TALITA ESTÉFANI ZUNINO SANTANA

NEW INSIGHTS ON GENETIC MODELING OF GROWTH AND REPRODUCTIVE TRAITS IN TROPICAL CROSSBRED AND NELLORE CATTLE

Dissertation presented to the Animal Science Graduate Program of the Universidade Federal de Viçosa, in partial fulfillment of the requirements for degree of *Magister Scientiae.*

VIÇOSA
MINAS GERAIS – BRASIL
2019

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
S232n
2019

Santana, Talita Estéfani Zunino, 1993-
New insights on genetic modeling of growth and reproductive traits in tropical crossbred and Nellore cattle / Talita Estéfani Zunino Santana. – Viçosa, MG, 2019.
ix, 53 f. : il. (algumas color.) ; 29 cm.

Texto em inglês.
Orientador: Fabyano Fonseca e Silva.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Bovinos de corte. 2. Teoria bayesiana de decisão estatística. 3. Genômica - Modelos matemáticos. 4. Aprendizado do computador. I. Universidade Federal de Viçosa. Departamento de Zootecnia. Programa de Pós-Graduação em Zootecnia. II. Título.

CDD 22. ed. 636.212

TALITA ESTÉFANI ZUNINO SANTANA

NEW INSIGHTS ON GENETIC MODELING OF GROWTH AND REPRODUCTIVE TRAITS IN TROPICAL CROSSBRED AND NELLORE CATTLE

Dissertation presented to the Animal Science Graduate Program of the Universidade Federal de Viçosa, in partial fulfillment of the requirements for degree of *Magister Scientiae.*

APPROVED: July 19th, 2019.

Luiz Otávio Campos da Silva
(Co-adviser)

Renato Domiciano Silva Rosado

Fabyano Fonseca e Silva
(Adviser)

“Climb the mountain not to plant your flag, but to embrace the challenge, enjoy the air and behold the view. Climb it so you can see the world, not so the world can see you.”

David McCullough Jr

ACKNOWLEDGMENTS

To the Universidade Federal de Viçosa and the Animal Science Department, which had provided me free education with the best professionals in Brazil.

To the National Research Council of Brazil (CNPq), for the scholarship.

To the beef cattle breeding program GENEPLUS – EMBRAPA on behalf of Dr. Paulo Roberto Costa Nobre, for granting the database and support in the execution of the research.

To the Fazenda Bama (J. L. Agropecuária Ltda.), for crossbred cattle data and for accepted me for my required internship undergraduate. I am forever thankful for this opportunity that Ms. Célia provided me.

I would like to express my gratitude to Dr. Fabyano Fonseca e Silva for his mentorship and guidance throughout my master's degree. I value the discussions we had regarding the Bayesian and Machine Learning methods.

To my co-adviser Dr. Luiz Otávio Campos da Silva, for scholarship provided for me in the GENEPLUS program and by encouraging scientific research. I appreciate the time; the efforts promoted and provide insight about the animal breeding.

To my co-advisers Dr. Gilberto Romeiro de Oliveira Menezes and Dr. Mário Luiz Chizzotti, for their collaboration in my dissertation.

To Dr. Renato Domiciano Silva Rosado, for accepting being member of my defense committee.

To Dr. Camila Ferreira Azevedo, for the friendship and theorical foundations for the development of the dissertation.

To the Cleysinho, for the friendship and fundamental contribution in the Machine Learning study.

To my friends from GDMA, for companionship and contributions in the development of the dissertation.

A big thank to my whole family, especially my parents Marcos and Márcia, who are always believing me, for all support and being an example for me, and my uncles André and Tânia, who are always doing their best for helping me.

SUMMARY

ABSTRACT

SANTANA, Talita Estéfani Zunino, M.Sc., Universidade Federal de Viçosa, July, 2019. **New insights on genetic modeling of growth and reproductive traits in tropical crossbred and Nellore cattle.** Adviser: Fabyano Fonseca e Silva. Co-advisers: Luiz Otávio Campos da Silva, Gilberto Romeiro de Oliveira Menezes and Mário Luiz Chizzotti.

In genetic evaluations of farm animals, infinitesimal linear models are frequently assumed, which do not consider source of non-additive and non-linear effects, it might reduce the predictive ability, mainly in populations of crossbred animals. In this context, there have been increasing interest in prediction methods that allow access these effects, above all, without assume statistical presuppositions. For predict breeding values in crossbred populations the key point use methods that allow assess non-additive effects (heterosis, complementarity and epistatic losses). However, these effects are highly correlated and frequently assumed as equally relevant. In the sense, a variable selection model (BayesB) was implemented to estimate non-additive effects as well as obtain breeding values for weaning weight in a population with 16,126 beef cattle corresponding to twenty-six crosses compositions. The BayesB proved to be a powerful method to reduce the estimation problems coming from non-additives covariates, and effects frequently assumed as important (maternal non-additive genetic effects and both breed additive effects are not relevant) were statistical reset, opposing the empirical presets assumed in several studies. In addition to benefits statistical promoted by dimensionality reduction, the BayesB model might reduce computational demand and processing time given that enable estimate non-additive effects and predict breeding values in single step, in other words, without additional analysis as it is currently done. It makes the BayesB model very attractive for application in breeding programs of crossbred beef cattle. On the other hand, in the genome-wide selection field, new statistical methods have been proposed in order to minimize the side effects (high-dimensionality and multicollinearity) coming from simultaneous estimation of SNPs. However, the studies applied to genomic classification with machine learning are few. In the sense, the artificial neural network (ANN) methods have been highlighted, however, scenarios with larger genomic data set analyzed by machine learning

(ML) algorithms, as ANN, imply in an expensive computational processing. For this reason, searching ML algorithms simplest, was proposed a study of genome-enabled classification by several machine learning frameworks for stayability trait in Nellore cattle. In this study, was performed SNPs selection to fitting three sires' genomic data set (one, three and five thousand markers), in order to evaluate the impact of structure data set in the classification of daughters. Moreover, was included biological noise in phenotypes in other to challenge to learning algorithms. In this sense, was verify that ML frameworks simplest, as Naïve Bayes, are better to elaborate methods to solve complex issue of classification.

RESUMO

SANTANA, Talita Estéfani Zunino, M.Sc., Universidade Federal de Viçosa, July, 2019. **Novas percepções em modelagem genética de característica de crescimento e reprodutiva em bovinos cruzados tropicais e Nel,ore**. Adviser: Fabyano Fonseca e Silva. Co-advisers: Luiz Otávio Campos da Silva, Gilberto Romeiro de Oliveira Menezes and Mário Luiz Chizzotti.

Nas avaliações genéticas de animais de produção modelos lineares infinitesimais são frequentemente assumidos, os quais não consideram efeitos de origem não-aditiva e não-linear o que pode reduzir a capacidade preditiva, principalmente em populações de animais cruzados. Neste contexto, há um crescente interesse em métodos de predição que permitem acesso a esses efeitos, sobretudo, sem assumir pressupostos estatísticos. Para predizer valores genéticos em populações cruzadas, o ponto-chave é utilizar métodos que permitem avaliar efeitos não-aditivos (heterose, complementaridade e perdas epistáticas). No entanto, esses efeitos são altamente correlacionados (o que implica em uma condição estatística desfavorável) e frequentemente assumidos como igualmente relevantes. Neste sentido, implementou-se um modelo de seleção de variáveis (BayesB) para estimar efeitos não-aditivos, bem como obter valores genéticos para peso à desmama em uma população com 16.126 bovinos de corte correspondentes a vinte e seis composições de cruzas. O BayesB provou ser um método poderoso para reduzir os problemas de estimativa provenientes de covariáveis não-aditivas, e efeitos comumente assumidos como importantes (efeitos genéticos não-aditivos maternos e ambos efeitos aditivos da raça não são relevantes) foram estatisticamente irrelevantes, o que contrapõe as predefinições empíricas assumidas em vários estudos. Além dos benefícios estatísticos promovidos pela redução de dimensionalidade, o modelo BayesB pode reduzir a demanda computacional e o tempo de processamento por permitir estimar efeitos não-aditivos e predizer valores genéticos em uma única etapa, ou seja, sem analises adicionais como é atualmente realizado. Isto torna o modelo BayesB muito atrativo para aplicação em programas de melhoramento genético de bovinos de corte cruzados. Por outro lado, no campo da seleção genômica ampla, novos métodos estatísticos vêm sendo propostos para minimizar os efeitos

colaterais (alta dimensionalidade e multicolinearidade) advindos da estimativa simultânea de SNPs. No entanto, os estudos aplicados à classificação genômica com aprendizado de máquina são poucos. Neste sentido, os métodos de redes neurais artificiais (RNA) têm tido grande visibilidade, no entanto, cenários com maiores conjuntos de dados genômicos analisados por algoritmos de aprendizado de máquina (ML), como RNA, implicam em um dispendioso processamento computacional. Por esta razão, buscando algoritmos ML mais simples para análise de dados genômicos, foram utilizados os métodos AdaBoost - ADA, Bernoulli Naïve Bayes - NB, Decision Tree - DT, Nearest Neighbors - KN, Multilayer Perceptron – MLP, e Support Vector Machine para Classification (SVC) para a classificação genômica de stayability em bovinos Nelore. Neste estudo foi realizada seleção de SNPs para ajustar diferentes conjuntos dados genômicos proveniente dos touros (mil, três mil e cinco mil marcadores), a fim de avaliar o impacto da estrutura de dados na classificação das filhas. Além disso, foi incluído ruído biológico nos fenótipos a fim de desafiar os algoritmos de aprendizado. Nesse sentido, verificou-se que os métodos de ML mais simples, como Naïve Bayes, são superiores à métodos mais elaborados para resolver questões complexas de classificação.

Chapter 1

1.1 General Introduction

Statistical modelling has been constant object of study among breeders in order to increase the genetic gain by means of selection, mainly, by prediction of breeding value more accurate. In genetic evaluation, both infinitesimal linear models are frequently assumed and, both in crossbred and genomic approaches, the non-additive effects are not included, which might reduce the predictive ability, mainly in crossbred populations (Bertoli et al., 2018). In this context, there has been increasing interest in prediction methods that allow access non-additive effects, including non-linear, without statistical presuppositions.

For crossbred populations, without genomic information, the most elaborate models involve ridge regression (Bertoli et al., 2016) and Bayesian models with heteroscedastic variances and reaction norm (Oliveira et al., 2016). The key point in prediction of breeding value in these individuals is allow estimate non-additive effects by racial composition, in other words, attribute heterozygosity, epistazygosity and complementarity to covariates. However, these covariates are assumed as equally relevant and, since all the covariates are on single information (racial composition), the presence of multicollinearity is naturally observed. Thus, selection variable models without assumed prepositions about genetic covariates, as BayesB (Meuwissen et al., 2001) might provide related to genetic architecture of non-additive effects and improve the prediction quality.

In practice, choose the breeds that will compose crosses is a timeless doubt that permeates around beef cattle farms due to variety of pure and synthetic breeds. For majority of production system, the weaning weight is the main characteristic to evaluate both calf and dam.

Based on this information, chapter 2 described the study conducted in a tropical crossbred population with the goal of implement a BayesB model to estimate genetic covariates as well as infer on the predictive ability and variance components for weaning weight trait.

Continuing with the new statistical approaches, now to genome-enable predictions, artificial neural network (ANN) methods have been highlighted because are admittedly universal approximators of complex functions (Hornik et al., 1989), in the sense, several studies both to prediction and to classification were conducted in livestock field.

However, scenarios with larger genomic data set analyzed by machine learning (ML) algorithms imply in an expensive computational demand in terms of processing time and data storage capacity in the Random Access Memory - RAM, for this reason, the dimensionality reduction or, in ML terms, feature extraction have been strongly encouraged before data modelling (de los Campos et al., 2010). For these reasons, searching ML algorithms simplest, methods as Naïve Bayes (Long et al., 2009), Support Vector Machine (Long et al., 2011), Neuro-Fuzzy (Shahinfar et al., 2012), Random Forest (Mokry et al., 2013) and XgBoost (Li et al., 2018) have been studied.

In this sense, the chapter 3 refers to study conducted genomic information to stayability trait in Nellore cattle, where was performed SNPs selection to fitting three sires' genomic data set (one, three and five thousand markers), in order to classifier the daughters' stayability trait.

1.2 Objectives

1.2.1 Main objective

Implement new statistical approaches for analyzes of prediction and classification of growth and reproductive traits in tropical crossbred and Nellore cattle.

1.2.2 Specific objectives

- Implement BayesB method to modelling non-additive effects from weaning weight trait in tropical crossbred cattle.
- Estimate non-additive effects maternal and direct assumed as covariate in BayesB method.

- Implement machine learning approaches to classifier stayability trait by genomic data set in Nellore cattle
- Perform markers selection
- Assess evaluation metrics based in confusion matrix to compare machine classifiers
- Compare predictive ability of classifiers run with different genomic data set structures

1.3 References

Bertoli, C. D., J. B. Neto, C. McManus, J. A. Cobuci, G. S. Campos, M. L. Piccoli, and V. Roso. 2018. Modelling non-additive genetic effects using ridge regression for an Angus – Nellore crossbred population. Anim. Prod. Sci. 59:823–834. doi:https://doi.org/10.1071/AN17439.

Bertoli, C. D., J. Braccini, and V. M. Roso. 2016. Comparing methodologies to estimate fixed genetic effects and to predict genetic values for an Angus × Nellore cattle population. J. Anim. Sci. 94:500–513. doi:10.2527/jas.2015-9344.

de los Campos, G., D. Gianola, G. J. M. Rosa, K. A. Weigel, and J. Crossa. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. (Camb). 92:295–308. doi:10.1017/S0016672310000285.

Hornik, K., M. Stinchcombr, and H. White. 1989. Multilayer feedforward networks are universal approximators. Neural Networks. 2:359–366.

