

Supervised Learning Algorithms Evaluation on Sweet Potato Data for Production Indices Considering Human Consumption

Avaliação de Algoritmos de Aprendizagem Supervisionada em Dados de Batata-Doce Para Índices de Produção Considerando o Consumo Humano

Mardem Arantes de Castro¹, Ranulfo Mascari Neto¹, Joaquim Quinteiro Uchôa¹, Jesimar da Silva Arantes¹, Orlando Gonçalves Brito², Valter Carvalho de Andrade Júnior², Jeferson Carlos de Oliveira Silva², Renato Ramos da Silva^{1*}

Abstract: The sweet potato holds significant importance as a root crop cultivated globally, serving purposes in both human and animal nutrition, ethanol fuel production, and ornamental cultivation. Several genetic experiments have been undertaken to identify superior root varieties. Within this study, a refined statistical methodology has been devised to mitigate environmental factors and promote equitable comparisons among diverse varieties. Through the integration of supplementary environmental data into the experimental framework, a deeper understanding of each variety can be attained. This research signifies an inaugural endeavor towards scrutinizing post-harvest sweet potato data, aiming to unveil novel insights.

Keywords: Sweet Potato — Supervised learning algorithms — Machine Learning

Resumo: A batata-doce é uma importante cultura de raiz produzida em todo o mundo, utilizada para consumo humano e animal, produção de etanol combustível e para fins ornamentais. Experimentos genéticos foram conduzidos para selecionar variedades de raízes superiores. Neste estudo, foi desenvolvido um método estatístico sofisticado para minimizar as influências ambientais e facilitar comparações justas entre diferentes variedades. Ao incorporar dados ambientais adicionais no desenho experimental, é possível obter mais informações sobre cada variedade. Esta pesquisa representa um passo inicial na análise de dados pós-colheita da batata-doce para descobrir novas evidências.

Palavras-Chave: Batata doce — Algoritmos de aprendizado supervisionado — Aprendizado de máquina

¹ Department of Applied Computing, Institute of Exact and Technological Science, Federal University of Lavras, Lavras, Minas Gerais, Brazil

² Department of Agriculture, Lavras School of Agriculture Science, Federal University of Lavras, Lavras, Minas Gerais, Brazil

*Corresponding author: renato.ramos@ufla.br

DOI: <https://doi.org/10.22456/2175-2745.140067> • Received: 06/05/2024 • Accepted: 27/06/2024

CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introduction

The sweet potato (*Ipomoea batatas* (L.) Lam.) is cultivated in 111 countries worldwide, yielding a total production of 105.19 million tons across approximately 8.62 million hectares of land [1]. Approximately 61.5% of this production is concentrated in Asia, 33.7% in Africa, 3.8% in the Americas, and 1% in other regions globally. China leads in production, contributing approximately 47.62 million tons in the year 2021 [1].

The sweet potato is a tropical root crop that holds significant socioeconomic importance globally, particularly recognized for its vital role in ensuring food security, especially within underdeveloped nations [2]. Food security refers to the

condition wherein all individuals, at any given time, possess both physical and economic access to nutritious food that fulfills the requirements for an active and healthy lifestyle [3]. Sweet potatoes play a substantial role in human nutrition, providing essential energy, vitamins, and minerals [4]. Consequently, sweet potatoes are regarded as a promising solution for addressing nutritional deficiencies among population groups experiencing deficits in calorie and nutrient intake.

As previously noted, sweet potatoes constitute a highly nutritious and readily available food source, possessing significant socioeconomic importance. Consequently, there arises a need to optimize their productivity, consistently striving to

enhance both quantity and quality of production. To address this objective, our project endeavors to identify the optimal combination of traits within a given sweet potato plant that maximizes the likelihood of achieving high productivity. To tackle this challenge, we employ machine learning techniques.

Machine Learning constitutes a subfield of computer science dedicated to developing algorithms that rely on a dataset of examples. It involves a two-step process in problem-solving: 1) Acquisition of a dataset comprising examples, and 2) Construction of an algorithm that embodies a statistical model based on the provided dataset. This statistical model operates under certain assumptions regarding problem-solving strategies [5].

The collection of examples is referred to as a dataset. In our scenario, the dataset comprises examples from a sweet potato experiment conducted in the olericulture sector of the Federal University of Lavras (UFLA), located in Lavras, MG, Brazil.

In this instance, the objective is to construct a statistical model capable of accurately predicting the commercial root yield of the sweet potato experiment, utilizing the data gathered from the experiment. Within this dataset, each instance represents a distinct sweet potato, possessing numerous attributes utilized in the prediction of productivity.

This paper is structured as follows: Section 2 outlines the significance of sweet potatoes and details the experimental methodology employed to generate analytical data. This section also covers the fundamentals of the subfield of machine learning utilized in this study, namely supervised learning, including various models and their evaluation methods. Section 3 discusses the analytical procedures undertaken, detailing the steps and resources involved. Section 4 presents a comprehensive discussion of the analysis results. Finally, Section 5 provides the conclusion.

