**Brazilian Journal of Agriculture** **v.98, n.1, p. 23 – 37, 2023**
DOI: 10.37856/bja.v98i1.4321

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

Belo Afonso Muetanene[1], Luiz Alexandre Peternelli[2], Policarpo Carneiro[2], Felipe Lopes da Silva[2] Danilo Pereira Barbosa[3], José Ivo Ribeiro Júnior[2]

[1] Universidade Lúrio, Faculdade de Ciências Agrárias, Moçambique. E-mail: floriafonso@gmail.com
[2] Universidade Federal de Viçosa, Viçosa, Minas Gerais State, Brazil. E-mails: peternelli@ufv.br, policarpo@ufv.br, felipe.silva@ufv.br, jivo@ufv.br
[3] Instituto Federal Goiano, Goiás State, Brazil, E-mail: danilo.barbosa@ifgoiano.edu.br

## ABSTRACT

The present study aimed to compare the following selection indices: Smith and Hazel multiplicative, Mulamba and Mock's, and the support vector machines algorithm (SVM) for sugarcane families selection. We considered the genotypic values for family means of the tons of stalks per hectare per family (GVFTSH) as the ideal selection approach to select sugarcane families. We used the dataset from Moreira et al. (2021), in that study, the authors conducted five experiments, in each experiment 22 sugarcane families were evaluated, we constructed the selection indices via a mixed models approach, adopting a selection percentage of 18% of the top families for the selection process. The selection indices were used to conduct an indirect selection of the tons of stalks per hectare per family (TSH) through the total number of stalks per plot (NS), stalks diameter (SD, in centimeters) and stalk height (SH, in meters). For the support vector machines (SVM), the explanatory traits were as follows: number of stalks (NS), stalk diameter (SD) and stalk height (SH), the response trait was the TSH, the selection criterion was to select only sugarcane families with a production of TSH higher than the overall mean. 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. In this study, for the selection via SVM, the selected families were ranked based on their decreasing probability of being classified as selected, and the SVM best parameters were obtained via grid search. In general, the Smith and Hazel index using the broad sense heritability as economic weight presented the best performance, as it presented the highest coincidence coefficient values with the GVFTSH in 80% of the experiments. In our study, the SVM had worse performance than the selection indices, mainly when compared to Smith and Hazel index using the broad sense heritability as economic weight. The lower performance for support vector machines obtained, is 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.

**Keywords**: Synthetic data, indirect selection, yield prediction, machine learning, BLUP

## 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 628 106 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 & 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 & 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 & 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 networks (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 have 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.

## MATERIAL AND METHODS

### *Dataset*

The dataset for this study came from Moreira et al. (2021), a study related to the sugarcane breeding program developed at the Federal University of Viçosa, MG, Brazil, and conducted at the Centre 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 meters). The response trait was the TSH (tons of stalks per hectare).

### *Selection indices approach*

For the selection indices, data 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 $\left(\boldsymbol{G} = \boldsymbol{I\sigma_g^2}\right)$; $\boldsymbol{b}$ is the vector of blocks effects (assumed random) where $\boldsymbol{b} \sim N\left(\boldsymbol{0}, \boldsymbol{I\sigma_b^2}\right)$**,** and $\boldsymbol{e}$ is the vector of residuals, $\boldsymbol{e} \sim (\boldsymbol{0}, \boldsymbol{R})$, where $\boldsymbol{R}$ is the residual covariance matrix ($\boldsymbol{R} = \boldsymbol{I\sigma_e^2}$). $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ represent the incidence matrices of the corresponding effects.

We estimated the broad sense heritability ($h^2$), genotypic variance ($\sigma_g^2$), 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 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 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 favor 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 considered the genotypic values for family means of the tons of stalks per hectare per family (GVFTSH) as the ideal selection approach. Thus, to evaluate the selection indices and support vector machines' performance, we computed the coincidence coefficient (CC) between the GVFTSH with each selection index 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

### *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 $\Phi: R^d \to H^f$, 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 $\gamma$, C (penalty parameter) controls the regularisation, a low C allows to have a reduced margin in the hyperplane, for $\gamma$, a higher value tends to overfit (HARRISON, 2020). We tested for $\gamma$ 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 $\boldsymbol{\Sigma} = \boldsymbol{CC^T}$, 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 $\boldsymbol{j}$. 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 1.