Li, B., N. Zhang, Y. G. Wang, A. W. George, A. Reverter, and Y. Li. 2018. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front. Genet. 9:1–20. doi:10.3389/fgene.2018.00237.

Long, N., D. Gianola, G. J. M. Rosa, and K. A. Weigel. 2011. Application of support vector regression to genome-assisted prediction of quantitative traits. Theor. Appl. Genet. 123:1065–1074. doi:10.1007/s00122-011-1648-y.

Long, N., D. Gianola, G. J. M. Rosa, K. A. Weigel, and S. Avendaño. 2009. Comparison of classification methods for detecting associations between SNPs and chick mortality. 14:1–14. doi:10.1186/1297-9686-41-18.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819–1829. doi:11290733.

Mokry, F. B., R. H. Higa, M. de Alvarenga Mudadu, A. Oliveira de Lima, S. L. C. Meirelles, M. V. G. Barbosa da Silva, F. F. Cardoso, M. Morgado de Oliveira, I. Urbinati, S. C. Méo Niciura, R. R. Tullio, M. Mello de Alencar, and L. Correia de Almeida Regitano. 2013. Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. BMC Genet. 14:1–11. doi:10.1186/1471-2156-14-47.

Oliveira, M. M., M. L. Santana, and F. F. Cardoso. 2016. Multiple-breed reaction norm animal model accounting for robustness and heteroskedastic in a Nelore-Angus crossed population. Animal. 10:1093–1100. doi:10.1017/S1751731115002815. HORNIK, K., M. Stinchcombr, and H. White. 1989. Multilayer feedforward networks are universal approximators. Neural Networks. 2:359–366.

Shahinfar, S., H. Mehrabani-Yeganeh, C. Lucas, A. Kalhor, M. Kazemian, and K. A. Weigel. 2012. Prediction of breeding values for dairy cattle using artificial neural networks and neuro-fuzzy systems. Comput. Math. Methods Med. 2012. doi:10.1155/2012/127130.

Chapter 2

Bayesian variable selection of non-additive genetic effects for breeding values prediction in tropical crossbred beef cattle

Abstract: Twenty-six Brazilian beef cattle crossbreds were used to investigate additive and non-additive genetic effects on weaning weight under a Bayesian variable selection framework. The proposed model assumed non-additive genetic covariates (direct and maternal breed additive, direct and maternal heterozygosity, direct and maternal complementarity, and direct epistazygosity), additive genetic and systematic environmental effects. Variable selection applied to non-additive genetic covariates revealed that maternal and both breed effects were not relevant in the present study. Posterior means (standard deviations) of direct non-additive genetic effects were 11.68 (7.62), 5.88 (3.77), -6.19 (7.55), respectively for heterozygosity, complementarity and epistazygosity. Additive genetic (159.95 kg²) was slightly higher than non-additive genetic variance (129.36 kg²), reflecting the greater differences of crossbreds in relation to pure Nellore breed. Heavier animals at weaning comes from R(RN), ST(SN), N(AN), ST(AN) and C(RN) crosses (A: Aberdeen Angus, C: Charolais, N: Nellore, S: Senepol, ST: Simmental), with predicted weaning weight equal to 242.16, 239.81, 239.51, 236.41 and 236.03 kg, respectively. These results reinforce the superiority of crossbred in relation to pure Nellore breed (predicted weight equal to 194.80 kg). To the best of our knowledge, there are no previous studies on Bayesian variable selection of non-additive genetic covariates, for which the multicollinearity issues is widely reported. The proposed model was efficient and can be indicated for genetic evaluation of crossbred animals derived from several breeds used in Brazilian breeding programs.

Keywords: heterosis, complementarity, multicollinearity, weaning weight.

2.1 Introduction

Crossbreeding has been widely used to combine the different characteristics for which each breed was previously selected (Falconer and Mackay, 1996), in addition to promoting greater production flexibility in the most diverse management systems. The superiority of crossbred animals over pure breeds is mainly due to non-additive genetic effects (heterosis, epistatic loss and complementarity). However, it is also necessary the prediction of additive genetic effects for all animals in the pedigree (Cunningham and Connolly, 1989; Rodriguez-Almeida et al., 1997).

Non-additive genetics effects are widely estimated through coefficients under a crossbred model, considering the retention of heterosis and epistatic loss as linearly proportional to heterozygosity and epistazygosity, respectively (Gregory and Cundiff, 1980). In this context, many researchers have described high difficulty to jointly estimate these non-genetics effects together due to the presence of multicolinearity. It implies in increasing of the variance estimators and estimates bias.

To work around this issue as well as obtain further accurate estimates in genetic evaluation of crossbred beef cattle, interesting models have been proposed, since linear regression model (Dillard et al., 1980; Arthur et al., 1999; Roso and Fries, 2000) as others, for instance: linear-mixed model considering dominance and additive breed effects as (Rodriguez-Almeida et al., 1997; Roso et al., 2005b), modelling to breed composition (Splan et al., 2002 and Toral et al., 2011), models of ridge regression (Pimentel et al., 2006; Bertoli et al., 2016 and 2018), heterogeneous residual variances (Cardoso et al., 2005) and also measuring environment genotype interaction by reaction norm and heteroscedastic variances (Oliveira et al., 2016). However, in all reported studies, non-genetics are assumed as fixed effects, which are considered as equally relevant and do not are submitted a jointly variable selection toll.

In summary, our hypothesis is based on pointing out for null or very small non-additive genetic effects using Bayesian variable selection approaches. We believe that the proposed framework is feasible to increase

both predictive ability and breeding value accuracy from crossbred models. Faced with a similar problem in genomic-wide selection, Meuwissen et al. (2001) proposed the BayesB model, which assumes variable selection based on an additional parameter π (probability of covariates be equal to zero).

Following this orientation, we proposed BayesB-based model for non-genetic covariates to estimate direct and maternal genetic covariates in the genetics evaluation of tropical crossbred cattle, as well as to jointly infer on the predictive ability and variance components for weaning weight trait.

2.2 Material and methods

2.2.1 Data

Were used weaning weight records of crossbred calves born between 2012 and 2016 on the Bama farm, located at Juara, Mato Grosso state, Brazil. The trait was adjusted to 240 days, adapted to according to Beef Improvement Programs (2010) guidelines. Contemporary groups (CG) were defined as born in the same year and management group. After preliminary edits, the dataset consisted of 16,126 records, with mean 215.25 ± 35.97 kg and 223 CG. Table 1 shows the genetic compositions of the animals. Due to the considerable number of sire unknown (12,266 or 54.4%) in pedigree file, was assumed that different levels of the random effect are uncorrelated, in other words, the relationship matrix is proportional to the identity matrix.

2.2.2 Direct and maternal breed additive and non-additive genetic effects covariates

To estimate the breed additive (deviation from the taurine breeds), heterosis, epistatic loss and complementarity effects were utilized the following coefficients, assumed as covariates: Direct (ad) and maternal (am) breed additive effects were equal to the expected fractions of Nellore genes in the composition of the calf and their dam, respectively, proposed by Rodríguez-almeida et al. (1997). A sum of zebuine x taurine ($H_{zt} = 2\left[\sum_{j=1}^{z} P_j -\right.$

$\left(\sum_{j=1}^{z} P_j\right)^2\right]$) and taurine x taurine ($H_{tt} = \left(\sum_{k=1}^{t} P_k\right)^2 - \sum_{k=1}^{t} P_k^2$) heterozygosity coefficients (Gregory and Cundiff, 1980; Roso and Fries, 2000) were assumed for estimated the heterosis effects, where P_j represent the fraction zebuine and P_k the fraction taurine. For each animal was calculated the direct (hd) and maternal (hm) heterozygosity. As described by Fries et al. (2002), the epistatic losses is proportional to heterozygosity observed on parents, hence the direct epistazygosity (ed) is given by ed = ½ ($H_s + H_d$), where H_s is the sire's heterozygosis and H_d is the dam's heterozygosis (Bertoli et al., 2016). Since the dams are daughters of pure parents, the maternal epistazygosity was not considered. The effects by direct (cd) and maternal (cm) complementarity were defined as cd = ad(1 – ad) and cm = am(1 – am), according (Fries et al., 2000). Table 2 shows the correlation among genetic coefficients (All coefficients obtained for each crossbred are illustrated in Table 3).

2.2.3 Model development

The linear-mixed model fitted was:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_1 a} + \sum_{\mathrm{i=1}}^{\mathrm{na}} \mathbf{Z_2 na} + \mathbf{e}$$

where: $\boldsymbol{y}$ is the vector of phenotypic values; $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are, respectively, the incidence matrix and the correspondent vector of systematic environmental effects (dam's age at calf weaning linear and quadratic, dam's weight at calf weaning, sex and contemporary group); $\boldsymbol{Z_1}$ is the incidence matrix of the random effect; $\boldsymbol{a}$ is the vector additive genetic effects (breeding values); $\boldsymbol{Z_2}$ is the incidence matrix of genetics covariates; $\boldsymbol{na}$ is the vector breed additive and non-additive genetic effects (ad, am, hd, hm, ed, cd and cm) and $\boldsymbol{e}$ is the vector of residuals.

The prior distributions for the systematics and random effects were given by

$$\boldsymbol{\beta}|\mu_\beta, \boldsymbol{I_\beta}\sigma_\beta^2 \sim N(\mu_\beta, \boldsymbol{I}\sigma_\beta^2)$$

$$\boldsymbol{a}|\boldsymbol{I_a}\sigma_a^2 \sim N(0, \boldsymbol{I}\sigma_a^2)$$

Genetic covariates were estimated using BayesB model proposed by Meuwissen et al. (2001). BayesB performs variable selection subjectively

determining π value, the proportion of covariates with null effects, thus the probability of the covariate having effect greater than zero is equal to 1 - π. This model assumes as prior a normal mixture distribution where

$$\boldsymbol{na}|\pi \sim (1-\pi)N(0,\sigma^2_{na}) + \pi N(0,\sigma^2_{na}=0)$$

The prior distributions for the hyperparameters, describe by de los Campos et al. (2009) and Pérez and de los Campos (2013), following

$$\sigma^2_a \sim \chi^{-2}(v_a, s^2_a)$$

$$\sigma^2_{na} \sim X^{-2}(v_{na}=5, S^2_{na}=0{,}04\,)\text{, with probability } (1-\pi)$$

$$\pi \sim \beta(p_0, \pi_0)\text{, with } p_0 > 0 \text{ and } \pi_0 \epsilon\, [0{,}1]$$

$$\sigma^2_e \sim \chi^{-2}(v_e=-2, S^2_e)\text{, where } v_e = 2 \text{ leading to a uniform distribution}$$

Applying Bayes' theorem, the posterior distributions in general formulation is described by

$$P(\beta, a, na, \sigma^2_a, \pi, \sigma^2_{na}, \sigma^2_e|y)\ \alpha\ P(y|\beta, a, na, \sigma^2_a, \pi, \sigma^2_{na}, \sigma^2_e) \times P\left(\beta|\mu_\beta, \sigma^2_\beta\right) \times$$

$$P(a|I\sigma^2_a) \times P\left(\sigma^2_a|v_a, S_a\ \right) \times P(na|\pi\ , \sigma^2_{na}) \times P\left(\pi|\mathrm{p}_0, \pi_0\ \right) \times P\left(\sigma^2_{na}|v_{na}, S_{na}\right) \times$$

$$P\left(\sigma^2_e|v_e, S_e\ \right)$$

2.2.4 Bayesian computation and analysis

To data processing was used Monte Carlo Markov Chain (MCMC) algorithms implemented in the BGLR package (Bayesian Generalized Linear Regression), available in the R software. All the parameters were estimated by Gibbs sampling, except $\boldsymbol{na}$, which requires sampling by Metropolis-Hastings algorithm, once its complete conditional distributions do unknown probability distributions (Chib and Greenberg, 1995).

The MCMC chains length of 200,000 iterations, considering a burn-in period of 50,000 and thinned every 10 iterations were saved. We evaluated, in each chain, the posterior mean, standard deviation and HPD interval 95%. The global convergence was verifying by Geweke test (Geweke, 1992). Furthermore, to breeding values was calculated accuracy, adapted from Mrode (2014), expressed by

$$r_{\hat{a},y} = \frac{cov(\hat{a}, y)}{(\sigma_{\hat{a}}\sigma_y)}$$

Where the correlation $(r_{\hat{a},y})$ between the breeding value predicted $(\hat{a})$ and the phenotypic value observed (y) is equal to covariance between of predicted breeding value and observed phenotypic value above product of predicted breeding value deviation and observed phenotypic value deviation (σ_y).

2.2.5 Crossbred genetic parameters estimate

After identifying the significant genetic covariates, the broad-sense heritability was estimated following

$$H^2 = \frac{\sigma_g^2}{\sigma_p^2}$$

Where the genetic variance (σ_g^2) is given by sum of additive genetic effects (σ_a^2), dominance (interactions between alleles at the same locus) genetic effects (σ_d^2), and epistatic losses (interactions between alleles at different loci) genetic effects (σ_i^2), and phenotypic variance (σ_p^2) is defined by sum of genetic variance (σ_g^2) and residual variance (σ_e^2) estimated.

The association between the pairs of genetic covariates is given by

$$r_{(1,2)} = \frac{\sigma_{na(1,2)}}{\sqrt{\sigma_{na(1)}^2 + \sigma_{na(2)}^2}}$$

Where $\sigma_{na(1,2)}$ represents genetic covariance between pair of non-additive genetic effects.

2.3 Results

2.3.1 Direct and maternal breed additive and non-additive genetic effects estimates

The mean ± SD probability of the genetic covariate effects is equal to zero (π) estimated by BayesB model was equal to 0.38 ± 0.12, which promoted heavy shrinkage around zero to all maternal effects (am, hm and cm) and to direct breed additive effect (ad), evincing the direct non-additive effects (Figure 1 and 2). Direct genetic effect coefficients (non-null effects) were estimated in

28.90 ± 5.53, 35.60 ± 10.7894 and -39.32 ± 7.10 to the heterozygosity (hd), complementarity (cd) and epistazygosity (ed), respectively.

