUDY

2.1 MATERIAL AND METHODS

2.2 Dataset

For this study, we used the dataset from Ferreira et al. (2022), a study related to the sugarcane breeding programme, conducted at the Centre for Sugarcane Research and Breeding, at the Federal University of Viçosa, Oratórios, Minas Gerais (20°25'S, 42°48'W, 494 m of altitude). The experiment was conducted in a randomized complete block design and was comprised by 60 families. For our study, we considered as explanatory traits: number of stalks per meters (NS), stalks diameter (SD, measured in millimeters,) and stalks height (SH, measured in meters) and the response trait as the TSH (measured in tons of stalks per hectare).

2.3 Selection indices approach

We constructed the selection indices via mixed models approach on the Selegen software (RESENDE, 2002), the mixed model used was $\boldsymbol{y} = \boldsymbol{Xr} + \boldsymbol{Zg} + \boldsymbol{Wb} + \boldsymbol{e}$, where: $\boldsymbol{y}$ is the observations vector $\left(\boldsymbol{y} \sim \boldsymbol{N}(\boldsymbol{X_r}, \boldsymbol{V})\right)$; $\boldsymbol{r}$ is the vector of replications effect (assumed fixed) summed to the overall mean, $\boldsymbol{g}$ is the vector of sugarcane families effects (assumed random), $\boldsymbol{g} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$ where $\boldsymbol{G}$ is the genetic covariance matrix of the families $\left(\boldsymbol{G} = \boldsymbol{I}\boldsymbol{\sigma}_{\boldsymbol{g}}^{\boldsymbol{2}}\right)$; $\mathbf{b}$ is the vector of blocks effects (assumed random) where $\boldsymbol{b} \sim \boldsymbol{N}\left(\boldsymbol{0}, \boldsymbol{I}\boldsymbol{\sigma}_{\boldsymbol{b}}^{\boldsymbol{2}}\right)$ and, $\boldsymbol{e}$ is the vector of residuals, where $\boldsymbol{e} \sim (\boldsymbol{0}, \boldsymbol{R})$ where $\boldsymbol{R}$ is the residual covariance matrix ($\boldsymbol{R} = \boldsymbol{I}\boldsymbol{\sigma}_{\boldsymbol{e}}^{\boldsymbol{2}}$). $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{W}$ represent the incidence matrices of the corresponding effects. About selection indices (SI), the Smith-Hazel (SMITH, 1936; HAZEL, 1943), multiplicative (SUBANDI et al., 1973) and the Mulamba and Mock's indices (MULAMBA; MOCK, 1978) were used to select sugarcane families for tons of stalks per hectare (TSH) based on the indirect traits number of stalks, stalks diameter and stalks height. All the selection indices were computed according to Pedrozo et al. (2009). The Smith-Hazel index (SHI) is given by:

$SHI = (w_{NS}\ x\ NS)(PGV\ x\ NS) + (w_{SD}\ x\ SD)(PGV\ x\ SD) + (w_{SH}\ x\ SH)(PGV\ x\ SH)$,

where:

PGV is the predicted genotypic value, $NS =$ number of stalks, $SD =$ stalks diameter and $SH =$ stalks height, w_{NS} is the NS economic weight, the same for w_{SD} and w_{SH}. For the SHI we used the genotypic standard deviation as the economic weight (PEDROZO et al., 2009).

The multiplicative index (MI) is given by:

$$MI = (PGV\ x\ NS)(PGV\ x\ SD)(PGV\ x\ SH).$$

The Mulamba and Mock's index (MMI) is based on the sum of ranks**.** Initially it ranks the genotypes, for each trait, by assigning higher absolute values to those of better performance and then the values assigned to each trait are summed to obtain the sum of the ranks, which indicates the classification of the genotypes (CRUZ; CARNEIRO, 2003). The smaller the sum, the better the performance of a genotype for the various traits (ALMEIDA et al., 2014). The Mulamba and Mock's index is given by:

$MMI = (r\ x\ PGV\ x\ NS) + (r\ x\ PGV\ x\ SD) + (r\ x\ PGV\ x\ SH)$, where:

r is the genotype's rank.

We adopted a selection percentage of 18% of the top families for the selection process. We predefined the genotypic values for family means of TSH (tons of stalks per hectare) as the best selection method. Thus, to evaluate the selection indices and support vector machines' performance, we computed the coincidence coefficient (CC) between the genotypic values for family means of TSH with each selection indices and with the support vector machines.

The coincidence coefficient (CC) was computed according to the following formula:

$CC = \frac{A}{B}$, where:

A = number of families selected simultaneously by both selection methods involved in each computation (selection indices, support vector machines, and genotypic values for family means of tons of stalks per hectare).

B= the total number of families which pretend to be selected

2.4 Selection via support vector machines

The support vector machines (SVM) classifier performs binary classification, i.e., it separates a set of training vectors for two different classes $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$, where $x_i \in R^d$ denotes vectors in a d-dimensional feature space and $y_i \in \{-1, +1\}$ is a class label (HUANG et al., 2017). The SVM model is generated by mapping the input vectors onto a new higher dimensional feature space denoted as $\Phi: R^d \rightarrow H^f$, where $d < f$. Then an optimal separating hyperplane in the new feature space is constructed by a kernel function $K(x_i, x_j)$, which is the product of input vectors x_i and x_j where $K(x_i, x_j) = \Phi(x_i).\Phi(x_j)$ (HUANG et al., 2017). A kernel

is a function that quantifies the similarities between two observations (JAMES et al., 2013), we used the *radial basis function* (rbf) kernel, where $K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2}$, and $\gamma > 0$ (PEDREGOSA et al., 2011). The rbf kernel has good performance in nonlinearly separable problems and shows good performance in most cases (GÉRON, 2019).

We used the two parameters: C and γ, C (penalty parameter) controls the regularisation, a low C allows to have a reduced margin in the hyperplane, for γ, a higher value tends to overfitting (HARRISON, 2020). We tested for γ the values: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and for C: 0.25, 0.5, 0.75, 1. For the SVM, the explanatory traits were as follows: number of stalks (NS), stalk diameter (SD) and stalk height (SH), the response trait was the tons of stalks per hectare (TSH), the selection criterion was to select only sugarcane families with a production of TSH higher than the overall mean, a value of one was assigned in case of selection and zero otherwise. To improve the SVM performance, we initially standardized the explanatory traits by $x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$, where x'_{ij} is the standardized trait value, x_{ij} is the original trait value and S_j is the trait standard deviation.