**Table 1.** 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 2.

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

| Scenarios | Parameter | BPV |
|---|---|---|
| SVM1 | $\gamma$ | 0.25 |
| | $C$ | 0.25 |
| SVM2 | $\gamma$ | 0.6 |
| | $C$ | 0.25 |
| SVM3 | $\gamma$ | 0.75 |
| | $C$ | 0.4 |
| SVM4 | $\gamma$ | 0.3 |
| | $C$ | 0.25 |
| SVM5 | $\gamma$ | 0.7 |
| | $C$ | 1 |

SVC: support vector classifier, present on the scikit-learn Python package; C: penalty parameter; $\gamma$: 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.

To compare the selection indices with the support vector machines, we considered the genotypic values for family means of tons of stalks per hectare per family (GVFTSH) as the ideal selection procedure and so, coincidence coefficient was computed involving the three procedures, namely: selection indices, support vector machines and GVFTSH.

## 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 3.

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 & MACKAY, 1997).

**Table 3.** 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 per family) 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 4 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 with the genotypic values for family means of the tons of stalks per hectare per family (GVFTSH) depended a lot on the trait used as economic weight in all the experiments.

**Table 4.** 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 the tons of stalks per hectare per family (GVFTSH) in each of the five experiments.

| | | MI | MMI | GVFTSH |
|---|---|---|---|---|
| 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 Smith and Hazel index using the broad sense heritability as economic weight (SHI$^{h}$) presented the best performance, as it presented the highest CC values with the GVFTSH in 80% of the experiments.

On the contrary, Smith and Hazel using the genetic standard deviation (SHI$^{sd}$) 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 GVFTSH 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. Table 5 illustrates the selection indices' coincidence coefficients related to the SVM with the selection indices and the genotypic values for family means of the tons of stalks per hectare per family (GVFTSH).

Observing the CC (coincidence coefficient) of the SVM with the GVFTSH (Table 5, 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 GVFTSH presented higher values when compared to SVM (SVM2, SVM3, SVM4 and SVM5). This means that the SVM models had worst performance than the $SHI^h$ and MMI. Only in experiments 3 and 5 different behaviour 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 GVFTSH (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 the GVFTSH). 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 ($\text{SHI}^{h}$).

**Table 5.** Coincidence coefficient for the families selected by the Smith-Hazel (SHI), multiplicative (MI), Mulamba and Mocks (MMI) indices, genotypic values for family means of the tons of stalks per hectare per family (GVFTSH) and by the SVM for the sugarcane families in each of the five experiments.

| Evaluation | SI | SVM2 | SVM3 | SVM4 | SVM5 |
|---|---|---|---|---|---|
| Experiment 1 | SHIh | 0.25 | 0.75 | 0 | 0.25 |
| | SHIcv | 0 | 0.75 | 0 | 0 |
| | SHIsd | 0 | 0.5 | 0 | 0 |
| | MI | 0 | 0.75 | 0 | 0 |
| | MMI | 0.5 | 0.75 | 0 | 0.5 |
| | GVFTSH | 0.5 | 0.5 | 0 | 0.5 |
| | SI | SVM1 | SVM3 | SVM4 | SVM5 |
| Experiment 2 | SHIh | 0.5 | 0.5 | 0 | 0.25 |
| | SHIcv | 0.5 | 0.5 | 0 | 0.25 |
| | SHIsd | 0.75 | 0.75 | 0 | 0.5 |
| | MI | 0.5 | 0.5 | 0 | 0.25 |
| | MMI | 0.5 | 0.5 | 0 | 0.25 |
| | GVFTSH | 0.5 | 0.5 | 0 | 0.25 |
| | SI | SVM1 | SVM2 | SVM4 | SVM5 |
| Experiment 3 | SHIh | 0.75 | 0.5 | 0 | 0.5 |
| | SHIcv | 0.75 | 0.5 | 0 | 0.5 |
| | SHIsd | 0.5 | 0.25 | 0 | 0.25 |
| | MI | 0.75 | 0.5 | 0 | 0.5 |
| | MMI | 0.5 | 0.5 | 0 | 0.5 |
| | GVFTSH | 0.75 | 0.5 | 0 | 0.5 |
| | SI | SVM1 | SVM2 | SVM3 | SVM5 |
| Experiment 4 | SHIh | 0.5 | 0.25 | 0.25 | 0.25 |
| | SHIcv | 0.75 | 0.5 | 0.25 | 0.5 |
| | SHIsd | 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 |
| | GVFTSH | 0.5 | 0.25 | 0.25 | 0.25 |
| | SI | SVM1 | SVM2 | SVM3 | SVM4 |
| Experiment 5 | SHIh | 0.5 | 0 | 0.75 | 0 |
| | SHIcv | 0.5 | 0 | 0.75 | 0 |
| | SHIsd | 0.5 | 0.25 | 0.75 | 0 |
| | MI | 0.5 | 0 | 0.75 | 0 |
| | MMI | 0.5 | 0 | 0.75 | 0 |
| | GVFTSH | 0.5 | 0 | 0.75 | 0 |