The sum of non-additive genetic effects estimated ranged between 7.69 and 23.35 kg for all the crossbreds studied. Heterosis effects and epistatic losses were higher to tricross animals offspring of sire 3/8 Nellore, followed by tricross crossbred. Nevertheless, in terms of complementarity effect, the crosses F1 stood out.

Considering, only, their non-additive genetic effects, the better crossings obtained were of crosses with Nellore (F1), followed by tricross crossing. The advantaged of F1 owing to absence of epistatic losses effects. The non-additive genetic effects are detailed illustrated in Table 3.

2.3.2 Variance components and genetic parameter estimates

The estimated components of variance and genetic parameter are presents in Table 4 and variance structure to non-additive genetic effects is decompose in Table 5. Additive genetic variance was slightly larger than non-additive genetic variance, in which heterosis and epistatic losses explain similarly the bigger fraction non-additive variance, while complementarity effect represents 11% of non-additive variance, that is a short fraction.

Correlation coefficients show strong positive association between heterosis and no-additive complementarity, and opposed associations between epistatic losses effect and other effects.

Narrow-sense heritability (h^2) estimated is in agreement with studied trait and presented moderate magnitude, while the broad-sense heritability (H^2) approach was estimated high magnitude, due to increase of non-additive genetic variance, given by heterosis and epistatic losses effects.

2.3.3 Breeding values and accuracy

The breeding values mean was estimated in 0.0008 with 0.35 of accuracy and 6.06 standard deviation. However, analyzing each crossbred we note that offspring of dams ½ Senepol ½ Nellore (SN) have negative breeding values ranged -7.85 to -0.37 kg, unlike of most crosses. Accordingly, caution

should be exercised in choosing SN dams when weaning weight is main selection criterion. All EBVs with their SD describe in Table 3.

2.3.4 Better crossbreds

Heavier animals at weaning comes from R(RN), ST(SN), N(AN), ST(AN) and C(RN) crosses (A: Aberdeen Angus, C: Charolais, N: Nellore, S: Senepol, ST: Simmental), with predicted weaning weight (Table 3) in 242.16, 239.81, 239.51, 236.41 and 236.03 kg, respectively. This reinforcing the superiority of crosses animals in relation to pure breed Nellore, since the pure weighed 194.80 kg, in mean.

2.4 Discussion

2.4.1 Direct and maternal breed additive and non-additive genetic effects estimates

BayesB model reveals (π= 0.38) that maternal non-additive genetic covariate and both breed additive covariate are not relevant to genetic evaluation of crossbred animals when direct non-additive covariates are present, opposing, in general, all the studies described so far. In other words, the information maternal and the additive deviation from the taurine breeds are not necessary to identify the genetic differences between purebred and crossbred animals and, therefore, can be excluded from the model. The reduction dimensionality enables a further favorable statistical condition, once excluded covariates have a strong linear correlation (Table 2).

Coefficients estimated to the non-null genetic covariates were greater than obtained using mixed model by Bueno et al. (2012) to weaning weight 205 days, per ridge regression by Bertoli et al. (2018) to weaning gain and too by Lopes et al. (2010) to yearling weight. Possibly this result was achieved due to a more leptokurtotic distribution assumed to genetic covariates, evidencing significance direct non-additive effects. According to with study by Meuwissen et al. (2001), the authors describe that better results are obtained when

variance components different are assumed for each covariate, comparing with a normal distribution with small variance (Bayesian Ridge Regression).

The great improvement outcome of reduction dimensionality is how much lower the number of correlated linearly effects to be estimated, further accurate is the estimation process considering the same number of records (Meuwissen et al., 2001) and lower computational demand is required. Moreover, from Bayes' Theorem, when an informative prior is assumed, parsimonious models promote a more appropriate Bayesian learning (learn from the data).

The increase in superiority in offspring is denominated hybrid vigor or heterosis. Thus, the adaptive power lost with the inbreeding depression (consanguinity) tends to be restored by crossing (Falconer and Mackay 1996), in Mendelian terms, heterosis effect is largely associated with dominance effects (Gregory et al., 1991) being favorable the characteristics of economic interest, mainly, those with low heritability. On the other hand, epistatic losses effects are negative and resulting from combined action of genes, in different loci, that usually are not interacting with each other (Kinghorn, 1987).

Our results to heterosis and epistatic losses effects (Table 3) according with the theory and with many researches relational to traits of weaning, as: (Arthur et al., 1999; Fries et al., 2000; Cardoso et al., 2004; Roso et al. 2005a and 2005b; Pimentel et al., 2006; Bertoli et al., 2018) and evidence the inversely proportional behavior between them. Nevertheless, the epistatic losses outweighed the gains obtained by heterosis effects.

Lastly, direct complementarity was estimated positively and represents multiplicative effect on additive actions of genes from crossing. Although complementarity effect is not well-known, in our approach was further important than breed additive effects, for instance, effects contemplated almost without exception in all works developed until then. Bertoli et al. (2016) also pointed out the relevance of inclusion complementarity, with caution, in ridge regression models to genetic evaluate of Angus x Nellore animals. Whenever cattle breeders, students and researchers discuss crossbred the hybrid vigor and complementarity between breeds terms are presents. Indeed, by model BayesB was possible to prove that the differences around crosses can be explained only with direct non-additive effects.

Thus, the better crossbred must be chosen on the balance between heterosis and epistatic losses, and an increase of complementarity. To weaning weight F1 dams, mainly, Aberdeen Angus x Nellore dams enabled calves heavier even with disadvantage of epistatic losses.

2.4.2 Variance components and genetic parameter estimates

The narrow-sense heritability was estimated in 0.29, similar to reported in literature, as 0.32 to preweaning weight gain in crosses with taurine breeds (Roso et al., 2005b), around 0.20 to postweaning gain in Nellore x Hereford, (Cardoso et al., 2005) and around 0.30 to postweaning gain in Angus x Nellore (Oliveira et al., 2016).

The variance explained by non-additive genetic effects evidence the great differences of cross breeds in relation to pure breeds. Thus, estimate the broad-sense heritability in crossbred animals is justified. Agreement with (Cardoso et al., 2005) that proved heterogeneity genetic variance, mainly in F1, considering distributions on heavier-tailed for heterogeneous residual variances to study postweaning gain in Nellore x Hereford population.

In Brazil, although crossbred animals breeding programs used heterosis, epistatic losses and complementarity effects (direct and maternal) to estimated breeding values, apparently do not estimate your variances (GenSys, 2018) which implies in underestimated genetic gain. Therefore, measure crossbred heritability indicates important application for selection programs on account of genetic gain is proportional to heritability.

2.4.3 Breeding values and accuracy

Even without informative relationship matrix, the breeding values were predicting with great accuracy. EBVs around zero were already expected given the presupposition of animal model, which assumes zero for the mean of the random effects. However, caution should be exercised in choosing Senepol x Nellore dams when weaning weight is main selection criterion, given the estimated negative breeding values for their offspring.

2.4.4 Benefits of model proposed

This is the first study that assumed selection of variables to fixed genetic effects in probability terms. In addition to benefits statistical promoted by dimensionality reduction, the BayesB model might reduce computational demand and processing time given that enable estimate non-additive effects and predict breeding values in single step, in other words, without additional analysis as it is currently done by ridge regression models (biased estimators), eliminating the subjectivity resulting from choice of k constant. It makes the BayesB model very attractive for application in breeding programs of crossbred beef cattle.

2.5 References

Arthur, P. F., H. Hearnshaw, and P. D. Stephenson. 1999. Direct and maternal additive and heterosis effects from crossing Bos indicus and Bos taurus cattle: Cow and calf performance in two environments. Livest. Prod. Sci. 57:231–241. doi:10.1016/S0301-6226(98)00172-9.

Bertoli, C. D., J. Braccini, and V. M. Roso. 2016. Comparing methodologies to estimate fixed genetic effects and to predict genetic values for an Angus × Nellore cattle population. J. Anim. Sci. 94:500–513. doi:10.2527/jas.2015-9344.

Bertoli, C. D., J. B. Neto, C. McManus, J. A. Cobuci, G. S. Campos, M. L. Piccoli, and V. Roso. 2018. Modelling non-additive genetic effects using ridge regression for an Angus – Nellore crossbred population. Anim. Prod. Sci. 59:823–834. doi:https://doi.org/10.1071/AN17439.

Bueno, R. S., R. de A. Torres, J. B. S. Ferraz, P. S. Lopes, J. P. Eler, G. B. Mourão, M. de Almeida e Silva, and E. C. de Mattos. 2012. Métodos de estimação de efeitos genéticos não-aditivos para características de peso e perímetro escrotal em bovinos de corte mestiços. Rev. Bras. Zootec. 41:1140–1145. doi:10.1590/S1516-35982012000500009.

Cardoso, F. F., R. A. Cardellino, and L. T. Campos. 2004. Componentes de (Co)variância e parametros genéticos de caracteres pós-desmama em bovinos da raça angus. Rev. Bras. Zootec. 33:313–319.

Cardoso, F. F., G. J. M. Rosa, and R. J. Tempelman. 2005. Multiple-breed genetic inference using heavy-tailed structural models for heterogeneous residual variances. J. Anim. Sci. 83:1766–1779. doi:doi.org/10.2527/2005.8381766x.

Chib, S., and E. Greenberg. 1995. Understanding the Metropolis-Hastings Algorithm. Am. Stat. 49:327–335. doi:10.2307/2684568.

Cunningham, E. P., and J. Connolly. 1989. Efficient design of crossbreeding experiments. Theor. Appl. Genet. 78:381–386. doi:10.1007/BF00265300.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 182:375–385. doi:10.1534/genetics.109.101501.

Dillard, E. U., O. Rodriguez, and O. W. Robison. 1980. Estimation of additive and nonadditive direct and maternal genetic effects from crossbreeding beef cattle. J. Anim. Sci. 50:653–663. doi:10.2527/jas1980.504653x.

Falconer DS and Mackay TFC 1996. Introduction to quantitative genetics, 3th edition. Harlow, Essex, UK: Longmans Green, 3.

Federation, B. I. 2010. Guidelines For Uniform Beef Improvement Programs.

Fries, L. A., D. J. Johnston, H. Hearnshaw, and H. U. Graser. 2000. Evidence of epistatic effects on weaning weight in crossbred beef cattle. Asian-Australasian J. Anim. Sci. 13:242.

Fries, L. A., F. S. Schenkel, V. M. Roso, F. V. Brito, J. L. P. Severo, and M. L. Piccoli. 2002. “Epistazygosity” and Epistatic effects. 7th World Congr. Genet. Appl. to Livest. Prod. 2002, Montpellier, Fr. doi:10.4161/cc.21720.

GenSys 2018. Sumário de Touros Conexão Delta G e Natura. Retrieved on 1 May 201 URL https://gensys.com.br/sumarios/

Geweke, J. 1992. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. 1–31.

Gregory, K. E., and L. V. Cundiff. 1980. Crossbreeding in beef cattle: Evaluation of systems. J. Anim. Sci. 51:1224–1242. doi:https://doi.org/10.2527/jas1980.5151224x.

Gregory, K. E., L. V Cundiff, and R. M. Koch. 1991. Breed effects and heterosis in advanced generations of composite populations for preweaning traits of beef cattle. J. Anim. Sci. 69:947. doi:10.2527/1991.693947x.

Kinghorn, B. P. 1987. The nature of 2-locus epistatic interactions in animals: evidence from Sewall Wright's guinea pig data. Theor. Appl. Genet. 73:595–604. doi:10.1007/BF00289200.

Lopes, J. S., P. R. N. Rorato, T. Weber, R. O. Araújo, M. A. Dornelles, and J. G. Comin. 2010. Avaliação do desempenho na pós-desmama para uma população bovina multirracial Aberdeen Angus x Nelore utilizando-se diferentes modelos genéticos. Arq. Bras. Med. Vet. e Zootec. 62:1439–1447.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics. 157:1819–1829. doi:11290733.

Miar, Y., G. S. Plastow, H. L. Bruce, S. S. Moore, O. N. Durunna, J. D. Nkrumah, and Z. Wang. 2014. Estimation of genetic and phenotypic parameters for ultrasound and carcass merit traits in crossbred beef cattle. Can. J. Anim. Sci. 94:273–280. doi:10.4141/cjas2013-115. Available from: http://pubs.aic.ca/doi/abs/10.4141/cjas2013-115

Mrode, R. A. 2014. Linear Models for the Prediction of Animal Breeding Values. London, UK.

Oliveira, M. M., M. L. Santana, and F. F. Cardoso. 2016. Multiple-breed reaction norm animal model accounting for robustness and heteroskedastic in a Nelore-Angus crossed population. Animal. 10:1093–1100. doi:10.1017/S1751731115002815.

Pérez, P., and G. de los Campos. 2013. BGLR : A Statistical Package for Whole Genome Regression and Prediction. Genetics. 198:483–495. doi:10.1534/genetics.114.164442.

Pimentel, E. D. C. G., S. A. De Queiroz, R. Carvalheiro, and L. A. Fries. 2006. Estimativas de efeitos genéticos em bezerros cruzados por diferentes modelos e métodos de estimação. Rev. Bras. Zootec. 35:1020–1027. doi:10.1590/S1516-35982006000400012.

Rodriguez-Almeida, F. A., L. D. Van Vleck, and K. E. Gregory. 1997. Estimation of direct and maternal breed effects for prediction of expected

progeny differences for birth and weaning weights in three multibreed populations. J. Anim. Breed. Genet. 75:1203–1212. doi:10.2527/1997.7551203x.