We also produced synthetic data via multivariate simulation to improve the SVM training performance, as we only had 60 sugarcane families, a number of families insufficient to train the SVM model, this procedure was also performed by Peternelli et al. (2018) and Moreira et al. (2021). To generate the synthetic data, we performed a simulation based on the covariance matrix $\boldsymbol{\Sigma}$ (positive definite) of the variables NS, SD, SH, and TSH. The Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma}$ was used to generate $\boldsymbol{\Sigma} = \boldsymbol{CC^T}$, where $\boldsymbol{C}$ is a lower triangular matrix $m\,x\,m$ which is the Cholesky factor. A normal multivariate vector $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{CZ}$ was simulated, where $\boldsymbol{\mu}$ is the mean vector of the variables (NS, SD, SH and TSH), $\boldsymbol{C}$ is the Cholesky factor from the covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{Z}$ is a vector of random independent and identically distributed (*iid*) variables with a standard normal distribution. Through this procedure, we generated 1000 row vectors of the type $[\boldsymbol{X_{i1}}, \boldsymbol{X_{i2}}, \boldsymbol{X_{i3}}, \boldsymbol{X_{i4}}]$, where $\boldsymbol{X_{ij}}$ ($i = 1$ to 1000, and $j = 1$ to 4) represents the simulated value of the variable (NS, SD, SH and TSH) for the individual j. This algorithm assures that all the variables have a covariance matrix $\boldsymbol{\Sigma}$ and mean vector $\boldsymbol{\mu}$ (CRESSIE, 1993; HAINING, 2005).

The simulation was conducted in each scenario separately, details on the simulation performed are presented in table 1.

Table 1- Illustration of the training and predictions performed in each scenario

Scenarios	Dataset used	Nr. of families	Simulation and Training	Predictions
1	R_1R_2	60	1000 families	R_3R_4
2	R_1R_3	60	1000 families	R_2R_4
3	R_1R_4	60	1000 families	R_2R_3
4	R_3R_4	60	1000 families	R_1R_2
5	R_2R_3	60	1000 families	R_1R_4
6	R_2R_4	60	1000 families	R_1R_3
7	R_1	60	1000 families	$R_2R_3R_4$
8	R_2	60	1000 families	$R_1R_3R_4$
9	R_3	60	1000 families	$R_1R_2R_4$
10	R_4	60	1000 families	$R_1R_2R_3$

R_1, R_2, R_3, R_4: experiment replications; R_iR_j: dataset comprised by R_i and R_j replications; $R_iR_jR_l$: dataset comprised by R_i, R_j and R_l replications; Nr. of families: number of families; Simulation and Training: we simulated 1000 families from each scenario individually, and the model training was based on the corresponding simulated dataset.

In this study, for the selection via SVM, the selected sugarcane families were ranked based on their decreasing probability of being classified as selected. For support vector machines (SVM) performance, the best parameter values obtained via grid search for each scenario are presented in Table 2.

Table 2- Parameters and best parameter values (BPV) in each scenario, using the support vector classifier, available on the scikitlearn Python package

Dataset used	Parameter	PVBP
R_1R_2	γ	0.3
	C	0.5
R_2	γ	0.3
	C	0.5
R_1R_3	γ	0.4
	C	0.5
R_1R_4	γ	0.5
	C	1
R_3R_4	γ	0.7
	C	0.5
R_2R_3	γ	0.4
	C	1
R_2R_4	γ	0.4
	γ	1
R_1	C	1
	C	0.8
R_3	γ	0.2
	C	0.25
R_4	γ	0.2
	C	0.25

R_1, R_2, R_3, R_4: replications; R_iR_j: dataset comprised by R_i and R_j replications; C: penalty parameter; γ: penalization parameter

2.5. RESULTS AND DISCUSSION

In Table 3, we present the coincidence coefficient (CC) of the selected families by the selection indices and by the genotypic values for family means of TSH (tons of sugarcane per hectare).

Table 3- Coincidence coefficient (CC) for the families selected by the Smith-Hazel (SHI), multiplicative (MI), Mulamba and Mock's (MMI) indices, SVM with the genotypic values for family means of TSH for the sugarcane families selection

	TSH of R_3R_4		TSH of R_2R_4
SHI^{sd}	0.18	SHI^{sd}	0.18
MI	0.27	MI	0.18
MMI	0.18	MM	0.18
SVM	0.63	SVM	0.54
	TSH of R2R3		TSH of R1R2
SHI^{sd}	0.18	SHI^{sd}	0
MI	0.09	MI	0
MMI	0.09	MM	0
SVM	0.36	SVM	0.54
	TSH of R1R4		TSH of R1R3
SHI^{sd}	0.27	SHI^{sd}	0.18
MI	0.27	MI	0
MMI	0.18	MM	0.18
SVM	0.27	SVM	0.45
	TSH of R2R3R4		TSH of $R_1R_3R_4$
SHI^{sd}	0.18	SHI^{sd}	0.27
MI	0.18	MI	0.18
MMI	0.27	MM	0.27
SVM	0.54	SVM	0.54
	TSH of R1R2R4		TSH of $R_1R_2R_3$
SHI^{sd}	0.09	SHI^{sd}	0.09
MI	0.09	MI	0
MMI	0.18	MM	0
SVM	0.36	SVM	0.54

R_1, R_2, R_3, and R_4: experiment replications; R_iR_j: dataset comprised by R_i and R_j replications; $R_iR_jR_l$: dataset comprised by R_i, R_j and R_l replications; SI: selection indices; SVM: support vector machines; TSH: tons of stalks per hectare; SHI^{sd}: Smith and Hazel index-based genetic standard deviation

The CC (coincidence coefficient) of SVM with TSH (genotypic values for family means of tons of sugarcane per hectare) were superior to the CC of TSH with the selection indices in 90% of the scenarios (Table 3), indicating that, in general, the SVM performance to select sugarcane families was better than the selection indices. SVM is a machine learning algorithm that uses offline learning to find the optimal hyperplane. The offline learning means that the SVM depends on the training data to find an equation that separates the two categories (HMEIDI et al., 2008). The lowest and the highest CC values between SVM and TSH were obtained in R_1R_4 (0.27) and R_3R_4 (0.63), respectively. In our study, sugarcane families were ranked by the SVM based on their decreasing probability of being classified as selected whereas, the selection

indices depend on the genotypic values to rank and then select the sugarcane families. The lowest CC values were found between the TSH value with MI (multiplicative index). These results differ from those of Pedrozo et al. (2009), evaluating the same selection indices. MI obtained the best CC value with the total soluble solids content per hectare (Brix production per hectare). However, in that study, the traits used to construct the SI were the number of stalks per meter, average stalks mass, and total soluble solids content. In a study to select sugarcane families using SVM, random forests logistic regression, k-nearest neighbor, and artificial neural networks, SVM outperformed all the other machine learning models (MOREIRA et al., 2021).