SI: selection indices; TSH: tons of sugarcane per hectare, $\text{SHI}^{cv}$: Smith and Hazel index (SHI) using as economic weight the genetic coefficient of variation, $\text{SHI}^{sd}$: SHI using as economic weight the genetic standard deviation, $\text{SHI}^{h}$: 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.

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.

## CONCLUSIONS

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 performances, specifically for the Smith and Hazel index which uses economic weights, using different economics weights resulted in different performances.

## REFERENCES

ALMEIDA, L. M.; VIANA, A. P.; AMARAL, J. AT; JÚNIOR, C. 2014. Breeding full-sib families of sugar cane using selection index. **Ciência Rural**, Santa Maria, Rio Grande do Sul, v. 44, p. 605–611. DOI: 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. 2007. Comparação de estratégias de seleção no melhoramento de populações F5 de soja. Revista Ceres, Viçosa, Minas Gerais, v. 54, p. 250–261.

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

BARBOSA, M. H. P.; RESENDE, M. D. V.; BRESSIANI, J. A.; SILVEIRA, C. I.; PETERNELLI, L. A. 2005. Selection of sugarcane families and parents by Reml/Blup. **Crop Breeding and Applied Biotechnology**, Viçosa, Minas Gerais, v. 5, p. 443-450. DOI: 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. 2018. Sustainability of sugar cane production in Brazil. A review. **Agronomy for Sustainable Development**, France, v. 13, p. 1-23. DOI: https://doi.org/10.1007/s13593-018-0490-x

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

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

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

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

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

ENTRINGER, G. C., VETORAZZI, J. C. F.; SANTOS, E. A., PEREIRA, M. G., VIANA, A. P. 2016. Genetic gain estimates and selection of S1 progenies based on selection indices and REML/BLUP in super sweet corn. **Australian Journal of Crop Science**, Australia, v. 10, p. 411–417. DOI: 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. 2022. Comparison of family selection methodologies used in the initial phase of sugarcane breeding. **Crop Science**, USA, v. 62, p. 679–689. DOI: 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. 2019. Comparison of non-parametric indexes to select soybean genotypes obtained by recurrent selection. **Semina: Ciências Agrarias**, Londrina, Paraná, v. 40, p. 1761-1774. DOI: 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. 2013. Ganho genético avaliado com índices de seleção e com REML/Blup em milho-pipoca. **Pesquisa Agropecuária Brasileira**, Brasília, v. 48, n. 11, p. 1464-1471. DOI: 10.1590/S0100-204X2013001100007

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

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

GESTEIRA, G. S.; BRUZI, A. T.; ZITO, R. K.; FRONZA, V.; ARANTES, N. E. 2018. Selection of early soybean inbred lines using multiple indices. **Crop Science**, USA, v. 58, p. 2494-2502. DOI: 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**, USA, v. 10, p. 1-15. DOI: 10.1371/journal.pone.0143197

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

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

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

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

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

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

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

JÚNIOR, A. T. A.; JÚNIOR, S. P. F.; RANGEL, R. M.; PENA, G. F.; RIBEIRO, R. M.; MORAIS, R. C.; SCHUELTER, A. R. 2010. 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**, São Paulo, Brazil, v. 9, n. 1, 340-347.

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. 2014. Indexes in the comparison of pre-commercial genotypes of common bean. **Ciência Rural**, Santa Maria, Rio Grande do Sul, v. 44, p. 1159-1165. DOI: https://doi.org/10.1590/0103-8478cr20121155

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

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

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

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. 2011. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, USA, v. 12, p. 2825-2830.

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

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

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

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

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

QUINTON, M.; MCMILLAN, I. 1995. The effect of index on selection on allele frequencies and future genetic gains when traits are correlated. **Theoretical and Applied Genetics**,

Germany, v. 93, p. 1335-1342. DOI: 10.1007/BF00223467.

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

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

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

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

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

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

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

Received in: April, 30, 2023.
Accepted in: June, 15, 2023.