Roso, V. M., and L. A. Fries. 2000. Avaliação das Heteroses Materna e Individual sobre o Ganho de Peso do Nascimento ao Desmame em Bovinos Angus x Nelore. Rev. Bras. Zootec. 29:732–737.

Roso, V. M., F. S. Schenkel, S. P. Miller, and L. R. Schaeffer. 2005a. Estimation of genetic effects in the presence of multicollinearity in multibreed beef cattle evaluation. J. Anim. Sci. 83:1788–1800.

Roso, V M, F. S. Schenkel, S. P. Miller, and J. W. Wilton. 2005b. Additive, dominance, and epistatic loss effects on preweaning weight gain of crossbred beef cattle from different Bos taurus breeds. J. Anim. Sci. 83:1780–1787. doi:10.2527/2005.8381780x.

Splan, R. K., L. V. Cundiff, M. E. Dikeman, and L. D. Van Vleck. 2002. Estimates of parameters between direct and maternal genetic effects for weaning weight and direct genetic effects for carcass traits in crossbred cattle. J. Anim. Sci. 80:3107–3111. doi:10.4148/2378-5977.1635.

Toral, F. L. B., V. M. Roso, C. V. de Araújo, and J. C. R. Filho. 2011. Genetic parameters and response to selection for post-weaning weight gain, visual scores and carcass traits in Hereford and Hereford×Nellore cattle. Livest. Sci. 137:231–237. doi:10.1016/j.livsci.2010.11.013.

Tables

Table 1. Number of calves by genetic compositions of sires and dams, established by the fraction Nellore

	Dams				
Sires	Four-eighths Angus Four-eighths Nellore	Four-eighths Red-Angus Four-eighths Nellore	Four-eighths Senepol Four-eighths Nellore	Nellore	Total
Angus	112	11	63	2,491	2,677
Red-Angus	0	2	2	331	335
Charolais	101	75	43	5	224
Senepol	1,100	610	491	1,586	3,787
Simmental	166	104	40	0	310
Three-eighths Nellore	1,822	1,086	613	147	3,668
Nellore	339	111	50	4,625	5,125
Total	3,640	1,999	1,302	9,185	16,126

367

Table 2. Pearson correlation coefficients (above diagonal) and Spearman correlation (below diagonal) among predictor variables of direct and maternal systematic effects[1]

	ad	am	hd	hm	ed	cd	Cm
ad		0.68	-0.94	-0.68	-0.57	-0.85	-0.68
am	0.85		-0.67	-1.00	-0.92	-0.43	-1.00
hd	-0.82	-0.77		0.67	0.66	0.94	0.67
hm	-0.85	-1.00	0.77		0.92	0.43	1.00
ed	-0.76	-0.96	0.83	0.95		0.47	0.92
cd	-0.35	-0.00	0.54	0.00	0.08		0.43
cm	-0.85	-1.00	0.77	1.00	0.96	0.00	

[1]ad= direct breed additive effect; am= maternal breed additive effect; hd= direct heterozygosity effect; hm= maternal heterozygosity effect; ed= direct epistazygosity effect; cd= direct complementarity effect; cm= maternal complementarity effect.

367

Table 3. Posterior means with standard deviation of breeding value (EBV) and weaning weight (WW), breed additive and non-additive genetics effects (genetic covariates used for estimate them), sum of non-additive genetics effect ($\sum_{i=1}^{na} na$) and ranking (R) for each crossing[1]

Crossing[2]	EBV	WW	ad	am	hd	hm	ed	cd	cm	$\sum_{i=1}^{na} na$	R
A(AN)	0.21 (6.52)	228.0 (24.96)	0 (0.25)	0 (0.5)	10.84 (0.38)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	7.69	13
C(AN)	0.30 (5.74)	228.28 (22.18)	0 (0.25)	0 (0.5)	18.06 (0.63)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	14.91	12
S(AN)	-0.01 (5.69)	221.68 (25.98)	0 (0.25)	0 (0.5)	18.06 (0.63)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	14.91	20
ST(AN)	0.96 (6.33)	236.41 (24.34)	0 (0.25)	0 (0.5)	18.06 (0.63)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	14.91	4
3/8N(AN)	0.13 (6.37)	227.70 (23.04)	0 (0.44)	0 (0.5)	18.77 (0.65)	0 (0.5)	-18.43 (0.47)	8.76 (0.25)	0 (0.25)	9.13	14
N(AN)	-0.20 (6.38)	239.51 (21.46)	0 (0.75)	0 (0.5)	10.84 (0.38)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	7.69	3
A(RN)	-0.23 (8.15)	232.64 (29.03)	0 (0.25)	0 (0.5)	18.06 (0.63)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	14.91	8
R(RN)	-0.03 (3.60)	242.16 (21.92)	0 (0.25)	0 (0.5)	10.84 (0.38)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	7.69	1
C(RN)	0.52 (6.78)	236.03 (18.69)	0 (0.25)	0 (0.5)	18.06 (0.63)	0 (0.5)	-9.83 (0.25)	6.675 (0.19)	0 (0.25)	14.90	5
S(RN)	0.73 (6.55)	229.95 (26.36)	0 (0.25)	0 (0.5)	18.06 (0.63)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	14.91	10
ST(RN)	0.96 (7.49)	235.36 (22.13)	0 (0.25)	0 (0.5)	18.06 (0.63)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	14.91	7
3/8N(RN)	0.12 (6.31)	227.46 (21.77)	0 (0.44)	0 (0.5)	18.77 (0.65)	0 (0.5)	-18.43 (0.47)	8.76 (0.25)	0 (0.25)	9.13	15
N(RN)	0.43 (5.95)	235.91 (21.51)	0 (0.75)	0 (0.5)	10.84 (0.38)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	7.69	6
A(SN)	-2.25 (6.41)	228.97 (19.79)	0 (0.25)	0 (0.5)	18.06 (0.63)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	14.91	11
R(SN)	-7.85 (12.34)	195.03 (36.86)	0 (0.25)	0 (0.5)	18.06 (0.63)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	14.91	25
C(SN)	-0.86 (5.99)	222.57 (24.04)	0 (0.25)	0 (0.5)	18.06 (0.63)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	14.91	19
S(SN)	-0.56 (6.50)	215.46 (24.84)	0 (0.75)	0 (0.5)	10.84 (0.38)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	7.69	21
ST(SN)	-0.37 (7.20)	239.81 (23.87)	0 (0.25)	0 (0.5)	18.06 (0.63)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	14.91	2
3/8N(SN)	-1.49 (6.33)	224.28 (22.94)	0 (0.44)	0 (0.5)	18.78 (0.65)	0 (0.5)	-18.43 (0.47)	8.76 (0.25)	0 (0.25)	9.14	17
N(SN)	-0.74 (6.54)	231.37 (21.52)	0 (0.75)	0 (0.5)	10.84 (0.38)	0 (0.5)	-9.83 (0.25)	6.68 (0.19)	0 (0.25)	7.69	9
A(N)	0.56 (6.18)	225.10 (22.65)	0 (0.5)	0 (1)	14.45 (0.5)	0 (0)	0 (0)	8.9 (0.25)	0 (0)	23.35	16
R(N)	0.03 (5.74)	209.84 (27.30)	0 (0.5)	0 (1)	14.45 (0.5)	0 (0)	0 (0)	8.9 (0.25)	0 (0)	23.35	23
C(N)	1.87 (12.45)	223.09 (22.03)	0 (0.5)	0 (1)	14.45 (0.5)	0 (0)	0 (0)	8.9 (0.25)	0 (0)	23.35	18
S(N)	-0.72 (5.69)	197.85 (20.20)	0 (0.5)	0 (1)	14.45 (0.5)	0 (0)	0 (0)	8.9 (0.25)	0 (0)	23.35	24

Crossing[2]	EBV	WW	ad	am	hd	hm	ed	cd	cm	$\sum_{i=1}^{na} na$	R
3/8N(N)	1.75 (5.37)	210.28 (22.94)	0 (0.69)	0 (1)	12.42 (0.43)	0 (0)	-9.20 (0.23)	7.65 (0.21)	0 (0)	10.87	22
N(N)	-0.01 (5.60)	194.80 (19.28)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.00	26
Mean	0.00 (6.06)	215.25 (27.54)	0	0	11.68 (7.62)	0	-6.19 (7.55)	5.88 (3.77)	0	11.36 (8.96)	-

[1]A= Angus; R= Red Angus; C= Charolais; S= Senepol; ST= Simmetal; 3/8N= Three-eighths Nellore and N= Nellore. ad= direct breed additive effect; am= maternal breed additive effect; hd= direct heterosis effect; hm= maternal heterosis effect; ed= direct epistatic loss effect; cd= direct complementarity effects and cm= maternal complementarity effect.

[2]Crossing denote breed sire x breed dam; mean in kg; Ranking ordered decreasing by posterior mean weaning weight.

367

Table 4. Posterior means of variance components (kg²) and genetic parameter for weaning weight trait in tropical crossbred[1]

Parameters	
σ_a^2	159.95
σ_d^2	58.03
σ_i^2	57.08
σ_c^2	14.25
σ_e^2	373.74
σ_P^2	663.05
h^2	0.2997
H^2	0.4148

[1]σ_a^2= additive genetic variance; σ_d^2= dominance variance; σ_i^2= epistatic losses variance; σ_c^2= genetic complementarity variance; σ_e^2= residual variance; σ_P^2= phenotypic variance; h^2 = narrow-sense heritability and H^2 = broad-sense heritability.

367

Table 5. Decomposition of the variance structure of the non-additive genetic effects estimates by BayesB model[1]

	hd	ed	cd
hd	58.03	-0.66	0.93
ed	-38.06	57.08	-0.47
cd	26.92	-13.40	14.25

[1]hd= direct heterosis; ed= direct epistatic losses and cd= direct complementarity. Variance (kg²) on the diagonal; Pearson correlation coefficient above diagonal and (co)variance (kg) below diagonal.

367

Figures

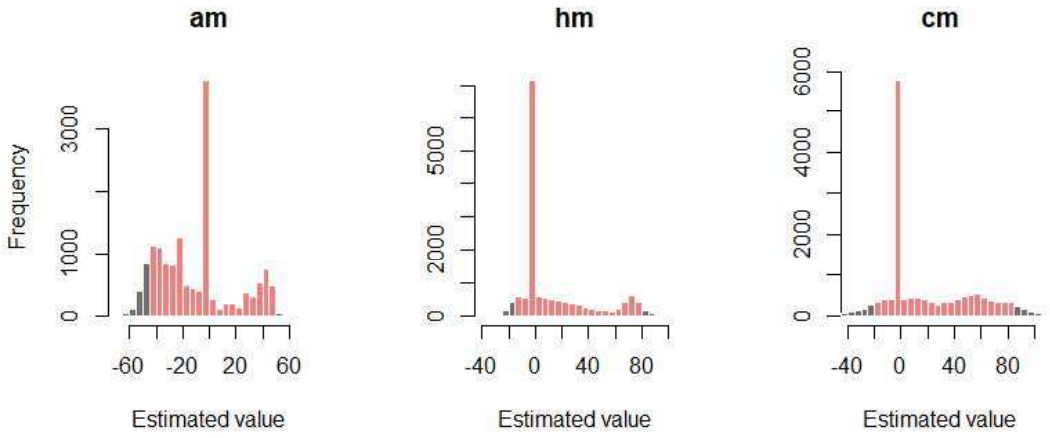


Figure 1. Posterior of maternal genetic covariate: breed additive (am), heterozygosity (hm) and non-additive complementarity (cm) for weaning weight (adjusted to 240 days) with BayesB model (colored region represents the highest posterior density interval – PD95%).

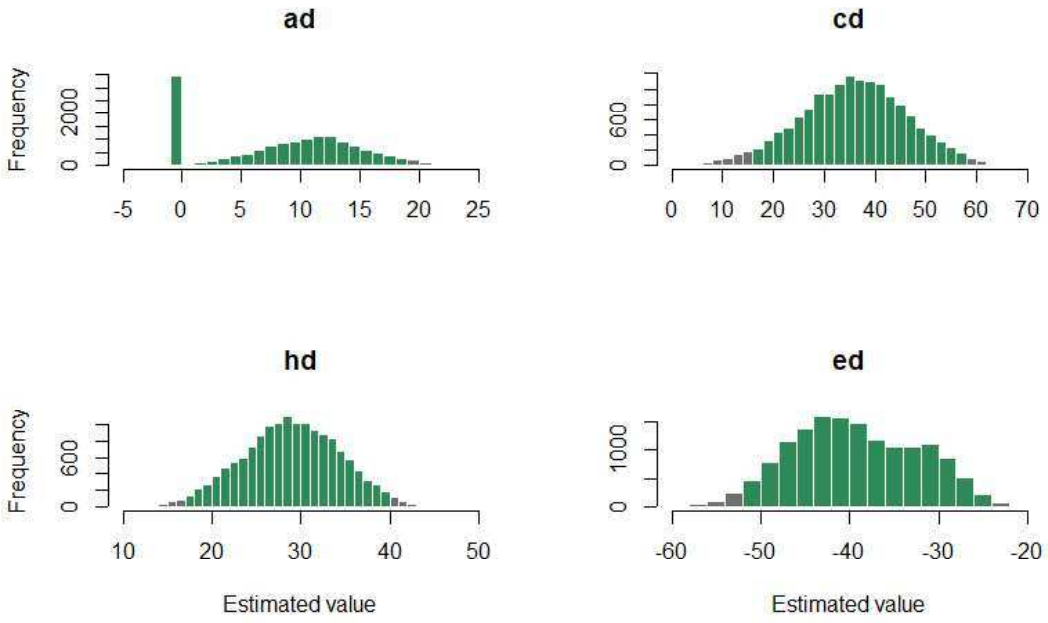


Figure 2. Posterior of direct genetic covariate: breed additive (ad), heterozygosity (hd), non-additive complementarity (cd) and epistazygosity (ed) for weaning weight (adjusted to 240 days) with BayesB model (colored region represents the highest posterior density interval – PD95%).