Table 4- Coincidence coefficient (CC) of the Smith-Hazel (SHI), multiplicative (MI), Mulamba and Mock's (MMI) indices and with the SVM (support vector machines) for the sugarcane families selection

SVM training	SI	R_3R_4	SVM training	SI	R_2R_4
		SVM			SVM
	SHI^{sd}	0.27		SHI^{sd}	0.27
R_1R_2	MI	0.45	R_1R_3	MI	0.27
	MMI	0.27		MMI	0.36
SVM training		R_2R_3	SVM training		R_1R_2
		SVM			SVM
	SHI^{sd}	0.27		SHI^{sd}	0
R_1R_4	MI	0.27	R_3R_4	MI	0
	MMI	0.27		MMI	0
SVM training		R_1R_4	SVM training		R_1R_3
		SVM			SVM
	SHI^{sd}	0.36		SHI^{sd}	0.36
R_2R_3	MI	0.45	R_2R_4	MI	0.27
	MMI	0.45		MMI	0.45
SVM training		$R_2R_3R_4$	SVM training		$R_1R_3R_4$
		SVM			SVM
	SHI^{sd}	0.18		SHI^{sd}	0.27
R_1	MI	0.27	R_2	MI	0.27
	MMI	0.36		MMI	0.36
SVM training		$R_1R_2R_4$	SVM training		$R_1R_2R_3$
		SVM			SVM
	SHI^{sd}	0.36		SHI^{sd}	0.27
R_3	MI	0.45	R_4	MI	0.27
	MMI	0.36		MMI	0.27

R_1, R_2, R_3, and R_4: experiment replications; R_iR_j: dataset comprised by R_i and R_j replications; $R_iR_jR_l$: dataset comprised by R_i, R_j and R_l replications; SI: selection indices, SHI^{sd}: Smith and Hazel's index using the genetic standard deviation as economic weight; Training with SVM: specifies the column containing the replications used to train the SVM used to predict the sugarcane families in the other replications

In table 4, the lowest CC (coincidence coefficient) between the selection indices with the SVM values were obtained in R_1R_2 (zero). In this case, the selection indices (SI) presented similar CC values with the SVM in 30% of the scenarios. The MMI

(Mulamba and Mock's index) showed the highest CC with the SVM in 40% of the scenarios. On the other hand, the Smith-Hazel index (SHI) presented the lowest CC with the SVM. For the specific case of the SHI, it also uses economic weights in the selection process, the economic weights may be the coefficient of genetic variation (CVg), the ratio CVg /CVe (CVe: coefficient of experimental variation), broad sense heritability and tentatively assigned weights (JÚNIOR et al., 2010, ALMEIDA et al., 2014). However, it is difficult to express the economic value of traits, (JAHUFER and CASLER, 2015), different economic weights result in different selections efficiencies as also verified by Almeida et. al (2014) in sugarcane.

Comparing only the three selection indices, the MMI (Mulamba and Mock's index) presented the best performance whereas the SHI the worst, the MMI and the MI indices have the advantage of not using the economic in their computation, they both rely on the genotypic and phenotypic data. Different results were obtained by Barbosa and Pinto (1998), where studying the efficiency of selection indices to identify superior clones of potato, the SHI (using the genetic standard deviation as economic weight) and MMI outperformed the MI. In another study, Freitas et al. (2013) studying the genetic gain evaluated with selection indices (WILLIAMS, 1962; PEŠEK and BAKER, 1969; Mulamba and Mock and Smith and Hazel indices) and with REML/Blup in popcorn observed that among the selection indices tested, the MMI provided the best results for selection of full-sib progenies and was outperformed only by the REML/Blup method. Júnior et al. (2010) studying selection two indices in a popcorn population from a fourth cycle of a recurrent selection program, observed that the MMI obtained better results than the Smith and Hazel index (using as economic weights: genetic standard deviation, the ratio CVg/CVe where CVe is the coefficient of experimental variation, heritability and tentatively assigned weights with magnitudes of 1, 10, 20, 1, 100, 100, 1, 1, 1, 1, 1, 15, 25, and 15, for the 14 traits evaluated).

Like the selection indices, the advantage of the SVM use in the sugarcane families selection process is that it doesn't demand to weigh the plants in the field which may improve the selection process (PETERNELLI et al., 2018; MOREIRA et al., 2021) by simplifying the harvest and harvest costs.

For this study, we conclude:

Support vector machines showed to be a good approach to select sugarcane families by learning from the data via multivariate simulation, having outperformed the selection

indices to select sugarcane families in most scenarios tested in this study. We need to emphasize that, in this case, the support vector machines were used to select sugarcane families from the same experiment.

Among the selection indices, the Mulamba and Mock and Smith and Hazel indices showed the best and worst performances to select sugarcane families, respectively.

3. SECOND CASE STUDY

3.1 MATERIAL AND METHODS

3.2 Dataset

The dataset for this study came from Moreira et al. (2021), a study related to the sugarcane breeding programme developed at the Federal University of Viçosa, MG, Brazil, and conducted at the Center for Sugarcane Research and Breeding, Oratórios, Minas Gerais (20°25'S, 42°48'W, 494 m of altitude).

In that study, the authors conducted five experiments, in each experiment 22 sugarcane families were evaluated. The explanatory traits were as follows: the total number of stalks per plot (NS), stalks diameter (SD, measured in centimeters), and stalk height (SH, measured in in meters). The response trait was the TSH (tons of stalks per hectare).