2. Concepts

This section aims to summarize significant concepts from the fields of agronomy and computer science, with a specific focus on their application in horticulture and machine learning. This overview intends to establish a foundation for comprehending the objectives of this study.

2.1 Sweet Potato Data

Sweet potatoes constitute a significant vegetable contributing to Brazilian food security, with approximately 20% of Brazilian municipalities relying on them for both subsistence and commercial purposes [6]. In terms of human nutrition, sweet potatoes stand out as a notable source of energy, vitamins, and minerals [4]. Furthermore, the plant displays potential as a functional food owing to its abundant content of bioactive compounds such as carotenoids and anthocyanins [7, 8, 9], antioxidants [7], and phytonutrients.

The crop displays considerable genetic diversity, which holds significant relevance for genetic breeding endeavors.

This attribute stems from the species' high potential for recombination. Consequently, botanical seeds present a substantial opportunity for utilization in breeding programs aimed at developing new varieties. Each obtained seed possesses the potential to generate a novel commercial genotype. This process of recombination, seed acquisition, and screening forms the foundation for the selection and breeding of sweet potato crops, often involving multiple stages of polycross and recurrent selection [10].

The experiment was conducted at the experimental area of the Olericulture Sector of the Federal University of Lavras (UFLA), located in Lavras, MG. The site is positioned at latitude 21°14'S, longitude 45°00'W, with an altitude of 918 meters. The soil type at the site is classified as a typical Distroferic Red Latosol (LVdf). According to Köppen's climate classification, the region falls under type Cwa, characterized by dry winters and predominant rainfall in the summer, with an average total annual precipitation of 1530 mm and an average annual temperature of 19.4°C [11]. Further details regarding the experimental conditions can be accessed in [12].

The experimental design employed was incomplete rows and columns, similar to the Honyaku design. The standard treatments consisted of 1,604 half-sibling sweet potato genotypes, derived from 31 progenies. Additionally, two control treatments from the Diamantina germplasm bank (UFVJM 58 and UFVJM 61) were incorporated. In order to quantify the pure experimental error, the controls were evaluated with repetitions, comprising 49 plots of the UFVJM 58 genotype and 47 plots of the UFVJM 61 genotype, resulting in a total of 1,723 plots installed in the experiment. They were randomized with the stipulation of maintaining a minimum proximity between related half-sibling genotypes (same mother) [12].

The incomplete block design enables optimization of the experimental area, resulting in savings in implementation, conduct, and evaluation efforts. It reduces the required number of controls, enhances efficiency in the evaluation of parameters, and maximizes results by allowing the simultaneous testing of numerous genotypes. The design plan for the incomplete block design considered the factors of row and column blocks, while also integrating constraints on the kinship of progenies within each matrix. This approach aimed to distribute treatments based on kinship among progenies and between siblings within the layout.

The experimental plots were arranged in windrows, each measuring 0.9 m x 1.0 m, resulting in a total area of 0.9 m² and representing the experimental unit (EU). Within each EU, three branches, each measuring 0.3 m in length and containing 6 to 8 buds, were planted, spaced 0.3 m apart. Overall, 17 windrows, each 90 m in length, were constructed, with a spacing of 1.0 m between them. The total area of the experiment was 0.18 ha (90 m x 20 m). In this design, experimental error is controlled in two directions: along the rows and columns. Thus, two types of blocks were established. The 17 windrow blocks measured 90.0 m x 1.0 m, with each block containing 100 experimental plots. The 100 transverse blocks (across the

windrows) had dimensions of 17.0 m x 1.0 m, each comprising 17 experimental plots.

2.2 Supervised learning

Machines can “learn” through various methods: supervised, semi-supervised, unsupervised, or reinforcement learning. Each type of learning employs distinct techniques and fulfills specific purposes. For the task of sweet potato regression, our team determined that supervised learning was the most appropriate approach.

In this type of learning, each instance of the dataset fed to the algorithm is labeled, meaning it is accompanied by its desired solution [13, 14]. The dataset provided to the machine learning algorithm is referred to as the training dataset or training set, and the process of constructing this model is termed training [14].

In supervised learning, two typical learning tasks prevail: classification and regression. Classification entails predicting a discrete value that represents the class to which an example belongs, given the dataset. For example, a common application is predicting whether an email is spam or not (i.e., spam filtering). In such scenarios, it is customary to assign an output of 0 for non-spam emails and an output of 1 for spam emails [13].

Regression, the objective of this project, entails predicting a target numeric value, such as the productivity of a plant, based on a set of features (such as color, mass, planting site, etc.) [13]. For both of these tasks, it is imperative to train the system by providing it with numerous examples of instances that require prediction (e.g., emails or plants). Nevertheless, a challenge arises: how can the algorithm generalize if it is already provided with the correct output? Fortunately, solutions to this problem exist.

2.2.1 Models

The project utilized three primary models: Random Forest, Support Vector Machines, and Lasso Linear Regression. These models will be introduced in the following subsection.

