Boosting algorithms for prediction in agriculture: An application of Feature importance and Feature Selection

Boosting algorithms for prediction crop damage

Viviane Costa Silva
https://orcid.org/0000-0001-5110-9063
Mateus Silva Rocha
https://orcid.org/0000-0002-6607-3064
Glaucia Amorim Faria
https://orcid.org/0000-0003-2474-4840
Silvio Fernando Alves Xavier Junior
https://orcid.org/0000-0002-4832-0711
Tiago Almeida de Oliveira
https://orcid.org/0000-0003-4147-7721
Ana Patricia Bastos Peixoto
https://orcid.org/0000-0003-0690-1144

ABSTRACT - The Agriculture sector has created and collected large amounts of data. It can be gathered, stored, and analyzed to assist in decision making generating competitive value, and the use of Machine Learning techniques has been very effective for this market. In this work, a Machine Learning study was carried out using supervised classification models based on boosting to predict disease in a crop, thus identifying the model with the best areas under curve metrics. Light Gradient Boosting Machine, CatBoost Classifier, Extreme Gradient, Gradient Boosting Classifier, Adaboost models were used to qualify the crop as healthy or sick. One can see that the LightGBM algorithm provided a better fit to the data with an area under the curve of 0.76 under the use of BORUTA variable selection.

Keywords: Artificial Intelligence; Machine Learning; BORUTA.

INTRODUCTION

Agriculture is the preeminent source of livelihood that forms the backbone of Brazil. The current challenges of water scarcity, runaway costs due to demand-supply and climate uncertainty require farmers to be equipped with resourceful agriculture. In particular, low crop yields due to climate change, poor irrigation facilities, reduced soil fertility and traditional agricultural techniques need to be addressed. Agricultural production depends on climatic conditions, biological and economic causes. Weather stations play a crucial role, as they provide information about weather events that directly impact productivity. Therefore, the prediction of crop health that affects yield is a significant problem that must be the focus of the investigation.

Today large agricultural systems compete to produce more due to unpredictable climate changes that drastically reduce water resources. The productive sector has created and stored large amounts of data that can be gathered, stored and analyzed to assist in decision making, generating competitive value. New advanced methods and approaches such as machine learning can be used by gathering information from past production results of an agricultural system to estimate future-time production (prediction).

LIAKOS et al. (2018) carried out a review work on machine learning applications in agriculture. The studies analyzed were categorized into crop management, including applications in yield prediction, disease detection, weed detection, crop quality and species recognition; livestock management, as well as applications in animal welfare and production for beef cattle, water management, and soil management.

The filtering and classification of the articles presented demonstrated how agriculture already benefits from ML techniques. Several papers for plant disease and crop damage are all based on image information using the classifiers Support Vector Machine (SVM), Artificial Neural Network (ANN), XY-Fusion to make the predictions with high accuracy (PANTEZI et al., 2017; MOSHOU et al., 2014; MOSHOU et al., 2004; MOSHOU et al., 2005; MOSHOU et al., 2006; FERENTINOS, 2018).

Machine learning is a technique utilized to predict crop yields in agriculture, and thus assist the farmer in making a forecast. Various techniques such as classification, regression and clustering are used to predict the yield of a crop. Artificial neural network classifiers or regressors, support vector machines (SVM), linear and logistic regression, decision trees, Naive Bayes are some of the algorithms used to implement the prediction. In this sense, this article aims to present variable selection and explanation techniques using boosting-based supervised classification algorithms to predict disease in a crop.