3.3 Selection indices approach

The analysis was conducted on the Selegen software (RESENDE, 2002), the mixed model used was $\boldsymbol{y} = \boldsymbol{Xr} + \boldsymbol{Zg} + \boldsymbol{Wb} + \boldsymbol{e}$, where: $\boldsymbol{y}$ is the observations vector $\left(\boldsymbol{y} \sim \boldsymbol{N}(\boldsymbol{X_r}, \boldsymbol{V})\right)$; $\boldsymbol{r}$ is the vector of replications effect (assumed fixed) summed to the overall mean; $\boldsymbol{g}$ is the vector of sugarcane families effects (assumed random), $\boldsymbol{g} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$, where $\boldsymbol{G}$ is the genetic covariance matrix of the families $\left(\boldsymbol{G} = \boldsymbol{I}\boldsymbol{\sigma}_{\boldsymbol{g}}^{\boldsymbol{2}}\right)$; $\boldsymbol{b}$ is the vector of blocks effects (assumed random) where $\boldsymbol{b} \sim N\left(\boldsymbol{0}, \boldsymbol{I}\boldsymbol{\sigma}_{\boldsymbol{b}}^{\boldsymbol{2}}\right)$, and $\boldsymbol{e}$ is the vector of residuals, $\boldsymbol{e} \sim (\boldsymbol{0}, \boldsymbol{R})$, where $\boldsymbol{R}$ is the residual covariance matrix ($\boldsymbol{R} = \boldsymbol{I}\boldsymbol{\sigma}_{\boldsymbol{e}}^{\boldsymbol{2}}$). $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ represent the incidence matrices of the corresponding effects.

We estimated the broad sense heritability (h^2), genotypic variance (σ_g^2), genetic coefficient of variation (CV_g), where: $CV_g(\%) = 100\frac{\sigma_g}{\bar{X}}$ is the genotypic standard deviation and $\bar{X}$ is the overall genotypic mean. All the selection indices were computed according to Pedrozo et al. (2009).

The Smith-Hazel (SMITH, 1936; HAZEL, 1943), multiplicative (SUBANDI et al., 1973) and the Mulamba and Mock's (MULAMBA; MOCK, 1978) selection indices were used to select sugarcane families for TSH based on the indirect traits number of stalks, stalks diameter, and stalks height. All the selection indices were computed according to Pedrozo et al. (2009). The Smith-Hazel index (SHI) is given by:

$$SHI = (w_{NS}\ x\ NS)(PGV\ x\ NS) + (w_{SD}\ x\ SD)(PGV\ x\ SD) + (w_{SH}\ x\ SH)(PGV\ x\ SH),$$

where:

PGV is the predicted genotypic value, $NS =$ number of stalks, $SD =$ stalks diameter, and $SH =$ stalks height; w_{NS} is the NS economic weight, the same for w_{SD} and w_{SH}.

For the SHI we tested the following economic weights: genotypic standard deviation, genotypic coefficient of variation, and the broad sense heritability (COSTA et al., 2008; ALMEIDA et al., 2014).

The multiplicative index (MI) is given by:

$$MI = (PGV\ x\ NS)(PGV\ x\ SD)(PGV\ x\ SH).$$

The Mulamba and Mock's index is based on the sum of ranks**.** Initially, it ranks the genotypes for each trait by assigning higher absolute values to those of better performance. Then, the values assigned to each trait are summed to obtain the rank sum, indicating the genotypes' classification (CRUZ; CARNEIRO, 2003). The smaller the sum, the better the performance of a genotype for the various traits (ALMEIDA et al., 2014).

The Mulamba and Mock's index (MMI) is given by:

$$MMI = (r\ x\ PGV\ x\ NS) + (r\ x\ PGV\ x\ SD) + (r\ x\ PGV\ x\ SH),$$

where: r is the genotype's rank. For all the selection indices, the selection was performed to favour families with higher NS, SD and SH. We adopted a selection percentage of 18% of the top families, i.e., from the evaluated 22 families (in each of the five experiments). We adopted a selection percentage of 18% of the top families for the selection process. We predefined the genotypic values for family means of TSH (tons of stalks per hectare) as the best selection method. Thus, to evaluate the selection indices and support vector machines' performance, we computed the coincidence coefficient (CC) between the genotypic values for family means of TSH with each selection indices and with the support vector machines.

The coincidence coefficient (CC) was computed according to the following formula:

$$CC = \frac{A}{B}, \text{ where:}$$

A = number of families selected simultaneously by both selection methods involved in each computation (selection indices, support vector machines, and genotypic values for family means of tons of stalks per hectare).

B= the total number of families which pretend to be selected

3.4 Selection via support vector machines

The support vector machines (SVM) classifier performs binary classification, i.e., it separates a set of training vectors for two different classes $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$, where $x_i \in R^d$ denotes vectors in a d-dimensional feature space and $y_i \in \{-1, +1\}$ is a class label (HUANG et al., 2017). The SVM model is generated by mapping the input vectors onto a new higher dimensional feature space denoted as $\Phi: R^d \rightarrow H^f$, where $d < f$. Then an optimal separating hyperplane in the new feature space is constructed by a kernel function $K(x_i, x_j)$, which is the product of input vectors x_i and x_j where $K(x_i, x_j) = \Phi(x_i).\Phi(x_j)$ (HUANG et al., 2017). A kernel is a function that quantifies the similarities between two observations (JAMES et al., 2013), we used the *radial basis function* (rbf) kernel, where $K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2}$, and $\gamma > 0$ (PEDREGOSA et al., 2011). The rbf kernel has good performance in nonlinearly separable problems and shows good performance in most cases (GÉRON, 2019).

We used the two parameters: C and γ, C (penalty parameter) controls the regularisation, a low C allows to have a reduced margin in the hyperplane, for γ, a higher value tends to overfitting (HARRISON, 2020). We tested for γ the values: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and for C: 0.25, 0.5, 0.75, 1. For the SVM, the explanatory traits were as follows: number of stalks (NS), stalk diameter (SD) and stalk height (SH), the response trait was the tons of stalks per hectare (TSH), the selection criterion was to select only sugarcane families with a production of TSH higher than the overall mean, a value of one was assigned in case of selection and zero otherwise. To improve the SVM performance, we initially standardized the explanatory traits by $x'_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}$, where x'_{ij} is the standardized trait value, x_{ij} is the original trait value and S_j is the trait standard deviation.