Random Forests consist of an ensemble of many decision trees. A decision tree, or tree-based model, is an algorithm that recursively partitions a dataset based on a measure such as information gain. It compares the dataset before and after splitting, selecting the attribute with the most information gain to divide the dataset. Once each tree is constructed, they collectively make predictions, and the average of these values is chosen as the Random Forest prediction [15, 16].

The Support Vector Machine is an algorithm initially designed for classification but can be adapted for prediction tasks such as regression. The concept is straightforward: given a dataset, it seeks to fit a regression line with margins that encompass as many instances as possible while constraining instances outside the margin [17, 18].

Lasso Regression is a form of regularized linear regression. Regularization is a technique that restricts the model, reducing its degrees of freedom to mitigate overfitting, a concept which will be discussed in the following subsection. Linear

regression is the simplest regression model, aiming to fit a straight line to the examples. Due to its simplicity, training this model is relatively fast [19, 20].

2.2.2 Evaluation

Once your model is trained, how do you determine its performance on new cases? In our scenario, the model may perform admirably for a specific set of plants in the experiment, yet it might struggle to generalize to new instances, rendering it inadequate. A proficient “learner” is a model capable of discerning the general pattern from the training samples in a manner that allows it to apply the learned pattern to every new example [21]. However, a model can “learn too well” from the training set, identifying patterns in the noise of the data itself, which may hinder its ability to generalize to new instances. Nonetheless, the model may exhibit exceptional performance on the training set [13]. In machine learning, this phenomenon is termed “overfitting”, and its counterpart, where the model fails to capture the data trend, is termed “underfitting” [21]. Fortunately, solutions exist for both of these issues.

For overfitting, solutions include selecting a new model with fewer parameters and less complexity, gathering more training data, or reducing the noise in the data (e.g., fix data errors and removing outliers) [13]. To address underfitting, options include selecting a more powerful model capable of capturing the general complex pattern of the data, providing better features to the algorithm, or reduce the constraints of the model (regularization) [13].

The sole method to ascertain whether the model will generalize effectively to new cases is by testing it on fresh data. However, deploying our model directly into production is not recommended, as it may yield unsatisfactory results and lead to user complaints. A more prudent approach involves evaluating the model before deployment by partitioning the data into two sets: the training set and the validation set. The model is trained on the training set and assessed on the validation set. This methodology aids in identifying overfitting by scrutinizing errors. If the training error is low and the test error is high, it suggests the model may be overfitting the training set.

There are several metrics available to evaluate the model’s performance. In our case, we will employ the Root Mean Squared Error (RMSE), given the nature of our regression task. RMSE quantifies the standard deviation of the residuals, which represent the distances between the data points and the regression line. This metric provides insight into the concentration of data points around the regression line [22].

3. Methodology

This section provides a detailed account of the methodology employed in the project, delineating each step. Such thoroughness is essential to substantiate the credibility of the results.

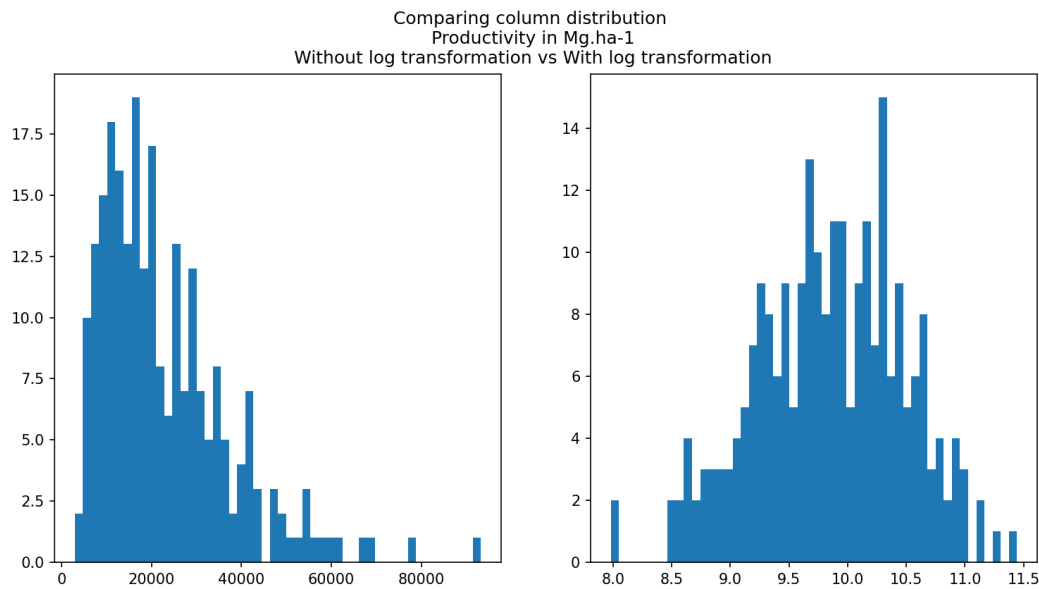


Figure 1. Distribution of production class.