MATERIAL AND METHODS

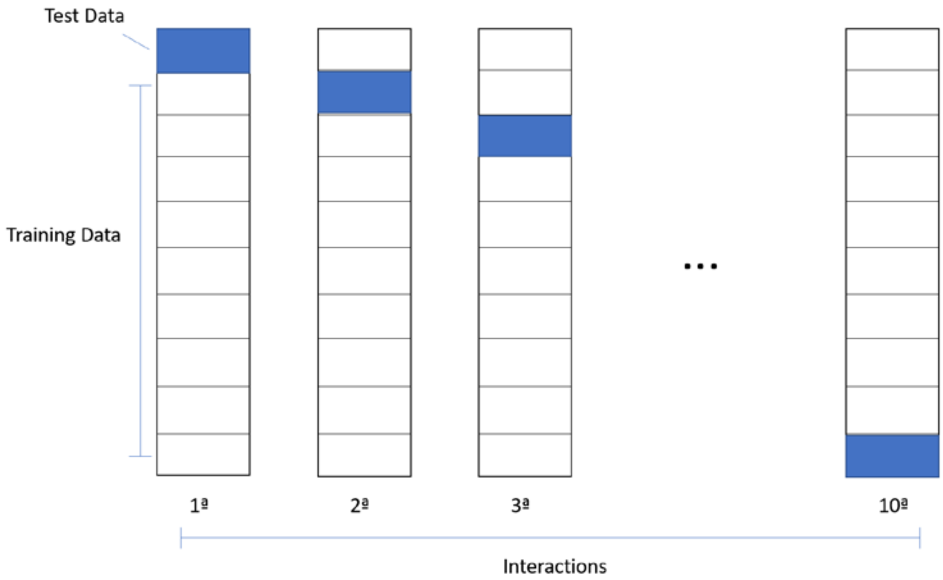
Dataset

The dataset used was obtained from the *hacktoon* competition (https://datahack.analyticsvidhya.com/contest/janatahack-machine-learning-in-agriculture/). The data are based on crops harvested by several farmers at the end of the harvest season and adapted for this analysis, in which the response/target variable (harvest damage - crop damage) was converted to two values (0 - healthy, 1- sick, in the original dataset it was 0, 1 and 2, healthy, pesticide damage and damage from other causes). The initial database has 148,168 with 88,858 non-null observations for the target variable. The features considered in the dataset are Estimated Insects Count - Continuous; Crop Type - Continuous; Soil type - Continuous; Pesticide Use Category - Continuous); Number of doses per week (Number Doses Week - Continuous); Number of weeks (Number Weeks Used - Discrete); Number Weeks Quit (Continuous); Period (Season - Continuous).

Data pre-processing techniques for model building

Data quality is one of the main concerns in Machine Learning, as it uses algorithms to extract knowledge while mining the data. Among the techniques used for this, Supervised Learning is done by dividing data into training and testing. During the data-training process, we created the model. The data test the model performance is evaluated. It is noteworthy that the division of data is done randomly. In this phase, the model is built from the data that are presented to the algorithm. For this, some techniques are considered, including Cross-validation, which will train and validate a model with the same dataset, dividing it into partitions (Figure 1). We also utilized other techniques in this work such as Features selection. In this case, the essential attributes that will be applied to train the selected model; in turn, in the Normalization of a data training, it subtracts the mean and divides it by the standard deviation of the data (Z-score technique), as it is indicated to standardize the database since it can contain variables with different scales in the data-training.

Figure 1 - Representation of cross-validation. Source: The Authors.



To carry out the performance evaluation we use the confusion matrix. Through the confusion matrix, it is possible to obtain performance metrics to assess how assertive the model is in the classifications (GÉRON, 2019). Hyperparameter optimization was performed using HyperOpt (FERNADES et al, 2021). In the model's validation, predictions are generated from the test data that is presented to the model. Such predictions are compared with the intended results to evaluate the model's performance with the areas under the curve (AUC), recall, Sensitivity, F1-score and other metrics (KHAN et al., 2019).