Chapter 3

Genome-enabled classification for stayability in Nellore cattle under a machine learning framework

Abstract: Currently, stayability stands as one of the most relevant female reproductive traits in beef cattle. This is a binary trait whose the success definition is given by a cow providing at least 3 calvings at 76 months of age. In response to the substantial development of high-density SNP markers, there is an increasing interest on machine learning (ML) methods for genome-enabled classifications. In this study, we use separately three sires' genotypic data set (with one, three and five thousand markers) in order to classify the daughters' stayability trait by different ML approaches (AdaBoost – ADA, Bernoulli Naïve Bayes – NB, Decision Tree - DT, Nearest Neighbors – KN, Multilayer Perceptron – MLP and Support Vector Machine for Classification – SVC). The evaluation of classification models' performance showed that it is possible to classify stayability using 1,000 markers, since it is more efficient in terms of computational demands, that is, processing time and data storage capacity in the Random Access Memory - RAM. The KN and NB methods provided greater classification ability, whereas the ADA, DT, MLP and SVC were affected by overfitting issues. The NB algorithm outperformed KN, and reported average accuracy, precision, and specificity of 0.62, 0.45, and 0.76, respectively. In summary, simplest tools (NB and KN) are robust and efficient machines recommend for genome-enabled classifications in animal breeding.

Keywords: classifier tools, genomic selection, naïve Bayes, neural network.

3.1 Introduction

With the coming of single nucleotide polymorphism (SNP) marker panels e with a meaning decrease at the costs of genotyping through these panels, the genome-wide selection became popular and accessible at the animal breeding field and plant. Thus, the ratio n:p (individuals:markers) increase rapidly over the years. The study realized by Meuwissen et al. (2001) revolutionized the manner to predict breeding values and, since then, new statistical methods have been proposed in order to minimize some issues such as high-dimensionality and multicollinearity, which coming from simultaneous estimation of all SNPs. In this context, several additive linear models have been proposed to this aim, such as BayesA, BayesB, BayesCπ and BLASSO, among others (Gianola et al., 2006; de los Campos et al., 2009). In this context, there has been increasing interest in predictive tools based on non-linear relations, that does not require the additive assumptions. Among them, the artificial neural network (ANN) have been highlighted because are admittedly universal approximators of complex functions (Hornik et al., 1989).

The several studies involving neural networks applied to genome-enabled prediction were conducted including marbling score trait (Okut et al., 2013), yield of milk, fat and protein in Jersey (Gianola et al., 2011) and Holstein cows (Ehret et al., 2015), heritability estimates using simulated data (Glória et al., 2016) and also classification of chick mortality (Long et al., 2009). Undoubtedly, scenarios with larger genomic data set analyzed by machine learning (ML) algorithms imply in an expensive computational demands, in terms of processing time and data storage capacity in the Random Access Memory – RAM, for this reason, the dimensionality reduction or, in ML terms, feature extraction have been strongly encouraged before data modeling (de los Campos et al., 2010) For these reasons, simplified ML methods such as Naïve Bayes (Long et al., 2009), Support Vector Machine (Long et al., 2011), Neuro-Fuzzy (Shahinfar et al., 2012), Random Forest (Mokry et al., 2013) and XgBoost (Li et al., 2018) have been studied.

In this sense, we previously performed SNPs selection to fitting three sires' genomic data set (one, three and five thousand markers), in order to

classify the daughters' stayability trait. Our aims with this paper were: 1) to choose the better genomic data structure to classify stayability using different ML methods (AdaBoost – ADA, Bernoulli Naïve Bayes – NB), Decision Tree - DT, Nearest Neighbors – KN, Multilayer Perceptron – MLP and Support Vector Machine for Classification – SVC), and 2) to recommend the algorithm with higher classification ability for animal breeding purposes.

3.2 Materials and Methods

Animal Care and Use Committee approval was not obtained for this study because the data set utilized was obtained from an existing database.

3.2.1 Phenotypic and genotypic data

The entire data set used in the present study came from the breeding program of beef cattle of Geneplus Consultoria Agropecuária Ltda – Embrapa, Brazil. The stayability trait refers to ability of cows to give birth to, at least, three viable calves until 76 consecutive months (Silva et al., 2018). Records from 44,626 daughters of genotyped bulls born between 1972 and 2014 with age at first calving between 20 and 60 months were used in the analyses. Value zero is assigned to failure and value one to success. Genomic data include information of 10,909 bulls. SNPs with minor allele frequency (MAF) minor that 0.05 and missing genotype frequency (call rate) greater that 0.95 were removed. After quality control were considered 309,788 SNPs located on autosome chromosomes. The SNP genotypes were coded as 0, 1 and 2 for the homozygous recessive, heterozygous and homozygous dominant, respectively.

3.2.2 Genetic evaluation

Considering a general approach for genomic selection we performed genetic evaluation by way of ssGBLUP method (single step Genomic Best Linear Unbiased Predictor) proposed by Legarra et al. (2014), using Bayesian linear-threshold model in order to estimate the effect of markers and to predict

the residual value associated to each phenotype. This model can be written as following:

$$\boldsymbol{l} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e},$$

where: $\boldsymbol{l}$ is the unobservable underlying liabilities for the categorical trait stayability; $\boldsymbol{\beta}$ the vector of systematic effects (contemporary group and class of age at first calving); $\boldsymbol{a}$ is the vector additive genetic effects (breeding values); $\boldsymbol{e}$ is the vector of residuals; e $\boldsymbol{X}$ e $\boldsymbol{Z}$ are incidence matrices related to $\boldsymbol{\beta}$ and $\boldsymbol{a}$, respectively. It is assuming, that prior $\boldsymbol{\beta}$ follows uniform distribution $\boldsymbol{P}(\boldsymbol{\beta})\ \alpha\ constant;\ \boldsymbol{a}|\sigma_a^2 \sim \boldsymbol{N}(0, \boldsymbol{H}\sigma_a^2)$, where $\boldsymbol{H}$ is the hybrid relationship matrix including both pedigree and genomic information and σ_a^2 is the variance additive genetic; and $\boldsymbol{e}|\sigma_e^2 \sim \boldsymbol{N}(0, \boldsymbol{I}\sigma_e^2)$ where $\boldsymbol{I}$ e σ_e^2 are the identity matrix and variance residual, respectively. The scaled inverse chi-squared distribution was defined prior for σ_a^2 e σ_e^2. The pedigree file used in the analysis contained 152,352 animals.

Contemporary groups (CG) were defined as groups of animals born in the same herd, year and season of birth (March to May, June to August, September to November and December to February).

Usually 13 classes are assigned to age at first calving (20-22, 23-25, 26-28, 29-31, 32-34, 35-37, 38-40, 41-43, 44-46, 47-49, 50-52, 53-55, 56-58, 59-60 months), seeking to reduce the number of classes was performed cluster analysis by self-organizing map of Kohonen, an artificial neural network based on competitive, cooperative and unsupervised learning (Kohonen, 1990). In a bidimensional architecture of neurons with topology 5x5, the number of calves, age at the first calving and stayability were presented at the input and within cluster sum of squares (Appendix 1) for all weight vectors were computed and used to obtain the best matching unit (BMU) according to Kohonen (2013). In each iteration, the input data was divided into subsets that share the same BMU, and a Gaussian neighborhood function was applied as a smoothing kernel to update the map. Through this iterative method, the corresponding weight vectors to the BMU were updated and adjusted according to their neighboring neurons. After the convergence (1000 epochs), the following five classes were obtained (Figure 1). This cluster analysis was

implemented in the kohonen package (Wehrens and Buydens, 2007) of the R software (R Core Team, 2019).

Systematic effects, variance components and breeding values (EBV) were inferred using the THRGIBBS1F90 software. Effect to all SNPs and residual value associated to each phenotype were obtained performing the POSTGSF90 software and PREDICTF90 software (Misztal et al., 2018), respectively. Analysis consisted of a single chain of 300,000 cycles, with a conservative burn-in period of 100,000 cycles and a thinning interval of 4 cycles. The convergence was verifying by Geweke test (Geweke, 1992) available in the boa package (Smith, 2007) of the R software.

3.2.3 Markers selection, inclusion of biological noise and label description

High-dimensional promoted by whole marker set is often associated to overfitting, due to the amount of redundant information which implies in less predictive ability and increase in computational resources (Ehret et al., 2015). For this reason, the markers selection was implemented before data modeling. The absolute effect of markers estimated by single-step genomic BLUP method were used to select 1,000, 3,000 and 5,000 key SNPs involved in stayability control (see Appendix 2 for more details), most them identified on chromosomes 1, 2, 3, 6, 9 and 12 (Figure 2).

According to (Teixeira et al., 2017) 147 candidate genes were detected on 1, 2, 5, 6, 9 and 20 chromosomes affecting the stayability phenotypes, which gives us support to exploit the markers selected under a genome-enabled classification viewpoint.

In order to develop a robust classifier was inserted biological noise, assessed by residual value obtained from genetic evaluation, on samples. The input was defined as sum of EBV and residual value evaluated in females and transformed in probability, considering the phenotypic probability to return binary class. Thus, to probabilities equal or above 0.2977 was assigned as success (1), whereas, failure (0).

3.2.4 Benchmark data sets and validation model

We sampled randomly 22 313 phenotypes obtained from the previous item to compose the benchmark data sets. Starting from selected markers three data sets were presented separately to the machine learning (ML) algorithms, in order to compare the impact of different data structures. Finally, data sets were defined according to number of inputs, set 1, 2 and 3 correspond to thousand, three thousand and five thousand markers, respectively. To evaluate the ability of the ML models in to distinguish stayability classes was performed 10-fold cross validation. The training and testing set were randomly selected to make ten subsets of genotypes associated at the phenotypes. During cross-validation process, each of the ten subsets generated served as testing in one round, with missing phenotypes. Thus, the averages observed on test sets (2,231 samples per test set) was used to quantifier the quality of classification.

3.2.5 Machine learning algorithms

The challenge of the learning algorithm is to induce a classifier to label new cases (testing set) with good accuracy (Mitchell, 1997). In this study, the following supervised ML algorithms were analyzed to generate classifiers models: AdaBoost (ADA), Bernoulli Naïve Bayes (NB), Decision Tree (DT), Nearest Neighbors (KN), Multilayer Perceptron (MLP) and Support Vector Machine for Classification (SVC).

3.2.5.1 AdaBoost

In boosting methods, base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble. ADA was introduced by Schapire and Freund (1997), its main concept is increase synaptic weight of examples difficult to classifier. Each subsequent weak learner is thereby forced to concentrate on the examples that are missed by the previous ones in the sequence (Hastie et al., 2009). The AdaBoost SAMME

algorithm — Stagewise Additive Modeling using a multi-class Exponential loss function was run in our analysis and can be seen detailly in Hastie et al. (2009).

3.2.5.2 Bernoulli Naïve Bayes

Developed to training and classification of data distributed according to multivariate Bernoulli distributions. The decision rule for NB is based on Bayes' theorem:

$$P(C = c|X_1 = x_1, \dots, X_p = x_p) = \frac{P\big(X_1 = x_1, \dots, X_p = x_p\big|C = c\big)P(C = c)}{P(x_1, \dots, X_p = x_p)}$$

Where $\boldsymbol{x_p}$ correspond to SNPs (0, 1 or 2) and $\boldsymbol{c}$ to stayability class (0 or 1). The NB explicitly penalizes the non-occurrence of a variable that is an indicator for class, assuming independence between the variables (Mccallum and Nigam, 1997; Raghavan and Schütze, 2010).

3.2.5.3 Decision Tree

Is a non-parametric supervised learning method, its structure of decision tree is composed of a root node, internal nodes, and leaf nodes. The internal nodes correspond to the values of the inputs, and each leaf node of the tree contains the probability distribution and the label of the class (Silva et al., 2019). Among the DT algorithms we used the CART – Classification and Regression Trees, that constructs binary trees using the input and threshold that yield the largest information gain at each node.

3.2.5.4 Nearest Neighbors

Neighbors-based classification is a type of non-generalizing learning, it does not attempt to construct a general internal model, but simply stores samples of the training data. Thus, the classification is assessed from a simple majority vote of the nearest neighbors of each point, defined as data class which has the most representatives within the nearest neighbors (Hastie et al., 2009).

3.2.5.5 Multilayer Perceptron

This algorithm is a neural networks (NN) class widely characterized as universal approximator of complex functions (Hornik et al., 1989), used to solve prediction problems and classification. The NNs are based on the use of activation functions that can intuitively approximate linear and nonlinear relationships between a response variable and its predictor variables. This is possible through the use of training algorithms, which update specific weights (subdivided into different levels called “neurons”) of each predictor variable in different levels of learning (called “layers”). Thus, when updates of weights do not culminate in significant improvements in prediction ability, can be assumed that the network is trained. Subsequently, it is necessary only to use new values of the independent variables so that the network provides the predictions inherent to them. In order to test different topologies, we evaluated MLP neural networks with one and two hidden layers as following: net1 (one layer with one neuron), net2 (one layer with two neurons), net3 (one layer with three neurons), net4 (one layer with four neurons), net5 (one layer with five neurons), net6 (one layer with ten neurons), net7 (one layer with fifteen neurons), net8 (two layers – with two and one neuron per layer), net9 (two layers – with two neurons in both layers), net10 (two layers – with three and two neurons per layer), net11 (two layers – with four and two neurons per layer). All nets were run using the Stochastic Gradient Descent (SGD) learning method, a simple yet very efficient approach to discriminative learning of linear classifiers. The activation functions linear and logistic were separately tested to all nets, a learning rate equal to 0.4 and 1,000 epochs were fixed.