We also produced synthetic data via multivariate simulation to improve the SVM training performance, as we only had 22 sugarcane families in each experiment, a number of families insufficient to train the SVM model, this procedure was also performed by Peternelli et al. (2018) and Moreira et al. (2021). To generate the synthetic data, we performed a simulation based on the covariance matrix $\boldsymbol{\Sigma}$ (positive definite) of the variables NS, SD, SH, and TSH. The Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma}$ was used to generate $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}^{\boldsymbol{T}}$, where $\boldsymbol{C}$ is a lower triangular matrix $m\,x\,m$ which is the Cholesky factor. A normal multivariate vector $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{C}\boldsymbol{Z}$ was

simulated, where $\boldsymbol{\mu}$ is the mean vector of the variables (NS, SD, SH and TSH), $\boldsymbol{C}$ is the Cholesky factor from the covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{Z}$ is a vector of random independent and identically distributed (*iid*) variables with a standard normal distribution. Through this procedure, we generated 1000 row vectors of the type $[\boldsymbol{X_{i1}}, \boldsymbol{X_{i2}}, \boldsymbol{X_{i3}}, \boldsymbol{X_{i4}}]$, where $\boldsymbol{X_{ij}}$ ($i = 1$ to 1000, and $j = 1$ to 4) represents the simulated value of the variable (NS, SD, SH and TSH) for the individual j. This algorithm assures that all the variables have a covariance matrix $\boldsymbol{\Sigma}$ and mean vector $\boldsymbol{\mu}$ (CRESSIE, 1993; HAINING, 2005).
The simulation was conducted in each experiment separately. Details on the simulation performed are presented in table 5.

Table 5- Illustration of the training and predictions observations performed in each experiment

Exp.	Nr. of families	Simulation	Training	Predictions $(n = 88)$
1	22	1000 families	1000 families	Experiments: 2, 3, 4 and 5
2	22	1000 families	1000 families	Experiments: 1, 3, 4 and 5
3	22	1000 families	1000 families	Experiments: 1, 2, 4 and 5
4	22	1000 families	1000 families	Experiments: 1, 2, 3 and 5
5	22	1000 families	1000 families	Experiments: 1, 2, 3 and 4

Exp.: experiment, Nr. of families: number of families

In this study, for the selection via SVM, the selected families were ranked based on their decreasing probability of being classified as selected. Regarding the support vector machine (SVM) best parameters obtained via grid search, in each scenario are presented in Table 6.

Table 6- Parameters and best parameter values (BPV) in each scenario, using the support vector classifier, available on the scikitlearn Python package

Scenarios	Parameter	BPV
SVM1	γ	0.25
	C	0.25
SVM2	γ	0.6
	C	0.25
SVM3	γ	0.75
	C	0.4
SVM4	γ	0.3
	C	0.25
SVM5	γ	0.7
	C	1

SVC: support vector classifier, present on the scikit-learn Python package; C: penalty parameter; γ: penalization parameter; SVM1: support vector machines, trained on experiment 1 and evaluated on the experiments 2, 3, 4, and 5; similar interpretations are valid for SM2, SVM3, SVM4, and SVM5.

3. 5 RESULTS AND DISCUSSION

Results on broad sense heritability, overall genotypic mean, genotypic standard deviation, and coefficient of genetic variation, for all the traits, in the five experiments are presented in Table 7.

The broad sense heritability measures the reliability of the measured phenotype value in predicting the true genotypic value (Almeida et al., 2014). Its values varied among all the traits in all the experiments. Heritability is a feature that depends not only on a unique trait but also on the population, the environmental conditions in which the individuals are involved, and the phenotype's measurement conditions (Falconer and Mackay, 1997).

Table 7- Broad sense heritability (h^2), overall genotypic mean ($\bar{x}$), genotypic standard deviation (S_g), and coefficient of genetic variation (CV_g) for the evaluated traits number of stalks (NS), stalks diameter (SD), stalks height (SH), and TSH (tons of stalks per hectare) for each of the five experiments

	Parameters	Evaluated traits			
		NS	SH	SD	TSH
Experiment 1	h^2	0.46	0.50	0.16	0.48
	S_g	22.17	0.17	0.76	24.37
	$\bar{x}$	108.35	2.48	25.20	109.14
	$CV_g(\%)$	20.46	6.97	2.99	22.32
Experiment 2	h^2	0.58	0.61	0.41	0.69
	S_g	23.33	0.23	3.97	33.70
	$\bar{x}$	107.59	2.46	25.35	111.30
	$CV_g(\%)$	21.68	9.15	4.95	30.28
Experiment 3	h^2	0.22	0.44	0.39	0.27
	S_g	11.85	0.14	1.02	15.14
	$\bar{x}$	114.45	2.53	25.28	110.59
	$CV_g(\%)$	10.35	5.47	4.02	13.68
Experiment 4	h^2	0.38	0.55	0.58	0.50
	S_g	20.56	0.21	1.69	26.23
	$\bar{x}$	104.95	2.53	25.23	104.40
	$CV_g(\%)$	19.58	8.27	6.71	25.12
Experiment 5	h^2	0.49	0.59	0.38	0.56
	S_g	20.19	0.22	1.15	27.54
	$\bar{x}$	105.86	2.58	25.28	107.12
	$CV_g(\%)$	19.06	8.44	4.54	25.70

Regarding the genetic coefficient of variation (CVg), the values were below 30% for all the traits except in experiment 2 (30.28% for TSH). High CVg can be interpreted as low experimental precision affecting the inferences that can be made for the observed traits (BARBOSA et al., 2005). In Table 8, we present the coincidence coefficient of the selected families by the selection indices and by the genotypic values for family means of TSH (tons of stalks per hectare).

Observing the coincidence coefficient (CC) among the selection indices, their performance varied among the five experiments. The lowest CC values were observed between SHI^{sd} and MMI (0.25) in experiment 3. It is essential to observe the CC between the SHI and TSH depended a lot on the trait used as economic weight in all the experiments.

Table 8- Coincidence coefficient for the sugarcane families selected by the Smith-Hazel (SHI), multiplicative (MI), Mulamba and Mock's (MMI) indices, and by the genotypic values for family means of TSH in each of the five experiments

		MI	MMI	TSH
Experiment 1	SHI	$0.75^{h,sd}$, 1^{cv}	0.5^{cv}, 0.75^{h},	$0.5^{cv,sd}$, 0.75^{h}
	MI		0.5	0.5
	MMI			0.75
Experiment 2	SHI	0.75^{sd}, $1^{h,cv}$	$0.75^{sd,}$ $1^{h,cv}$	0.75^{sd}, $1^{h,\ cv}$
	MI		1	1
	MMI			1
Experiment 3	SHI	0.5^{sd}, 0.75^{h}, 1^{cv}	0.25^{sd}, $0.75^{h,cv}$	0.5^{sd}, 0.75^{h}, 1^{cv}
	MI		0.75	1
	MMI			0.75
Experiment 4	SHI	0.5^{sd}, 0.75^{h}, 1^{cv}	0.5^{sd}, 0.75^{h}, 1^{cv}	0.5^{sd}, 0.75^{cv},1^{h}
	MI		1	0.75
	MMI			0.75
Experiment 5	SHI	0.75^{sd}, $1^{h,cv}$	0.75^{sd}, $1^{h,cv}$	$0.75^{h,cv,sd}$
	MI		1	0.75
	MMI			0.75

h stands for SHI using as economic weight the broad sense heritability; cv for SHI using as economic weight the genetic coefficient of variation; sd for SHI using as economic weight the genetic standard deviation; TSH is the tons of stalks per hectare

In general, the SHI using the broad sense heritability as weight presented the best performance, as it presented the highest CC values with the genotypic values for family means of TSH in 80% of the experiments.