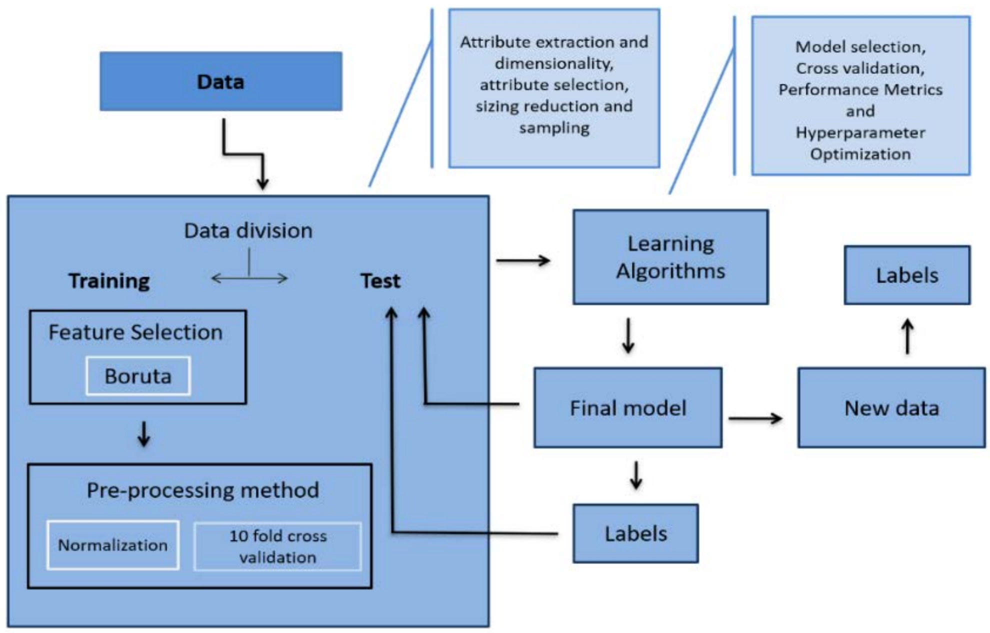
We use the *feature importance* to verify the model's behaviour in a scenario with variable restrictions. To perform the forecast we applied the Shapley Additive Explanation (SHAP) (BROWNLEE, J., 2020). SHAP values are a recent technique that allows quantitative estimation of model interpretability (Lundberg and Lee, 2017, Lundberg et al., 2018). To explain complex models, SHAP uses a simple additive function to interpret the model:

$$f(a) = g(a') = \emptyset_0 + \sum_{j=1}^{J} \emptyset_j a'_j,$$

Wherein $f(a)$ is the original machine learning model, $g(a')$ is the simplest explanation model, J is the number of simplified input features, $\emptyset j$ is the SHAP values measured in all possible input orders $a'j$ is the simplified input vector that indicates whether a particular attribute is present or not during the estimation. It is associated with the model prediction when all attributes are not considered in the estimation (LUNDBERG and LEE, 2017; LUNDBERG et al., 2018, MOLNAR, 2020). The BORUTA algorithm, whose construction is based on the Random Forest classifier, was used (KURSA et al., 2010).

We applied Machine Learning (ML) algorithms based on decision trees and boosting (boosting trees). They have an advantage in terms of fast training time for average datasets in comparison to other algorithms. This type of algorithm does not require considerable time to adjust the hyperparameters (hyperparameters tuning). The tested algorithms were: Adaboost Classifier (YING et al., 2013); Gradient Boosting (ZHANG and HAGHANI, 2015); XGboost is the abbreviation for eXtreme Gradient Boosting package (CHEN and GUESTRIN, 2015); LightGBM (KE et al., 2017) and CatBoost Classifier (PROKHORENKOVA et al, 2017). Figure 2 displays the steps followed to obtain the results. The analysis was performed using scikit-learn library from Python (PEDREGOSA et al., 2011).

Figure 2 - Machine Learning analysis workflow. Source: The Authors



RESULTS AND DISCUSSION

Table 1 displays the sample distribution for the crop damage outcome. The proportion of crop damage, whether by pesticides or other causes, is 16%, being an unbalanced outcome far from proportional equity in the response variable. Table 1 shows the descriptive statistics of the harvest variables. The database is formed by 4 quantitative variables. The number of weeks used and the number of weeks to leave do not present the mean, median and standard deviation values, as they do not present practical results. The outcome variable Harvest damage has two values 0 (healthy) and 1 (sick), with a mean of 0.16 and a standard deviation of 0.3707.