3.2.5.6 Support Vector Machine for Classification

This method has been applicated in machine learning for classification and regression, it is also a particular case of RKHS (González-Recio et al., 2014). The SVC method uses a pre-established kernel function to separate different classes of data, maximizing the distance between the closest points in relation to each class to find a separation line between data (Silva et al., 2019).

The ML methods employed were fitted using the Scikit library available in/ Python programming language (Pedregosa et al., 2011). All methods were run with the Scikit default parameters, except the MLP algorithm, which was previously described.

3.2.6 Performance evaluation metrics

Finally, to choose the better classifier, in terms of predictive ability, we calculated metrics widely known in ML approaches based on the confusion matrix. Given the predicted values as True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) for each test set, the following metrics were obtained:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Balanced\ accuracy = \left(\frac{TP}{TP + FP} + \frac{TN}{TN + FN}\right) \div 2$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$Sensitivity\ (recall) = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$\text{F1 score} = \frac{2TP}{2TP + FP + FN}$$

$$\text{Matthews correlation coefficient (MCC)} =$$

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$\text{False discovery rate (FDR)} = \frac{FP}{FP + TP}$$

3.3 Results and Discussion

Figure 3 displays the evaluation metrics obtained from testing data set with 1,000 SNPs for all learning algorithms. The metrics were accessed by average confusion matrix (Figura 4). In sequence, are presented the confusion matrix in terms of standard deviation (Figure 5) and percentage (Figure 6).

3.3.1 Comparison of SNP subsets

We goal in this study was, starting of sire's SNPs, predicted the daughter's phenotype using ML approaches. For this, different structures of genomic set (1k, 3k and 5k markers) were applied to train and test the algorithms employed, in order to evaluated the influence of variable predictors on the learning process. The Figure 3 and Table 1 show that the performance of classifiers was equivalent to all genomic set. This result might be attributed to polygenic inheritance of the character revealed by Manhattan plot (Figure 2), which implies in noisy input, since the SNPs, in general, are redundant. Indeed, mainly in genome-enabled predictions the features selection is often recommended for ML methods (Felipe et al., 2014; González-Recio et al., 2014; Ehret et al., 2015), which justifies our work to choose most important markers by real genetic evaluation. Although the number of markers is directly related to overfitting and consequently, to poor predictive ability (Glória et al., 2016), the number of SNPs to be used will depend on the characteristic studied. Therefore, we recommended 1,000 SNP's (inputs) to classifier stayability trait, so as to the subsequent results will refer to them.

3.3.2 Evaluation of neural networks models

Assessing the accuracy across MPL models (Table 2) the overfitting, even in the simplest topologies, was clearly observed applying to 0 label, since the estimated accuracy was equivalent to frequency of failure class. When NNs without regularization are implemented to prediction of continuous values the overfitting issue is generally verified with the increase in number of neurons and middle-layers (Gianola et al., 2011). However, this behavior is not valid to our results and reveals that classification error might have been evidenced by

imbalanced classes (Figure 7) as well as the strong association between the predictive variables, since predictive ability of NN did not depend of network architecture when the number of sample at the training was larger than the number of markers (Okut et al., 2011). For comparison purposes among ML models we choose, under simplistic viewpoint, the linear neural network (net1) was used for further analysis.

3.3.3 Prediction performance assessment

The Naïve Bayes (NB) approach outperformed, clarify, the other classification methods employed (Figure 3), followed by KN, SVC, DT, ADA and lastly MLP. In general, can be seen in Figure 4 that the learning algorithms predicted well the failure label, this might be attributed to inheritance mechanism of phenotype studied (Van Melis et al., 2007), which is considerably affected by farmer's choice in, for instance, challenge heifer for early reproduction. Investigating the evaluation metrics, mainly FDR (compute the type I errors) and MCC (balanced measure of confusion matrix), all models seem is performing similarly to random guessing, excepting NB. This, sure, is associated to imbalanced data (Figure 6) and can be clearly observed from balanced accuracy, that show values around 50%, and by confusion matrix at the standard deviation approach (Figure 5), which revel less SD to the success class in the methods suppressed by overfitting (SVC, DT, ADA and MLP). The NK and NB models performed similarly well, it's because both have training algorithms simplest. Comparatively, NB provide classification with twice as precision whereas KN had almost double odds of occurrence of type I errors (FDR). The NB have been used to solve complex issue of classification, as text categorization (Mccallum and Nigam, 1997; Eyheramendy et al., 2003), showing that simple methods also can be robust. The advantages of NB under classification models involving SNPs was also described by Long et al. (2009) and, although the conditional independence of markers given class assumed in NB might be violated (Domingos and Pazzani, 1996), it seems to be the key point to obtain accurate classifications against complex attributes.

3.4 References

de los Campos, G., D. Gianola, G. J. M. Rosa, K. A. Weigel, and J. Crossa. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. (Camb). 92:295–308. doi:10.1017/S0016672310000285.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 182:375–385. doi:10.1534/genetics.109.101501.

Domingos, P., and M. Pazzani. 1996. Beyond independence: Conditions for the optimality of the simple Bayesian Classier. Proc. 13th Int. Conf. Mach. Learn. 105–112.

Ehret, A., D. Hochstuhl, D. Gianola, and G. Thaller. 2015. Application of neural networks with back-propagation to genome-enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle. Genet. Sel. Evol. 47:1–9. doi:10.1186/s12711-015-0097-5.

Eyheramendy, S., D. D. Lewis, and D. Madigan. 2003. On the naive bayes model for text categorization. Proc. Artif. Intell. Stat. 32546:3–10.

Felipe, V. P. S., M. A. Silva, B. D. Valente, and G. J. M. Rosa. 2014. Using multiple regression, Bayesian networks and artificial neural networks for prediction of total egg production in European quails based on earlier expressed phenotypes. Poult. Sci. 94:772–780. doi:10.3382/ps/pev031.

Geweke, J. 1992. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. 1–31.

Gianola, D., H. Okut, K. A. Weigel, and G. J. M. Rosa. 2011. Predicting complex quantitative traits with Bayesian neural networks: A case study with Jersey cows and wheat. BMC Genet. 12:4–7. doi:10.1186/1471-2156-12-87.

Gianola, D., R. L. Fernando, and A. Stella. 2006. Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. Genetics. 173:1761–1776. doi:10.1534/genetics.105.049510.

Glória, L. S., C. D. Cruz, R. A. M. Vieira, M. D. V. de Resende, P. S. Lopes, O. H. G. B. D. de Siqueira, and F. Fonseca e Silva. 2016. Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. Livest. Sci. 191:91–96. doi:10.1016/j.livsci.2016.07.015.

González-Recio, O., G. J. M. Rosa, and D. Gianola. 2014. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livest. Sci. 166:217–231. doi:10.1016/j.livsci.2014.05.036.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning. Springer series in statistics., New York.

Hornik, K., M. Stinchcombr, and H. White. 1989. Multilayer feedforward networks are universal approximators. Neural Networks. 2:359–366.

Kohonen, T. 1990. The Self-organizing Map. 78:1464–1480.

Kohonen, T. 2013. Essentials of the self-organizing map. Neural Networks. 37:52–65. doi:10.1016/j.neunet.2012.09.018.

Legarra, A., O. F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step: a general approach for genomic selection $. Livest. Sci. 166:54–65. doi:10.1016/j.livsci.2014.04.029.

Li, B., N. Zhang, Y. G. Wang, A. W. George, A. Reverter, and Y. Li. 2018. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front. Genet. 9:1–20. doi:10.3389/fgene.2018.00237.

Long, N., D. Gianola, G. J. M. Rosa, and K. A. Weigel. 2011. Application of support vector regression to genome-assisted prediction of quantitative traits. Theor. Appl. Genet. 123:1065–1074. doi:10.1007/s00122-011-1648-y.

Long, N., D. Gianola, G. J. M. Rosa, K. A. Weigel, and S. Avendaño. 2009. Comparison of classification methods for detecting associations between SNPs and chick mortality. 14:1–14. doi:10.1186/1297-9686-41-18.

Mccallum, A., and K. Nigam. 1997. A Comparison of Event Models for Naive Bayes Text Classification.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819–1829. doi:11290733.

Misztal, I., S. Tsuruta, D. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. 2018. Manual for BLUPF90 family programs. University of Georgia.

Mitchell, T. M. 1997. Machine Learning. 1st ed. McGraw Hill Higher Education, New York.

Mokry, F. B., R. H. Higa, M. de Alvarenga Mudadu, A. Oliveira de Lima, S. L. C. Meirelles, M. V. G. Barbosa da Silva, F. F. Cardoso, M. Morgado de Oliveira, I. Urbinati, S. C. Méo Niciura, R. R. Tullio, M. Mello de Alencar, and L. Correia de Almeida Regitano. 2013. Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. BMC Genet. 14:1–11. doi:10.1186/1471-2156-14-47.

Okut, H., D. Gianola, G. J. M. Rosa, And K. A. Weigel. 2011. Prediction of body mass index in mice using dense molecular markers and a regularized neural network. Genet. Res. (Camb). 93:189–201. doi:10.1017/s0016672310000662.

Okut, H., X. L. Wu, G. J. Rosa, S. Bauck, B. W. Woodward, R. D. Schnabel, J. F. Taylor, and D. Gianola. 2013. Predicting expected progeny difference for marbling score in Angus cattle using artificial neural networks and Bayesian regression models. Genet. Sel. Evol. 45:1. doi:10.1186/1297-9686-45-34.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12:2825–2830.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

Raghavan, C. D. M. P., and H. Schütze. 2010. An introduction to information retrieval. Nat. Lang. Eng. 16:569. doi:10.1210/endo-38-3-156.

Schapire, R. E., and Y. Freund. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J. Comput. Syst. Sci. 55:119–139. doi:10.1088/0034-4885/55/7/004.

Shahinfar, S., H. Mehrabani-Yeganeh, C. Lucas, A. Kalhor, M. Kazemian, and K. A. Weigel. 2012. Prediction of breeding values for dairy cattle using

artificial neural networks and neuro-fuzzy systems. Comput. Math. Methods Med. 2012. doi:10.1155/2012/127130.

Silva, D. O., M. L. Santana, D. R. Ayres, G. R. O. Menezes, L. O. C. Silva, P. R. C. Nobre, and R. J. Pereira. 2018. Genetic parameters for stayability to consecutive calvings in Zebu cattle. Animal. 12:1807–1814. doi:10.1017/S1751731117003457.

Silva, J. C. F., R. M. Teixeira, F. F. Silva, S. H. Brommonschenkel, and E. P. B. Fontes. 2019. Plant Science Machine learning approaches and their current application in plant molecular biology: A systematic review. Plant Sci. 284:37–47. doi:10.1016/j.plantsci.2019.03.020.

Smith, B. J. 2007. boa: An R package for MCMC output convergence assessment and posterior inference. J. Stat. Softw. 21. doi:10.18637/jss.v021.i11.

Teixeira, D. B. A., Fernandes Júnior, G. A., Silva, D. B. S., Costa, R. B., Takada, L., Gordo, D. G. M., Bresolin, T., Carvalheiro, R., Baldi, F., Albuquerque, L. G. 2017. Genomic analysis of stay ability in Nellore cattle. Plos One. 12(6): e0179076. doi.org/10.1371/journal.pone.0179076.

Van Melis, M. H., J. P. Eler, H. N. Oliveira, G. J. M. Rosa, J. A. I. I. V. Silva, J. B. S. Ferraz, and E. Pereirat. 2007. Study of stayability in Nellore cows using a threshold model. J. Anim. Sci. 85:1780–1786. doi:10.2527/jas.2005-608.

Wehrens, R., and L. M. C. Buydens. 2007. Self- and Super-organizing Maps in R: The kohonen Package. JSS J. Stat. Softw. 21. Available from: http://www.jstatsoft.org/

Tables

Table 1. Average of performance results for the machine learning classifiers of stayability from test sets with three and five thousand markers.

Input variables (Attribute)	ML algorithm[1]	Evaluation metrics							
		Accuracy	Balanced accuracy	Precision	Sensitivity (recall)	Specificity	F1	MCC[2]	FDR[2]
Three thousand markers	ADA	0.7101	0.5067	0.0273	0.4428	0.7149	0.0508	0.0447	0.9727
	NB	0.6344	0.5769	0.4413	0.3830	0.7594	0.4099	0.1480	0.5587
	DT	0.7068	0.5067	0.0350	0.4043	0.7150	0.0643	0.0399	0.9650
	KN	0.6633	0.5340	0.2289	0.3637	0.7293	0.2800	0.0794	0.7711
	MLP	0.7121	0.5000	0.0000	0.0000	0.7121	0.0000	0.0000	1.0000
	SVC	0.7112	0.5062	0.0230	0.4732	0.7147	0.0437	0.0478	0.9770
Five thousand markers	ADA	0.7095	0.5047	0.0218	0.4325	0.7141	0.0413	0.0364	0.9782
	NB	0.6315	0.5736	0.4369	0.3788	0.7574	0.4056	0.1415	0.5631
	DT	0.7068	0.5067	0.0350	0.4041	0.7150	0.0643	0.0398	0.9650
	KN	0.6632	0.5337	0.2284	0.3632	0.7291	0.2796	0.0789	0.7716
	MLP	0.7121	0.5000	0.0000	0.0000	0.7129	0.0000	0.0000	0.9926
	SVC	0.7108	0.5061	0.0237	0.4634	0.7147	0.0450	0.0465	0.9763

[1]ADA= AdaBoost, NB= Bernoulli Naïve Bayes, DT= Decision Tree, KN= Nearest Neighbors, MLP= Multilayer Perceptron (with one layer, one neuron and identity activation function) and SVC= Support Vector Machine for Classification.

[2]MCC= Matthews correlation coefficient and FDR= False discovery rate.

Table 2. Average accuracy of multilayer perceptron classifier (MLP) fitted using the stochastic gradient descent (SGD) learning algorithm and differ topologies from test sets.