On the contrary, the SHI using the genetic standard deviation as the economic weight had the worst performance. MI and MMI were practically similar in selection efficiency, as they presented the highest CC values with the genotypic values for family means of TSH in 60% of the experiments. The higher the CC between two selection indices, the more similar their results will be (PEDROZO et al., 2009). These results differ from the ones Pedrozo et al. (2009) obtained, where the MI had better performance in selecting sugarcane genotypes than the other two selection indices used in our study. In an experiment with sugarcane families in the T1 stage, the best results were achieved by the SHI (Smith and Hazel index) and MMI (Mulamba and Mock's index) indices (ALMEIDA et al., 2014). In a similar study in popcorn, the MMI provided the best results for selecting full-sib progenies (FREITAS et al., 2013)

compared to the Smith and Hazel index. Mulamba and Mock's index was also the best selection index to select cotton genotypes among the Smith and Hazel. Table 9 illustrates the selection indices' coincidence coefficients related to the SVM.

Table 9. Coincidence coefficient for the families selected by the Smith-Hazel (SHI), multiplicative (MI), Mulamba and Mocks (MMI) indices, genotypic values for family means of TSH and by the SVM for the sugarcane families in each of the five experiments

Evaluation	SI	SVM2	SVM3	SVM4	SVM5
	SHI^{h}	0.25	0.75	0	0.25
	SHI^{cv}	0	0.75	0	0
Experiment 1	SHI^{sd}	0	0.5	0	0
	MI	0	0.75	0	0
	MMI	0.5	0.75	0	0.5
	TSH	0.5	0.5	0	0.5
	SI	SVM1	SVM3	SVM4	SVM5
	SHI^{h}	0.5	0.5	0	0.25
Experiment 2	SHI^{cv}	0.5	0.5	0	0.25
	SHI^{sd}	0.75	0.75	0	0.5
	MI	0.5	0.5	0	0.25
	MMI	0.5	0.5	0	0.25
	TSH	0.5	0.5	0	0.25
	SI	SVM1	SVM2	SVM4	SVM5
	SHI^{h}	0.75	0.5	0	0.5
	SHI^{cv}	0.75	0.5	0	0.5
Experiment 3	SHI^{sd}	0.5	0.25	0	0.25
	MI	0.75	0.5	0	0.5
	MMI	0.5	0.5	0	0.5
	TSH	0.75	0.5	0	0.5
	SI	SVM1	SVM2	SVM3	SVM5
	SHI^{h}	0.5	0.25	0.25	0.25
	SHI^{cv}	0.75	0.5	0.25	0.5
Experiment 4	SHI^{sd}	0.75	0.75	0.5	0.5
	MI	0.75	0.5	0.25	0.5
	MMI	0.75	0.5	0.25	0.5
	TSH	0.5	0.25	0.25	0.25
	SI	SVM1	SVM2	SVM3	SVM4
	SHI^{h}	0.5	0	0.75	0
	SHI^{cv}	0.5	0	0.75	0
Experiment 5	SHI^{sd}	0.5	0.25	0.75	0
	MI	0.5	0	0.75	0
	MMI	0.5	0	0.75	0
	TSH	0.5	0	0.75	0

SI: selection indices; TSH: tons of sugarcane per hectare, SHI^{cv}: Smith and Hazel index (SHI) using as economic weight the genetic coefficient of variation, SHI^{sd}: SHI using as economic weight the genetic standard deviation, SHI^{h}: SHI using as economic weight the broad sense heritability, SVM2:support vector machines, trained on the experiment 2 and evaluated on the experiments 1, 3, 4 and 5. Similar interpretations are valid for SM1, SVM3, SVM4 and SVM5.

Observing the CC (coincidence coefficient) of the SVM with the genotypic values for family means of TSH (Table 9), it is noticeable that its values varied across the experiments. For example, for experiment 1, the CC of SHI^{h}, MMI (0.75) with the

genotypic values for family means of TSH presented higher values when compared to SVM (SVM2, SVM3, SVM4 and SVM5). This means that the SVM models had worst performance than SHI^h and MMI. Only in experiments 3 and 5 different behavior was observed. In experiments 3 and 5, a CC value of 0.75 was achieved by SVM1 and SVM3, respectively. SHI^{cv} also observed the same CC values, MI in experiment 3 and by all selection indices in experiment 5. In some cases, no sugarcane family was simultaneously selected by the SVM model, and the genotypic values for family mean TSH (CC of zero).

SVM4 had the worst performance among the five machine learning models evaluated (having achieved a CC of zero with all the other selection indices and genotypic values for family means of TSH). As said before, SVM4 was trained based on the simulation data from experiment 4 and was evaluated in the other four experiments. In the study conducted by Moreira et al. (2021) on average, 98% of the sugarcane families selected by the BLUPIS method (considered as ideal) were also selected by the SVM classifier, however, there´s a difference in the methodology used. In our study, the sugarcane families were ranked by the SVM based on the decreasing probability of being classified as selected and the evaluation was performed separately for each experiment. On the other hand, the selection indices used in this study depends on the genotypic values to rank the families, another factor to be considered is the small sample size (22 families for each experiment) used to estimate the correlation matrix needed to perform the dataset simulation, this fact may have affected the SVM training. Despite the use of SVM in the sugarcane selection process allows to simplify the harvest and reduces its costs (PETERNELLI et al., 2018; MOREIRA et al., 2021), in our study the SVM had worse performance than the selection indices, mainly when compared to Smith and Hazel index using as economic weight the broad sense heritability (SHI^h).