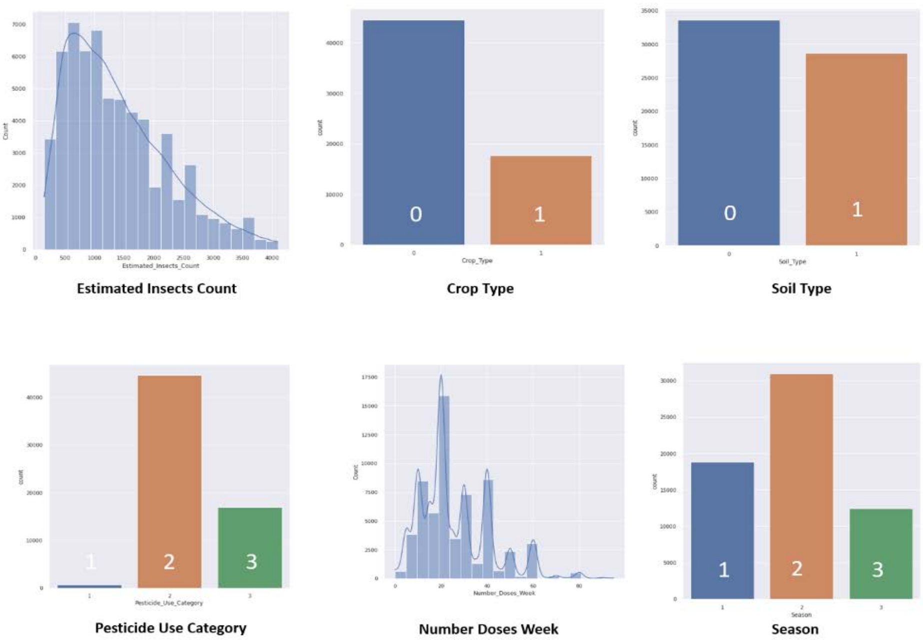
Table 1 - Sample distribution of outcome variables

Features	Min	Max	Avg. prop.	Median	Var.	Std.
Estimated Insects Count	150	4097	1398.21	1212	721,136.26	849,197.42
Crop Type	0	1	0.28	0	0.2039	0.45
Soil type	0	1	0.45	0	0.2481	0.49
Pesticide Use Category	1	3	2.26	2	0.2132	0.46
Number of doses per week	0	95	25.85	20	241.0382	15.52
Number of weeks	0	67	28.65	28	153.8508	12.40
Number Weeks Quit	0	50	9.56	7	97.6645	9.88
Period	1	3	1.89	2	0.4921	0.70
Crop Damage	0	1	0.16	0	0.1374	0.37

Min. = Minimum; **Max**. = Maximum; **Avg**. **prop**. =Average Proportion; **Var**. = Variance; **Std**.= Standard Error

Figure 3 exhibits the complementation of descriptive statistics in graphical form for better visualization.

Figure 3 - Descriptive graphics panel.



For the Models' evaluation with or without the variable selection algorithm called BORUTA, some metrics were used such as Accuracy, Sensitivity, F1 score, Accuracy-Sensitivity curve, Calibration curve, Density curve and area under the curve ROC.

Table 2 displays the fit with the model with all variables considered for prediction. The results of five different Machine Learning algorithms are presented in the test set. The CatBoost

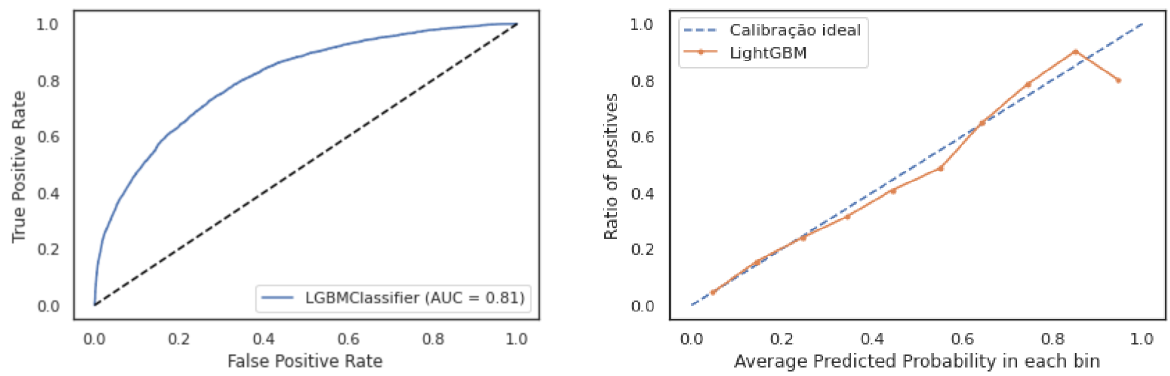
Classifier and Light Gradient Boosting Machine models presented similar results, with differences only in the third decimal places. A suitable difference was in *Recall* (sensitivity) that indicated LightGBM has a higher rate of accuracy in predicting crop damage.