Activation function	Neural networks[1]										
	net1	net2	net3	net4	net5	net6	net7	net8	net9	net10	net11
Identity	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
Logistic	0.72	0.72	0.72	0.71	071	0.70	0.70	0.72	0.72	0.72	0.71

[1]net1 = one layer with one neuron, net2= one layer with two neurons, net3= one layer with three neurons, net4= one layer with four neurons, net5= one layer with five neurons, net6= one layer with ten neurons, net7= one layer with fifteen neurons, net8= two layers - with two and one neuron per layer, net9= two layers - with two neurons in both layers, net10= two layers - with three and two neurons per layer and net11= two layers - with four and two neurons per layer.

367

Figures

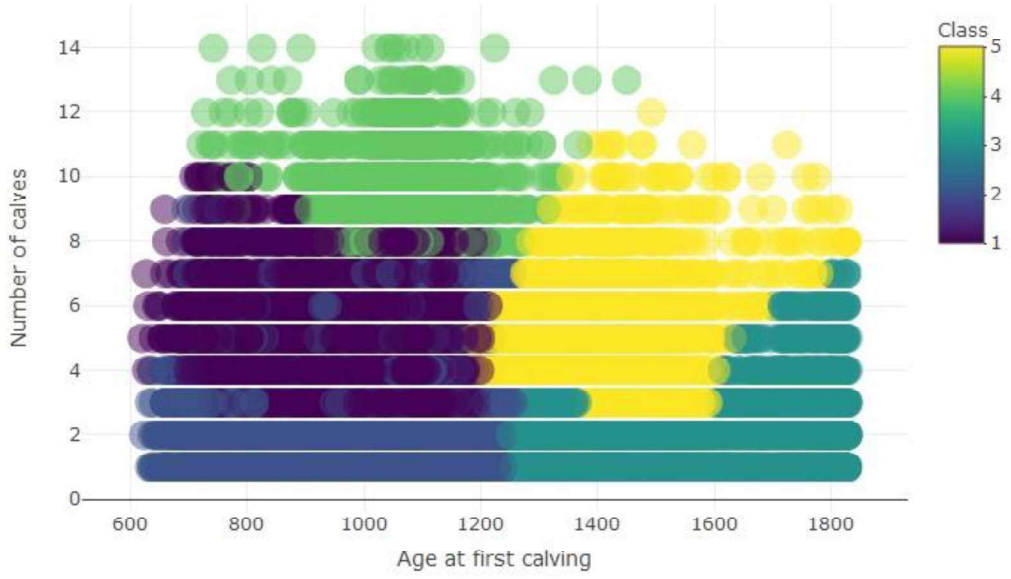


Figure 1. Classes obtained by Kohonen self-organizing map for all the individuals with phenotype, according to age at first calving (days) and number of calves.

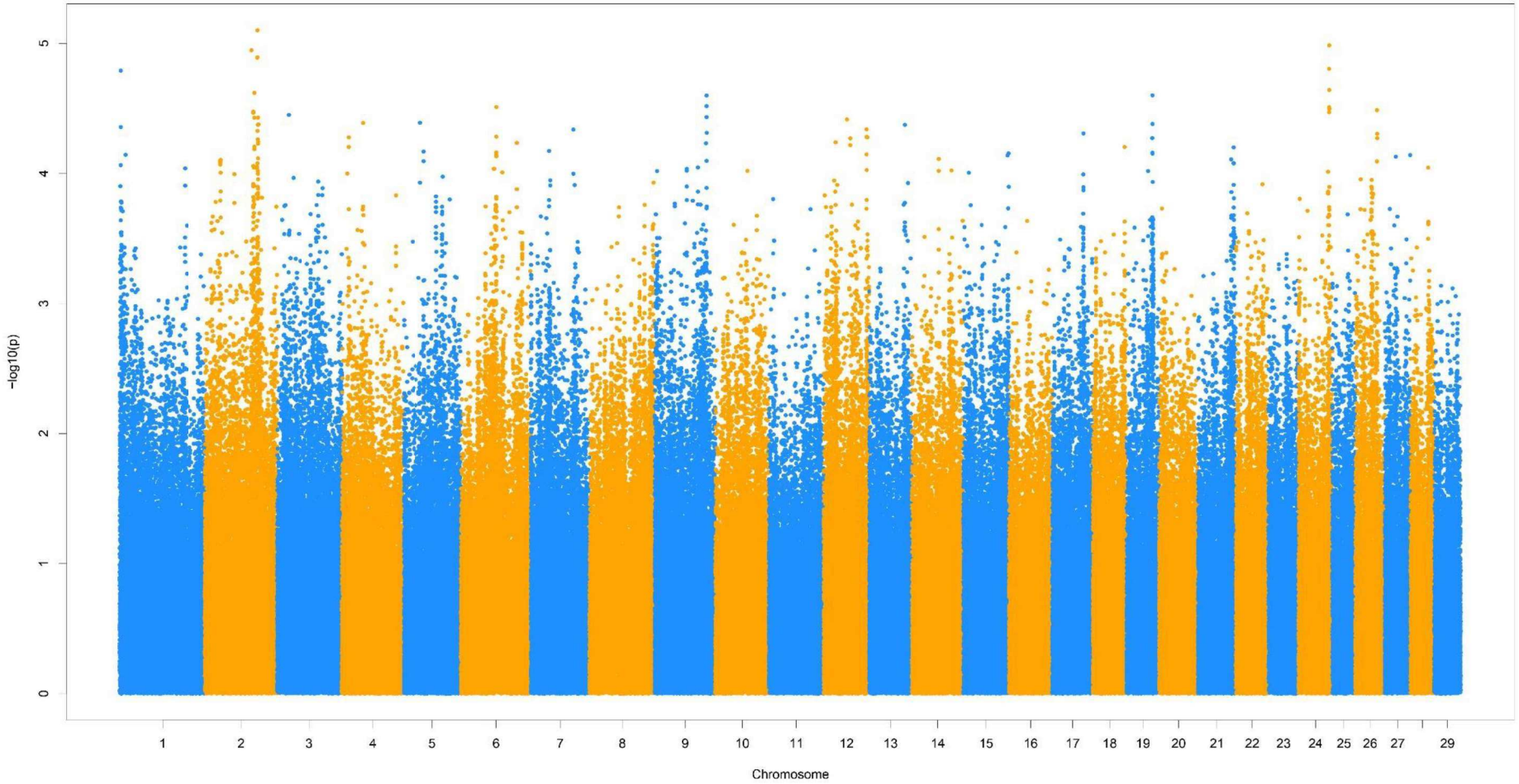


Figure 2. Manhattan plot of SNPs effects for stayability in Nellore cattle estimated by ssGBLUP method.

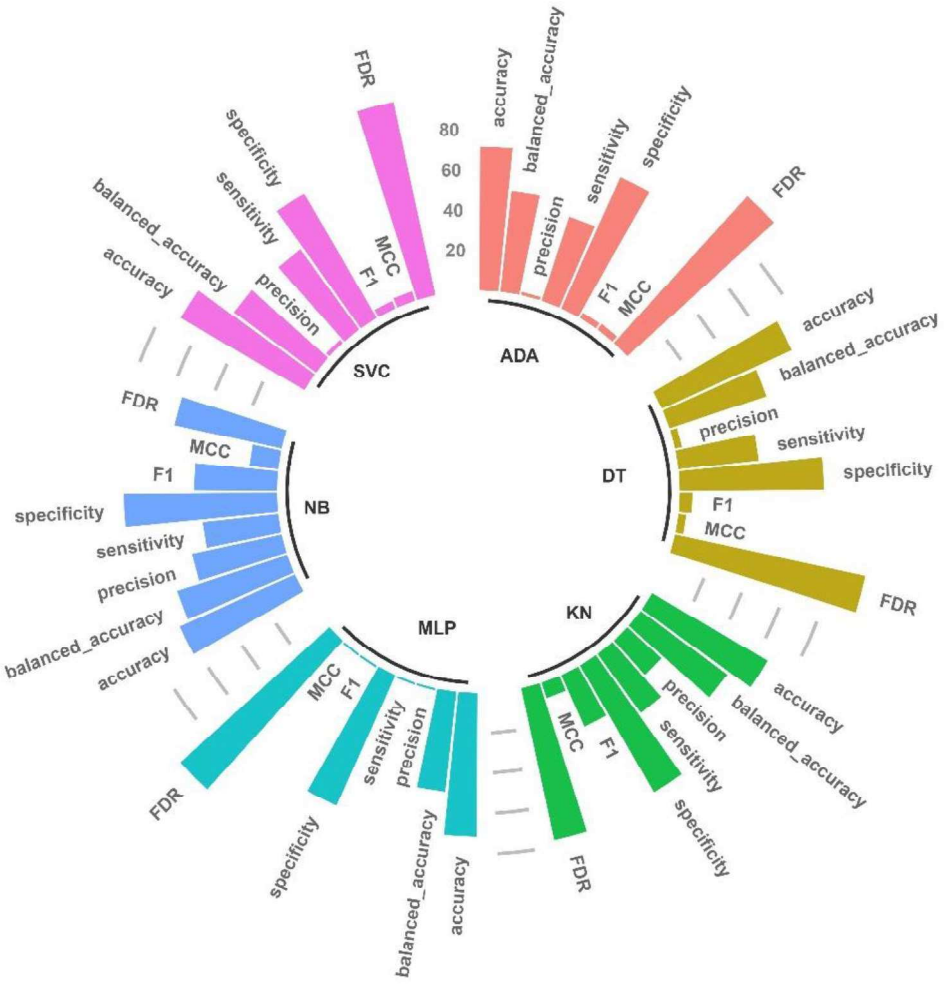


Figure 3. Average of evaluation metrics (accuracy, balanced accuracy, precision, sensitivity (recall), specificity, matthews correlation coefficient (MCC) and false discovery rate (FDR) for the machine learning classifiers (AdaBoost (ADA), Bernoulli Naïve Bayes (NB), Decision Tree (DT), Nearest Neighbors (KN), Multilayer Perceptron with one layer, one neuron and identity activation function (MLP) and Support Vector Machine for Classification (SVC)) from test sets with thousand markers

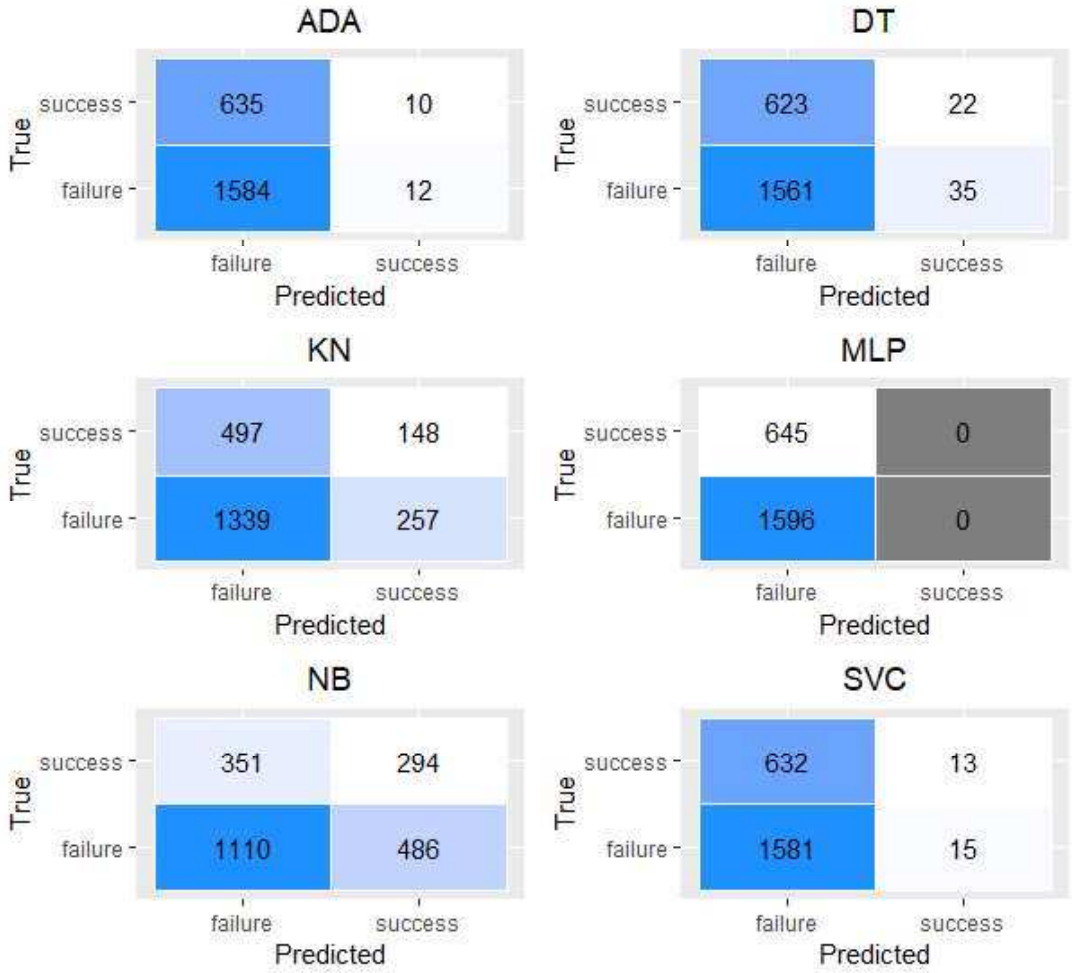


Figure 4. Average confusion matrix for the failure and success classes predicted by machine learning classifiers (AdaBoost - ADA, Bernoulli Naïve Bayes - NB, Decision Tree - DT, Nearest Neighbors - KN, Multilayer Perceptron with one layer, one neuron and identity activation function - MLP and Support Vector Machine for Classification - SVC) from test sets with thousand markers.