3.1 Data Information

Measurements of gas exchange (physiological data) were recorded using an infrared gas analyzer (IRGA), model LI-6400XT (Li-COR Biosciences, Lincoln, Nebraska, USA), equipped with a 6 cm² chamber and an artificial red/blue LED light source (LI6400-02B, Li-Cor, Lincoln, Nebraska, USA) at 140 days. On this date, leaf water potential (Ψ h expressed in Mpa) was also assessed using a Scholander-type pressure chamber (model 1000, PMS Instruments, USA). Chlorophyll index assessment using the Konica Minolta Spad-502plus sensor (AUCCloro) was performed at 90, 125, and 150 DAP, providing a chlorophyll index gradient from 0 to 99. Agronomic data collection was performed at harvest, 150 days after planting. On this date, characteristics related to root yield and quality were evaluated. Total root production was quantified by weighing all the roots from the plot on a scale. For commercial production of the plot, only roots with commercial size and shape compatible with the market, weighing between 100 and 800 g and free from insect damage, were weighed. The number of roots per plot was obtained by counting. Rejects were considered as the total root production minus the commercial production. Analyses related to root quality were performed using rating scales (values from 1 to 5). Ratings were assigned to each plot, evaluating characteristics such as insect damage, commercial standard, and root shape. The colors of the root skin and flesh were also analyzed using rating scales described by [23]. After obtaining the data, they were tabulated at the plot mean level, disregarding the effect per plant (clone).

The dataset consists of 300 instances with 144 attributes. Exploratory Data Analysis (EDA) revealed that out of these 144 attributes, 140 are numerical and 4 are categorical. During EDA, several issues were identified: many columns were

poorly distributed, with a large portion of their values being zero, some columns acted as identifiers, others had highly similar distributions, and some were directly related to the target attribute, necessitating their removal. Furthermore, preprocessing was essential due to numerous outliers in several attributes. For instance, in the target column intended for prediction, negative values were found. This doesn't make sense in our context (for example, sweet potato productivity cannot be negative), so these values were replaced with their respective positive counterparts. In other cases, outliers in other columns were replaced with the median of the column, as it is less influenced by extreme values compared to the mean.

Another preprocessing step involved applying the logarithm function to asymmetric distributions, which helps approximate them to a normal distribution, considered more suitable for machine learning models. Additionally, all numerical columns were normalized, and missing values were replaced with the median of each column. For categorical columns, the "OneHotEncoder" class from scikit-learn was used, creating a new column for each categorical value in those columns.

In Figure 1, it presented the distribution of production class.

3.2 Data Cleaning

Initially, the utilized dataset contained erroneous values (e.g., "2.4"). These inaccuracies were rectified manually. Subsequently, during a preliminary data analysis, it was noted that the dataset comprised 300 rows and 140 columns, of which 4 were initially classified as categorical columns. However, upon closer examination, only 2 categorical columns were utilized, as duplicates were identified and a column was des-

ignated to identify each row.

Subsequently, the numerical columns underwent analysis through the plotting of histograms for each one. These plots facilitated the identification of incorrect values within the columns, unusual distributions prompting the removal of the column, noteworthy columns exhibiting favorable distributions, and the identification of outliers.

Following the histogram analysis, numerous columns were discarded based on the following criteria: a high prevalence of zeros, distributions nearly identical to other columns, or columns calculated using the target column.

After the removal of columns, outliers were addressed through manual correction, either by replacing them with the median value or converting negative outliers to positive values.

3.3 Preprocessing

Another type of analysis conducted pertained to the skewness of the data. A common and accepted approach involves log-transforming columns to mitigate attribute skewness. Skewed data can compromise the performance of machine learning models.

Upon completion of data cleaning, it was deemed advantageous to partition a test dataset to simulate unseen values. This task was accomplished using the Scikit-Learn library's train-test-split function, where 20% of the data was randomly sampled as the test set, with the remaining 80% allocated to the training set.

Following the partitioning of data into training and testing sets, the features were categorized into three groups based on their treatment requirements. The first group comprised numeric attributes with high skewness, where null values were replaced with their median, followed by logarithmic transformation and standardization. The second group, consisting of numeric attributes with normal skewness, had null values replaced by their median and underwent standardization. The final group, encompassing categorical attributes, had null values replaced by their mode.

3.4 Model Selection

Next, the focus shifted to selecting the machine learning model best suited for this dataset. Three models were chosen for analysis: Random Forest (known for its generally robust performance), Support Vector Regression, and Lasso Regression (recognized for its efficacy with small datasets).

Regarding the models (random forest, lasso regression, and support vector regression), a randomized hyperparameter selection approach was employed. This involved selecting a range of values for each hyperparameter of each model, and then training the model using cross-validation with a randomly chosen hyperparameter configuration. This process was carried out for each of the models. The best results obtained for each model are:

Random Forest

- Number of trees: 104

- Maximum depth: 10

- Node splitting criterion: mean squared error

Lasso Regression

- Alpha (constant multiplying the L1 regularization term): 161

Support Vector Regression

- C (regularization parameter based on L2 regularization): 3863

Model selection was conducted by evaluating model errors through cross-validation and hyperparameter tuning. It was determined that Support Vector Regression (SVR) yielded the most favorable results, as presented in Table 1. Regarding the RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) metrics, it is ideal to obtain a lower value. In contrast, for the R^2 (Coefficient of Determination) metric, it is ideal to have a higher value (close to 1). Furthermore, both SVR and Random Forest provided a feature importance measure, which will be elaborated upon in the Results section.