The SVM such as any other machine learning models, its performance depends strongly on the training data. In a study to classify twenty grapevine varieties (GUTIÉRREZ et al., 2015), artificial neural networks outperformed SVM, however, in the same study, both machine learning models presented similar performances in classifying five grapevine varieties. In a similar study to select sugarcane families using SVM, random forests logistic regression, k-nearest neighbour and artificial neural networks, SVM outperformed all the other machine learning models (MOREIRA et al., 2021). For the specific case of the Smith and Hazel index which uses economic

weights in the selection process, it is difficult to express the economic value of traits, (JAHUFER and CASLER, 2015) and different economic weights result in different selections efficiencies as also verified by Almeida et. al (2014) in sugarcane.

In this study, we conclude:

In general, the Smith and Hazel index using the broad sense heritability as the economic weight was the best among all the selection indices and support vector machines to classify sugarcane families.

For the Smith and Hazel index, the traits used as different economic weight affected differently the selection index performance.

4. CONCLUSIONS

For the dataset from Ferreira et al. (2022), support vector machines showed to be a better approach to select sugarcane families by learning from the data via multivariate simulation. Using similar methodology on the dataset from Moreira et al. (2021), lower performance for support vector machines was obtained, probably due to the smaller sample size used to estimate the correlation matrix, impacting on the dataset simulation used to train the support vector machines.

In all the studies, the selection indices showed different performance, specifically for the Smith and Hazel index which uses economic weights, using different economics weights resulted in different performances.

5. REFERENCES

ALMEIDA, L. M.; VIANA, A. P.; AMARAL, J. AT; JÚNIOR, C. Breeding full-sib families of sugar cane using selection index. **Ciência Rural**, v. 44, p. 605–611, 2014. https://doi.org/10.1590/S0103-84782014000400005

BÁRBARO, I. M., CENTURION, M. A. P. C.; MAURO, A. O. D.; UNÊDA-TREVISOLI, S. H.; COSTA, M. M. Comparação de estratégias de seleção no melhoramento de populações F5 de soja. Revista Ceres, v. 54, p. 250–261, 2007.

BARBOSA, M. H. P.; PINTO, C. A. B. P. Eficiência de índices de seleção na identificação de clones superiores de batata. **Pesquisa Agropecuária Brasileira**, v. 33, n. 33, p. 149-156, 1998.

BARBOSA, M. H. P.; RESENDE, M. D. V.; BRESSIANI, J. A.; SILVEIRA, C. I.; PETERNELLI, L. A. Selection of sugarcane families and parents by Reml/Blup. **Crop Breeding and Applied Biotechnology**, v. 5, p. 443-450, 2005. http://www.alice.cnptia.embrapa.br/alice/handle/doc/316265

BORDONAL, R. O.; CARVALHO, J. L. N.; LAL, R.; FIGUEIREDO, E. B.; OLIVEIRA, B. G.; SCALA JR, N. L. Sustainability of sugar cane production in Brazil. A review. **Agronomy for Sustainable Development**, v. 13, p. 1-23, 2018. https://doi.org/10.1007/s13593-018-0490-x

CERÓN-ROJAS, J. J.; CROSSA, J.; SAHAGÚN-CASTELLANOS, J.; CASTILLO-GONZÁLEZ, F.; SANTACRUZ-VARELA, A. A Selection Index Method Based on Eigenanalysis. **Crop Science**, v. 46, p. 1711-1721, 2006. 10.2135/cropsci2005.11-0420

COMPANHIA NACIONAL DE ABASTECIMENTO (CONAB). **Acompanhamento da safra brasileira de cana-de-açúcar**. Primeiro levantamento da Safra 2021/2022. CONAB, 2021

COUTINHO, G.; PIO, R; SOUZA, F. B. M.; FARIAS, D. H.; BRUZI, A. T.; GUIMARÃES, P. H. S. Multivariate analysis and selection indices to identify superior quince cultivars for cultivation in the tropics. **HortScience**, v. 54, p. 1324-1329. 2019. https://doi.org/10.21273/HORTSCI14004-19

CRESSIE, N. A. C. Statistics for spatial data. New York: John Wiley & Sons, 1993.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos Biométricos Aplicados ao Melhoramento Genético.** Viçosa: Editora UFV, 2003.

ENTRINGER, G. C., VETORAZZI, J. C. F.; SANTOS, E. A., PEREIRA, M. G., VIANA, A. P. Genetic gain estimates and selection of S1 progenies based on selection indices and REML/BLUP in super sweet corn. **Australian Journal of Crop Science**, v. 10, p. 411–417, 2016. 10.21475/ajcs.2016.10.03.p7248

FERREIRA, P. H. S.; GONÇALVES, M. T. V.; TEIXEIRA, G.; PAULA, F. M.; OLIVEIRA, R. L.; BARBOSA, M. H. P.; PETERNELLI, L. A. Comparison of family selection methodologies used in the initial phase of sugarcane breeding. **Crop Science**, v. 62, p. 679–689, 2022. https://doi.org/10.1002/csc2.20685

FREIRIA, G. H.; PERINI, L. J.; ZEFFA, D. M.; NOVAIS, P. S., LIMA, W. F., GONÇALVES, L. S. A.; PRETE, C. E. C. Comparison of non-parametric indexes to select soybean genotypes obtained by recurrent selection. **Semina: Ciências Agrarias**, v. 40, p. 1761-1774, 2019. http://dx.doi.org/10.5433/1679-0359.2019v40n5p1761

FREITAS, I. L. J.; JUNIOR, A. T. A.; VIANA, A. P.; PENA, G. F.; CABRAL, P. S.; VITTORAZZI, C.; SILVA, T. R. C. Ganho genético avaliado com índices de seleção e com REML/Blup em milho-pipoca. **Pesquisa Agropecuária Brasileira**, v. 48, n. 11, p. 1464-1471, 2013. 10.1590/S0100-204X2013001100007

FALCONER, D. S.; MACKAY, T. F**. Introduction to quantitative genetics**. Edinburgh: Longman, 1997.

GÉRON, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. Rio de Janeiro: Alta Books Editora, 2019.

GESTEIRA, G. S.; BRUZI, A. T.; ZITO, R. K.; FRONZA, V.; ARANTES, N. E. Selection of early soybean inbred lines using multiple indices. **Crop Science**, v. 58, p. 2494-2502, 2018. 10.2135/cropsci2018.05.0295

GUTIÉRREZ, S.; TARDAGUILA, J.; FERNÁNDEZ NOVALES, J.; DIAGO, M. P. 2015. Support Vector Machine and Artificial Neural Network Models for the Classification of Grapevine Varieties Using a Portable NIR Spectrophotometer. **PLoS ONE**, v. 10, p. 1-15, 2015. 10.1371/journal.pone.0143197

HAINING, R. **Spatial Data Analysis-Theory and Practice**. Cambridge: Cambridge University Press, 2005.