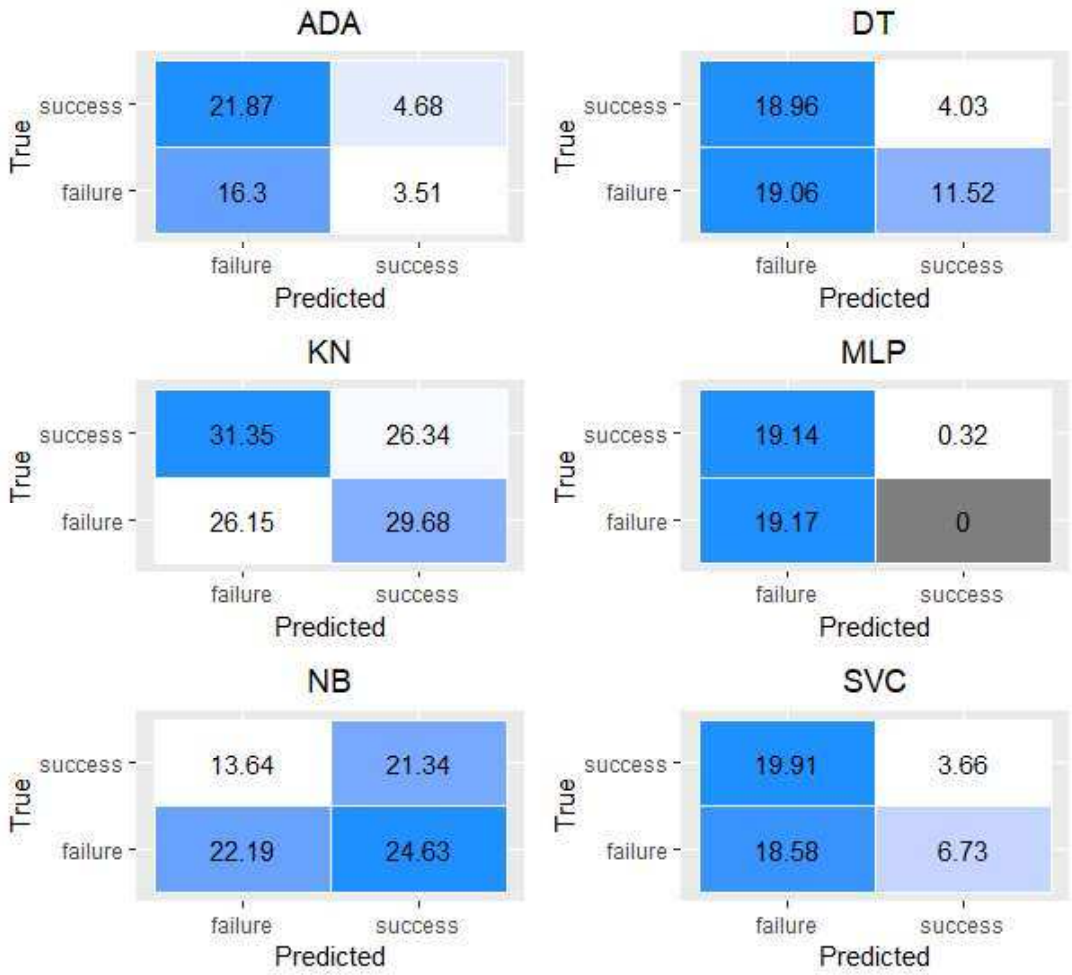


Figura 5. Confusion matrix in standard deviation for the failure and success classes predicted by machine learning classifiers (AdaBoost - ADA, Bernoulli Naïve Bayes - NB, Decision Tree - DT, Nearest Neighbors - KN, Multilayer Perceptron with one layer, one neuron and identity activation function - MLP and Support Vector Machine for Classification - SVC) from test sets with thousand markers.

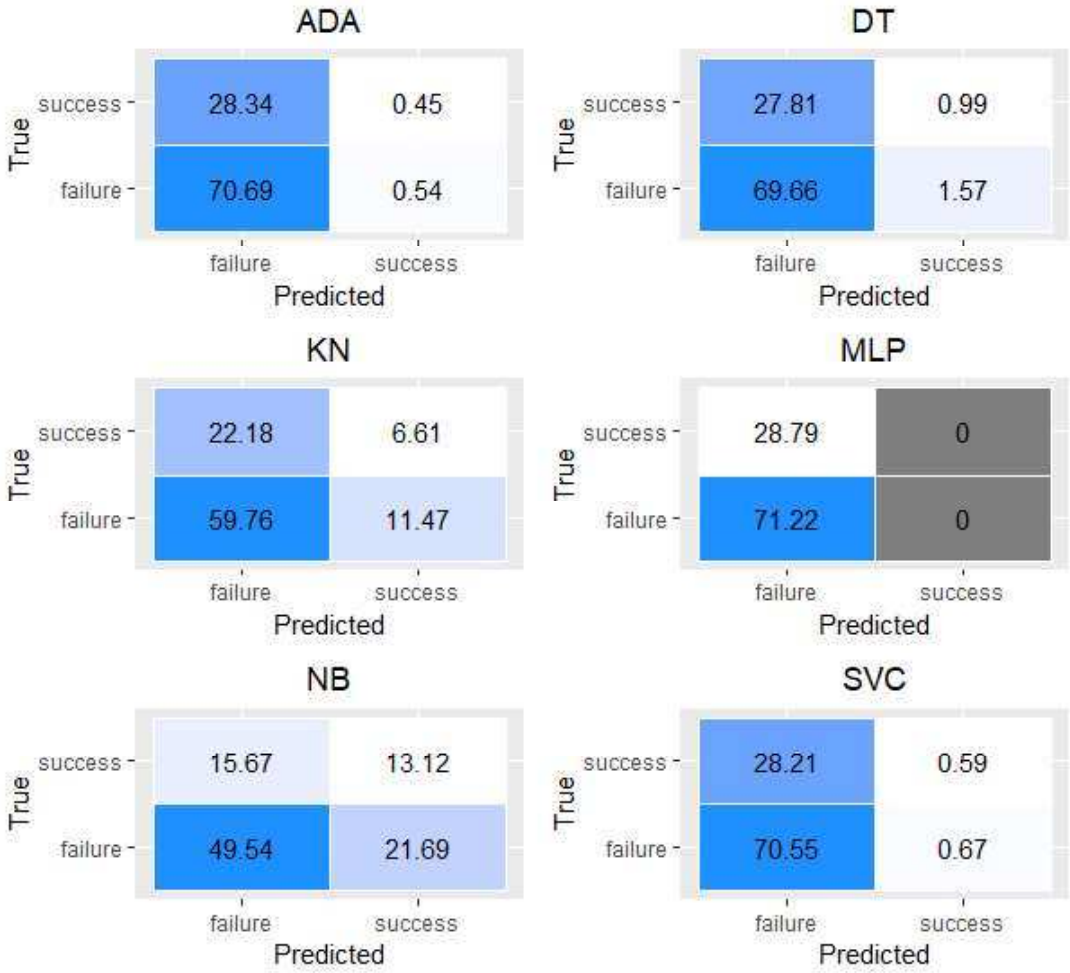
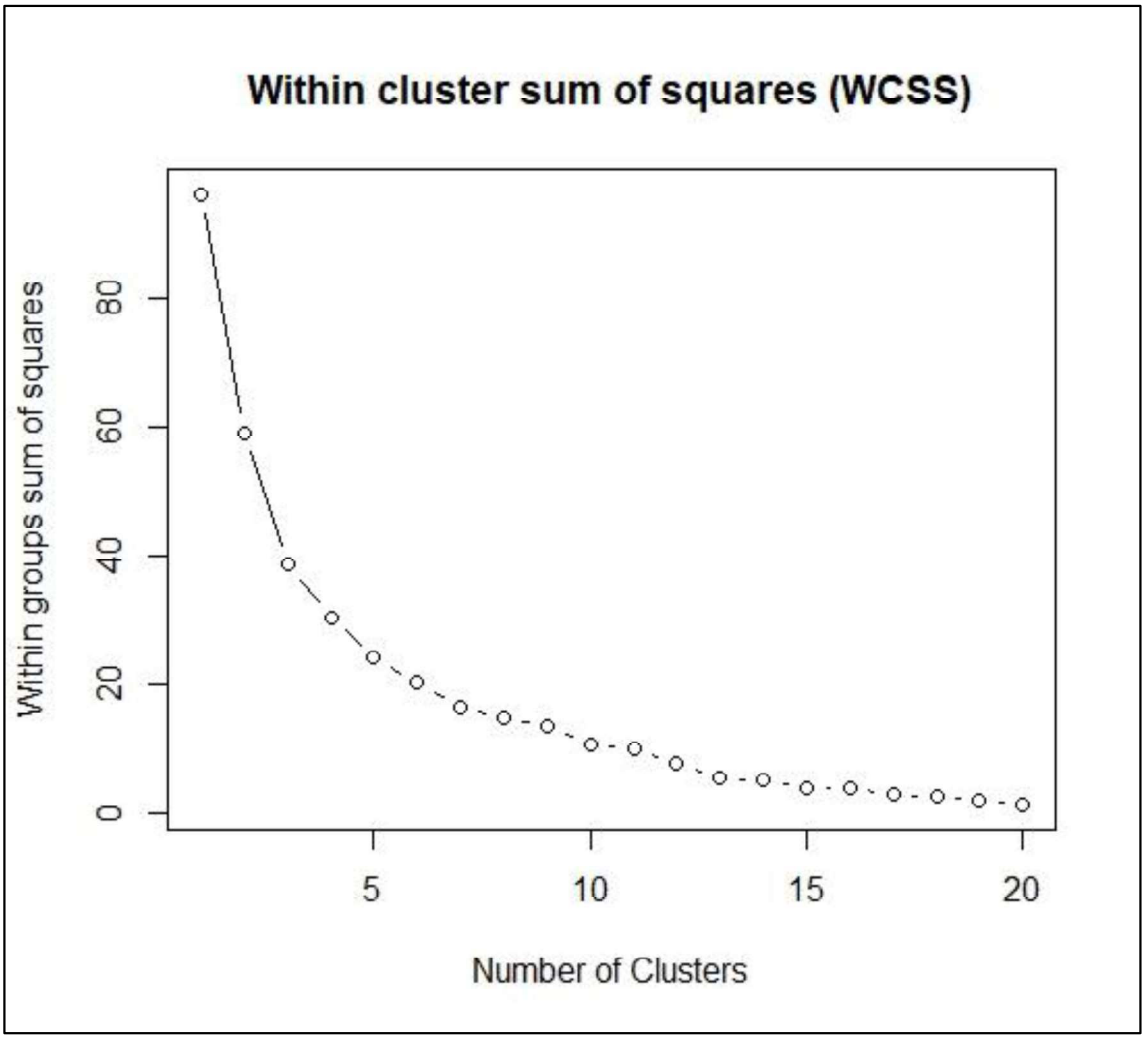


Figure 6. Confusion matrix in percentage for the failure and success classes predicted by machine learning classifiers (AdaBoost - ADA, Bernoulli Naïve Bayes - NB, Decision Tree - DT, Nearest Neighbors - KN, Multilayer Perceptron with one layer, one neuron and identity activation function - MLP and Support Vector Machine for Classification - SVC) from test sets with thousand markers.

Appendices

Appendix 1. Within cluster sum of squares for all classes used to obtain the best matching unit (BMU) by Kohonen map.



Appendix 2. Number of markers (n) selected per chromosome (Chr) in each genomic data set.

SNP_1k		SNP_3k		SNP_5k	
Chr	n	Chr	n	Chr	N
1	-	1	175	1	293
2	130	2	285	2	436
3	51	3	191	3	314
4	24	4	71	4	134
5	38	5	105	5	182
6	76	6	185	6	285
7	36	7	138	7	223
8	24	8	97	8	182
9	89	9	209	9	316
10	23	10	90	10	170
11	7	11	23	11	48
12	80	12	224	12	352
13	16	13	67	13	130
14	18	14	75	14	142
15	26	15	105	15	190
16	3	16	23	16	49
17	33	17	104	17	170
18	19	18	69	18	129
19	61	19	157	19	233
20	8	20	38	20	78
21	40	21	102	21	170
22	21	22	78	22	125
23	7	23	32	23	69
24	39	24	86	24	133
25	4	25	31	25	59
26	46	26	120	26	175
27	16	27	61	27	101
28	10	28	40	28	79
29	-	29	19	29	33

Chapter 4

4.1 General conclusions

This research addressed two new methods under viewpoint of genetic evaluation of crossbred cattle and genome-enabled classification, which allowed to estimate non-additive effects important in crossbred population and to challenge machine algorithms to learning with biological noise and sires' attributes to classify the daughters.

The singularity of crossbred animals is often associated to heterosis and the complementarity between breeds and the further away parental breeds more evident is the increase in productivity, however, researchers have described difficulty in quantifying these effects. In this sense, this study provided a powerful method to reduce the estimation problems coming from non-additives effects. The importance for each genetic covariate (non-additive effects) has been proven in variance terms and for the first time the broad-sense hereditability for weaning weight was calculated. Effects frequently assumed as important (maternal non-additive genetic effects and both breed additive effects are not relevant) were reset, opposite to several studies using Ridge Regression. In addition to benefits statistical promoted by dimensionality reduction, the BayesB model might reduce computational demand, processing time given that enable the estimation of non-additive effects in single step with breeding value prediction (random effects), in other words, without additional analysis as it is currently done, which makes BayesB model very attractive for application in breeding programs of crossbred beef cattle.

In the genome-wide selection field, new statistical methods have been proposed in order to minimize the side effects (high-dimensionality and multicollinearity) coming from simultaneous estimation of SNPs. However, the studies applied to genomic classification with machine learning are few. This is the first work that propose genome-enabled classification by several machine learning frameworks for reproductive trait and, mainly, challenging the algorithms learning to biological noise. In this sense, was seen that ML

frameworks simplest, as Naïve Bayes, might solve complex issue of classification.