Table 1. Comparison among regression algorithms.

	RMSE	MAE	R^2
Support Vector Regression	4656.85	2865.54	0.8724
Random Forest	7340.23	4032.57	0.6831
Lasso Regression	5241.10	3063.74	0.8384

3.5 Feature Selection

Finally, given the similarity observed among many features during the histogram analysis, it was deemed pertinent to conduct feature selection towards the project's conclusion. For this purpose, two methods were proposed: correlation-based and recursive feature elimination.

In the correlation-based method, a correlation matrix was computed, and if two features exhibited a correlation value exceeding 0.9, one of them would be eliminated. Consequently, 27 features were removed. However, it was observed that this method did not enhance the model's performance.

Conversely, the recursive feature elimination method operates as follows: Initially, it iterates through all the features and selects the one resulting in the lowest model error. Once selected, the feature is eliminated from the dataset and cannot be chosen again. This process is repeated until the model error ceases to decrease.

4. Results

In Figure 2, our model achieved a Root Mean Squared Error (RMSE) of 4656.85, indicating the average difference between the actual value and the prediction made by the model for a new instance. Additionally, the optimal performance was observed with 71 features. However, results close to the

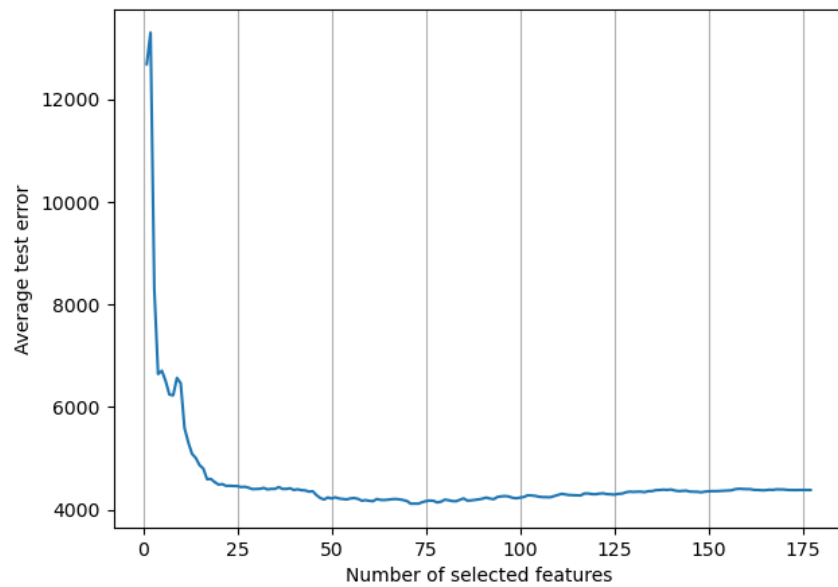


Figure 2. SVR model evaluation by features numbers.

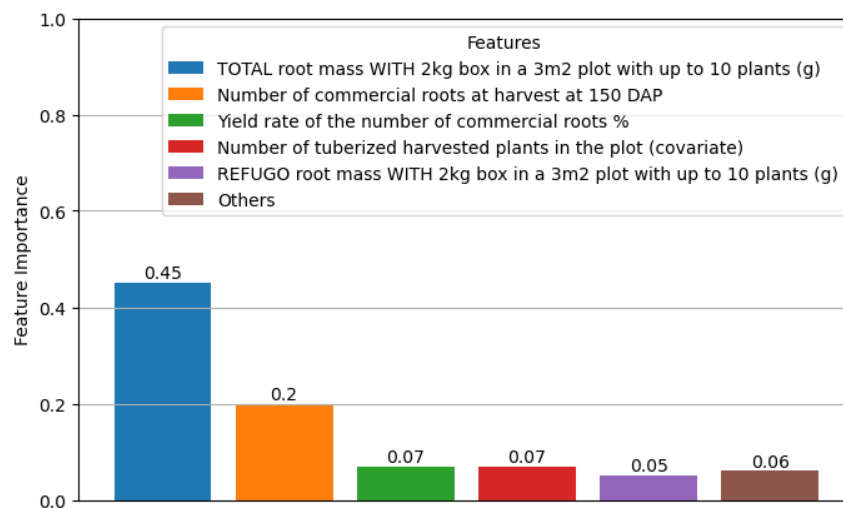


Figure 3. Feature selected from SVR model.

optimum are already obtained with just 20 features at a cost of worsening the result by just 8% in relation to the optimum.