HARRISON, M. **Machine Learning-Guia de Referência Rápida**. São Paulo: Novatec Editora Ltda, 2020.

HAZEL, L. N. The genetic basis for constructing selection indexes. **Genetics**, v. 28, p. 476-490, 1943

HMEIDI, I.; HAWASHIN, B.; EL-QAWASMEH, E. Performance of KNN and SVM classifiers on full word Arabic articles. **Advanced Engineering Informatics**, v. 22, p. 106–111, 2008. https://doi.org/10.1016/j.aei.2007.12.001

HUANG, MW; CHEN, C.-W.; LIN, WC; KE, S.-W.; TSAI, C. F. SVM and SVM Ensembles in Breast Cancer Prediction. **PLoS ONE**, v. 12, n. 1, p. 1-14, 2017. 10.1371/journal.pone.0161501

JAHUFER, M. Z. Z.; CASLER, M. D. Application of the Smith-Hazel Selection Index for Improving Biomass Yield and Quality of Switchgrass. **Crop Science**, v. 55, 2015. https://doi.org/10.2135/cropsci2014.08.0575

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning: With applications in R**. New York: Springer, 2013.

JÚNIOR, A. T. A.; JÚNIOR, S. P. F.; RANGEL, R. M.; PENA, G. F.; RIBEIRO, R. M.; MORAIS, R. C.; SCHUELTER, A. R. Improvement of a popcorn population using selection indexes from a fourth cycle of recurrent selection program carried out in two different environments. **Genetics and Molecular Research**, v. 9, n. 1, 340-347, 2010

MARINHO, C. D.; GRAVINHA, G. A.; SEBASTIÃO, L. C. A.; ALMEIDA, N. C.; DAHER, R. F.; BRASILEIRO, B. P.; PAULA, T. O. M; AMARAL J. A. T. Indexes in the comparison of pre-commercial genotypes of common bean. **Ciência Rural**, v. 44, p. 1159-1165, 2014. https://doi.org/10.1590/0103-8478cr20121155

MENDES, F. F.; RAMALHO, M. A. P.; ABREU, A. F. B. Índice de seleção para escolha de populações segregantes do feijoeiro-comum. **Pesquisa Agropecuária Brasileira**, v. 44, 1312-1318, 2009. https://doi.org/10.1590/S0100-204X2009001000015

MOREIRA, E. F. A; BARBOSA, M. H. P.; PETERNELLI, L. A. Can statistical learning models make early selection among sugar cane families easier and still efficient? **Crop Science**, v. 61, p. 456-465, 2021. 10.1002/csc2.20334

MULAMBA, N. N.; MOCK, J. J. Improvement of yield potential of the Eto Blanco maize (Zea mays L.) population by breeding for plant traits. **Egyptian Journal of Genetics and Cytology**, v. 7, p. 40-51, 1978.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.

PEDROZO, C. A.; BENITES, F. R. G.; BARBOSA, M. H. P.; RESENDE, M. D. V.; DA SILVA, F. L. Eficiência de índices de seleção utilizando a metodologia reml/blup no melhoramento da cana-de-açúcar. **Scientia Agraria**, v. 10, n. 1, p. 031-036, 2009. http://dx.doi.org/10.5380/rsa.v10i1.11711

PEŠEK, J.; BAKER, R. J. Desired improvement in relation to selection indices. **Canadian Journal of Plant Science**, v. 49, p. 803-804, 1969.

PETERNELLI, L. A.; MOREIRA, E. F. A.; NASCIMENTO, M.; CRUZ, C. D. Artificial neural networks and linear discriminant analysis in early selection among sugar cane families. **Crop Breeding and Applied Biotechnology**, v. 17, p. 299-305, 2017. https://doi.org/10.1590/1984-70332017v17n4a46

PETERNELLI, L. A.; BERNARDES, D. P.; BRASILEIRO, B. P.; BARBOSA, M. H. P.; SILVA, R. H. T. Decision Trees as a Tool to Select Sugar cane Families. **American Journal of Plant Sciences**, v. 9, p. 216-230, 2018. https://doi.org/10.4236/ajps.2018.92018

QIN, Y.; KARIMI, H. R.; LI, D.; LUN, S.; ZHANG, A. A Mahalanobis Hyperellipsoidal Learning Machine Class Incremental Learning Algorithm, **Abstract and Applied Analysis**, v. 2014, p. 1-5, 2014. https://doi.org/10.1155/2014/894246

QUINTON, M.; MCMILLAN, I. The effect of index on selection on allele frequencies and future genetic gains when traits are correlated. **Theoretical and Applied Genetics**, v. 93, p. 1335-1342, 1995. 10.1007/BF00223467.

RESENDE, M. D. V. Software Selegen-REML/BLUP. Curitiba: EMBRAPA, 2002

SINGH, R. K.; CHAUDHARY, B. D. **Biometrical Methods in Quantitative Genetic Analysis**. India: Kalyani Publisher, 2007.

SMIRDELE, E. C.; FURTINI, I. V.; SILVA; C. S. C.; BOTELHO, F. B. S.; RESENDE, M. P. M.; BOTELHO, R. T. C.; COLOMBARI F. J. M.; CASTRO, A. P;

UTUMI. Index selection for multiple traits in upland rice progenies. **Revista de Ciências Agrárias**, v. 42, p. 4-12, 2019. https://doi.org/10.19084/RCA18059

SMITH, F. H. A discriminate function for plant selection. **Annals of Eugenics**, v. 7, p. 240-250, 1936. https://doi.org/10.1111/j.1469-1809.1936.tb02143.x

SUBANDI, W.; COMPTON, A.; EMPIG, L. T. Comparison of the efficiencies of selection indices for three traits in two variety crosses of corn. **Crop Science**, v. 13, n. 2, p. 184-186, 1973.

VASCONCELOS, E. S.; FERREIRA, R. P.; CRUZ, C. D.; MOREIRA, A.; FREITAS, R. J. B. Estimativas de ganho genético por diferentes critérios de seleção em genótipos de alfafa. **Revista Ceres**, v.57, p. 205–210, 2010.

VENMUHIL, R.; SASSIKUMAR, D.; VANNIARAJAN, C.; INDIRANI, R. Selection indices for improving the selection efficiency of rice genotypes using grain quality traits. **Electronic Journal of Plant Breeding**, v. 11, p. 543-549, 2020. 10.37992/2020.1102.091