The first figure illustrates the performance of the model concerning the number of features utilized. It is evident that the model's performance experiences a notable improvement when the number of features exceeds 3. Subsequently, another significant drop is observed near the 20s mark. Remarkably, the lowest average test error is attained with 71 features, comprising the subset selected for the model.

In terms of feature importance, a ranking of the 5 most important features was made, according to the final model used, represented in Figure 3. The most important feature was found to be a measure of the mass of the plant roots. Moreover, all the other feature importance were added and represented by the "Others" category.

The second graph illustrates that the feature with the highest importance in the model results has more than double the value of the second most impactful feature. In this figure, we can see that the five main characteristics correspond for approximately 94% of the importance in predicting productivity. The rest of the characteristics correspond for approximately 6% of the importance of productivity, meaning that we can only use this small number of characteristics to describe the final behavior of productivity without major losses.

Sweet potato is an extremely versatile crop with multiple applications [24] such as human consumption, animal feed [25], biofuel production [26], bioplastic production, agro-industrial use, among others. In this context, genetic improvement programs should focus their research on the most agronomically relevant traits and purposes of use. The total

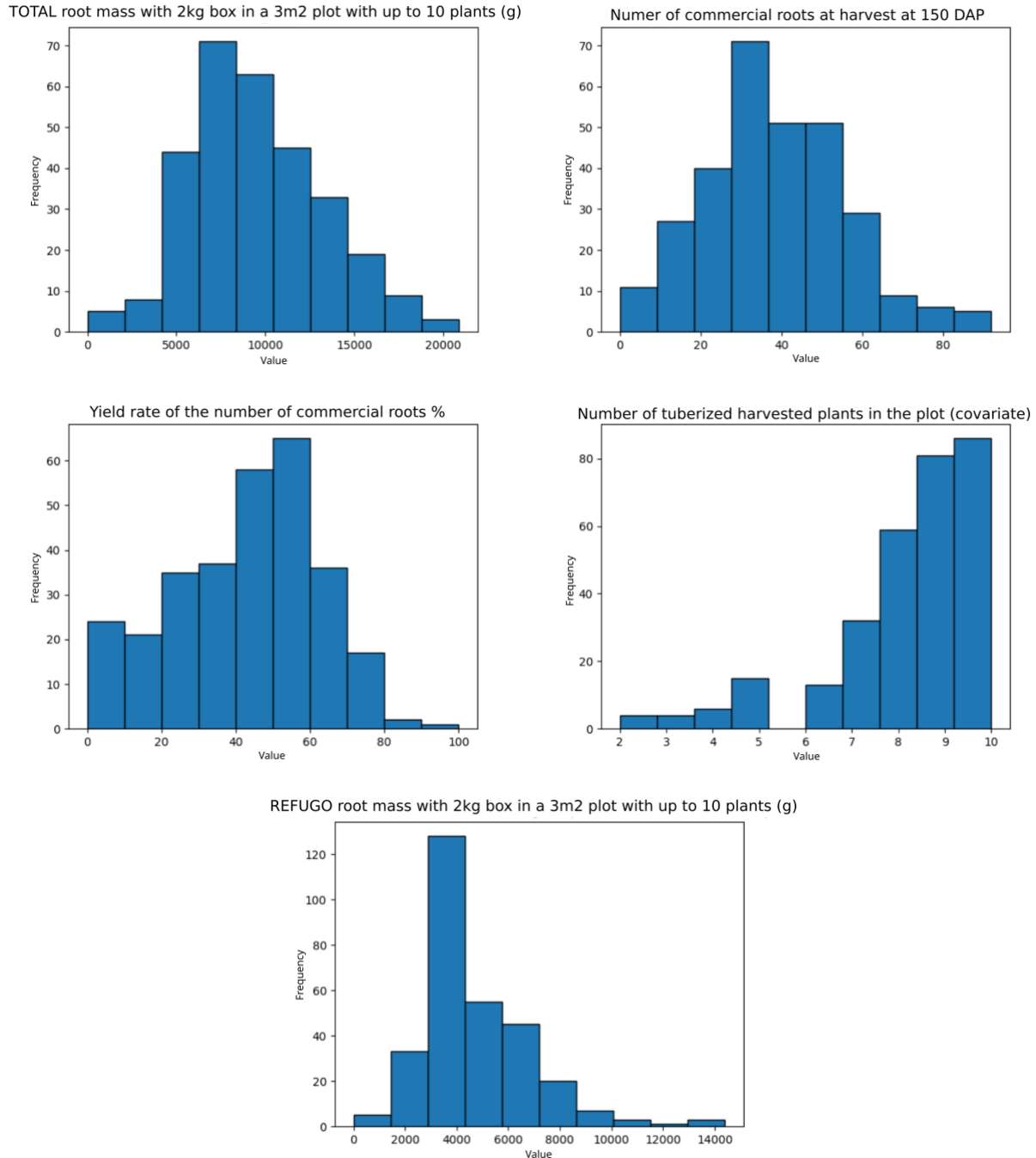


Figure 4. Histograms for features selected by SVR.

mass of roots, as illustrated in Figure 3, emerges as one of the most relevant variables for these studies. This aspect becomes prominent when selecting genotypes recommended for ethanol production, animal feed, as well as for the manufacture of processed foods in agro-industry, such as sweets and flours. This is because, in these applications, the visual quality of the roots is of little importance, as the main objective is starch production. On the other hand, when considering the human consumption market, characteristics related to high

root quality become fundamental, such as productivity and the number of commercial roots. Due to the high demands of this market, roots with excellent commercial appearance, adequate shape (usually fusiform), without damage caused by insects, and with an average weight between 100 and 800 g are required [27]. Roots that do not meet these criteria are not of interest, being classified as “reject” or “discard” and may be directed mainly for animal feed, either in their natural state or incorporated into other foods.

The third graph, Figure 4, presents a histogram for each of the features found by the method (SVR). These histograms show the distribution shape of each of the features.

5. Conclusion

This paper presents the initial contribution of machine learning algorithms to potato data exploration post-harvest for human consumption. While it is an ongoing project with preliminary results, it has revealed interesting insights that could complement [12].

Three methods were analyzed: Random Forest, Support Vector Regression, and Lasso Regression. Of the three methods, the one that presented the best results was the Support Vector Regression (SVR) method.

The primary outcome of this study, based on the aforementioned results, indicates that the model utilized can produce accurate predictions with a relatively low number of features (approximately 25), despite the availability of a large feature set (approximately 175). Moreover, five features demonstrate greater significance than all others. Consequently, these findings offer the potential to streamline experimental design for increased efficiency. This leads to a significant question: "Can the time of experimentation be reduced by planning data before harvest?" While this question remains unanswered for now, similar studies involving other cultivars have suggested its feasibility [28, 29].

Sweet potato is a crop with multiple agronomic uses, notable for both human and animal consumption of its roots. Various studies demonstrate that tuberous root production is the main criterion for selecting superior genotypes in the species. It is important to note that total root production is strongly associated with the search for genotypes suitable for industrial purposes, such as the production of animal feed, sweets, and flours. In these cases, the visual quality of the roots is less relevant, with a greater focus on starch yield and other qualitative traits such as taste and color. For the fresh market, which is the main focus of the species breeding programs, commercial production is extremely important, as it reflects the yield of roots that can be effectively marketed at a higher value. Additionally, the number of roots is a crucial trait, as it shows a high correlation with both total and commercial root productivity. Therefore, traits like these are extremely important for selecting high-performing clones, particularly in terms of high-quality root productivity. The variables identified as the most contributing to phenotypic differentiation among genotypes are indeed essential, as they reflect practical and indispensable data for selection.

Acknowledgements

The authors are grateful to PIVIC/UFLA and PIBIC/FAPEMIG programs for supporting this work.

Author Contributions

Mardem Arantes de Castro: performed the computations; contributed to the development of this study; methodology; preparation of the original drafting of the composition. Raulfo Mascari Neto: performed the computations; contributed to the development of this study. Joaquim Quinteiro Uchoa: conceived of the presented idea; contributed to the development of this study; discussed the results; proofreading. Jesimar da Silva Arantes: contributed to the development of this study; supervision; discussed the results; review-writing and editing. Orlando Gonçalves Brito: contributed to the development of this study; discussed the results; proofreading. Valter Carvalho de Andrade Junior: responsible for research into sweet potato cultivation at UFLA, conceived of the presented idea; discussed the results; proofreading. Jeferson Carlos de Oliveira Silva: Field experiment, data collection and proofreading. Renato Ramos da Silva: conceived of the presented idea; contributed to the development of this study; supervision; discussed the results; review-writing and editing.

References

- [1] FOOD, T.; (FAO), A. O. *FAO Statistics online database (FAOSTAT)*. 2023. Accessed: 09/08/2023. Disponível em: <https://www.fao.org/faostat/en/#data/QCL>.
- [2] FERNANDES, A. et al. Sistema de produção de batata-doce. Brasília, DF: Embrapa Hortaliças, 2021., 2021.
- [3] WHAT is Food Security? There are Four Dimensions. 2023. Disponível em: <https://www.worldbank.org/en/topic/agriculture/brief/food-security-update/what-is-food-security#:~:text=Based%20on%20the%201996%20World,an%20active%20and%20healthy%20life.> > Accessed: 09/11/2023.
- [4] VIEIRA, T. d. S. et al. Efeito da substituição da farinha de trigo no desenvolvimento de biscoitos sem glúten. *Brazilian Journal of Food Technology*, SciELO Brasil, v. 18, p. 285–292, 2015.
- [5] ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2020.
- [6] MELLO, A. F. S. A importância socioeconômica da batata-doce para a agricultura brasileira. *Brasília: Embrapa Hortaliças*, 2015. Accessed: 13/01/2024.
- [7] VIZZOTTO, M. et al. Physicochemical and antioxidant capacity analysis of colored sweet potato genotypes: In natura and thermally processed. *Ciência Rural*, v. 47, 01 2017.
- [8] LAURIE, S. M.; FABER, M.; CLAASEN, N. Incorporating orange-fleshed sweet potato into the food system as a strategy for improved nutrition: The context of south africa. *Food Research International*, v. 104, p. 77–85, 2018. ISSN 0963-9969. Balanced diets in food systems: emerging trends and challenges for human health and

- wellbeing. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0963996917305781>.
- [9] ZHANG, X. et al. The modulatory effect of anthocyanins from purple sweet potato on human intestinal microbiota in vitro. *Journal of agricultural and food chemistry*, v. 64, 03 2016.
- [10] GRÜNEBERG ROBERT MWANGA, M. A. W.; ESPINOZA, J. Breeding clonally propagated crops. In: *Plant breeding and farmer participation*. Rome, Italy: Food and Agriculture Organization of the United Nations, 2009. cap. 13, p. 275–322.
- [11] FERREIRA, A. A. A. D. L. G. de C. E. Classificação e tendência climática em lavras, mg. *Ciência e Agrotecnologia*, v. 31, n. 6, p. 1862–1866, 2007.
- [12] SILVA, J. C. D. O. *Seleção de clones de batata-doce para diferentes aptidões agrônomicas*. Tese (Master thesis) — Universidade Federal de Lavras, Lavras, MG, October 2019. Available at <http://repositorio.ufla.br/handle/1/37472>.
- [13] GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. [S.l.]: "O'Reilly Media, Inc.", 2022.
- [14] VEMURI, V. K. The hundred-page machine learning book. *Journal of Information Technology Case and Application Research*, Taylor & Francis, v. 22, n. 2, p. 136–138, 2020.
- [15] LIU, Y.; WANG, Y.; ZHANG, J. New machine learning algorithm: Random forest. In: LIU, B.; MA, M.; CHANG, J. (Ed.). *Information Computing and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 246–252. ISBN 978-3-642-34062-8.
- [16] BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, Oct 2001. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1010933404324>.
- [17] SOMVANSHI, M. et al. A review of machine learning techniques using decision tree and support vector machine. In: IEEE. *2016 international conference on computing communication control and automation (ICCUBEA)*. [S.l.], 2016. p. 1–7.
- [18] GAMMERMANN, A. Support vector machine learning algorithm and transduction. *Computational Statistics*, v. 15, n. 1, p. 31–39, Mar 2000. ISSN 1613-9658. Disponível em: <https://doi.org/10.1007/s001800050034>.
- [19] GERALDO-CAMPOS, L. A.; SORIA, J. J.; PANDO-EZCURRA, T. Machine learning for credit risk in the reactive peru program: a comparison of the lasso and ridge regression models. *Economies*, MDPI, v. 10, n. 8, p. 188, 2022.
- [20] MUTHUKRISHNAN, R.; ROHINI, R. Lasso: A feature selection technique in predictive modeling for machine learning. In: IEEE. *2016 IEEE international conference on advances in computer applications (ICACA)*. [S.l.], 2016. p. 18–20.
- [21] ZHOU, Z.-H. *Machine learning*. [S.l.]: Springer Nature, 2021.
- [22] GLEN., S. *RMSE: Root Mean Square Error*. 2023. From StatisticsHowTo.com: Elementary Statistics for the rest of us! <https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/>. Access in: november 28th.
- [23] COSTA, A. L. d. et al. Selection of superior sweet potato genotypes for human consumption via mixed models. *Bragantia*, SciELO Brasil, v. 81, p. e4122, 2022.
- [24] SILVA, J. C. de O. et al. Mixed model-based indices for selection of sweet potato genotypes for different agronomic aptitudes. *Euphytica*, Springer, v. 218, n. 7, p. 86, 2022.
- [25] DONATO, L. M. S. et al. Uso de ramas de batata-doce para produção de feno. *Ciência Animal Brasileira*, SciELO Brasil, v. 21, p. e–53493, 2020.
- [26] VIEIRA, A. D. et al. Avaliação agronômica de clones de batata doce com potencial para produção de etanol. *Brazilian Journal of Applied Technology for Agricultural Science/Revista Brasileira de Tecnologia Aplicada nas Ciências Agrárias*, v. 8, n. 1, 2015.
- [27] JÚNIOR, V. C. de A. et al. Características produtivas e qualitativas de ramas e raízes de batata-doce. *Horticultura brasileira*, SciELO Brasil, v. 30, p. 584–589, 2012.
- [28] AONO, A. H. et al. A divide-and-conquer approach for genomic prediction in rubber tree using machine learning. *Scientific Reports*, v. 12, n. 1, p. 18023, Oct 2022. ISSN 2045-2322. Disponível em: <https://doi.org/10.1038/s41598-022-20416-z>.
- [29] AONO, A. H. et al. A joint learning approach for genomic prediction in polyploid grasses. *Scientific Reports*, v. 12, n. 1, p. 12499, Jul 2022. ISSN 2045-2322. Disponível em: <https://doi.org/10.1038/s41598-022-16417-7>.