Table 2 - Metrics of the models evaluated in the test set (data-training).

Model	Accuracy	AUC	IC$_{AUC}$ [95%]	Recall	Espec.	Prec.	F1
Light Gradient Boosting Machine	0.85	0.81	[0.80;0.82]	0.25	0.97	0.66	0.36
CatBoost Classifier	0.85	0.81	[0.80;0.82]	0.24	0.97	0.66	0.36
Extreme Gradient Boosting	0.80	0.80	[0.80;0.82]	0.59	0.84	0.43	0.50
Gradient Boosting Classifier	0.83	0.77	[0.76;0.78]	0.31	0.93	0.48	0.37
Ada Boost Classifier	0.84	0.79	[0.79;0.80]	0.18	0.97	0.62	0.28

The ROC (Receiver Operator Characteristic) curve is calculated for different probability thresholds with the joint variation of sensitivity and specificity. Figure 4a shows the area under the curve (AUC) of the ROC curve of 0.81 for the LightGBM model.

Figure 4 - ROC curve and the calibration curve for the LightGBM model.



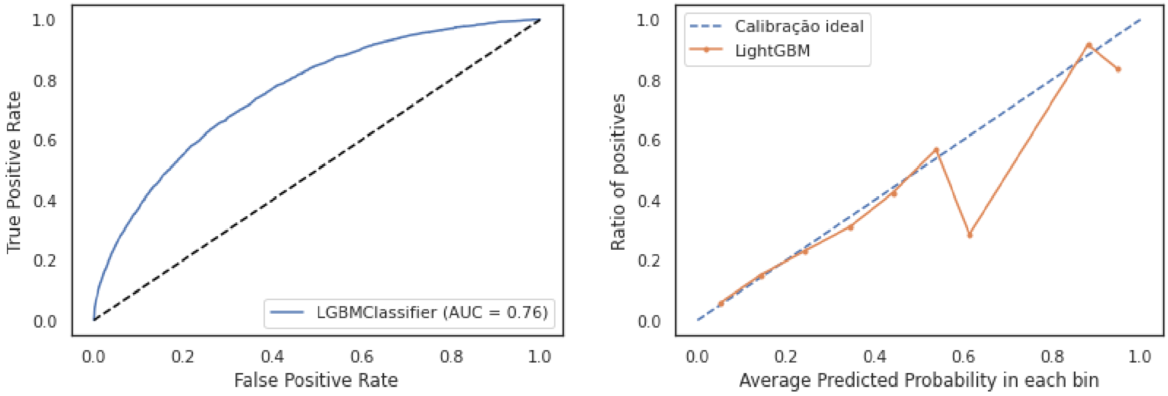
(a) (b)

Machine learning algorithms often predict probabilities of uncalibrated classes. The calibration chart evaluates the calibration of a model, and the methods proportionate better calibrate predictions for a problem. The calibration graph presents a linear blue line that involves the relative frequency of what was observed (y-axis) versus the predicted probability frequency (x-axis). In this type of graph, the closer the adjusted model is to the blue dotted line, the better the model will be calibrated. The calibration curve for the LightGBM model (Figure

4b) gives a remarkable result for the model being well-calibrated in all predicted probability ranges.

The LightGBM algorithm using the Boruta feature selection has the soil type, insect estimate and crop type variables as the most significant variables of the model. There was a 7% reduction in the area under the curve (AUC) value (Figure 5a).

Figure 5 - ROC curve and the model calibration curve under LightGBM Boruta.

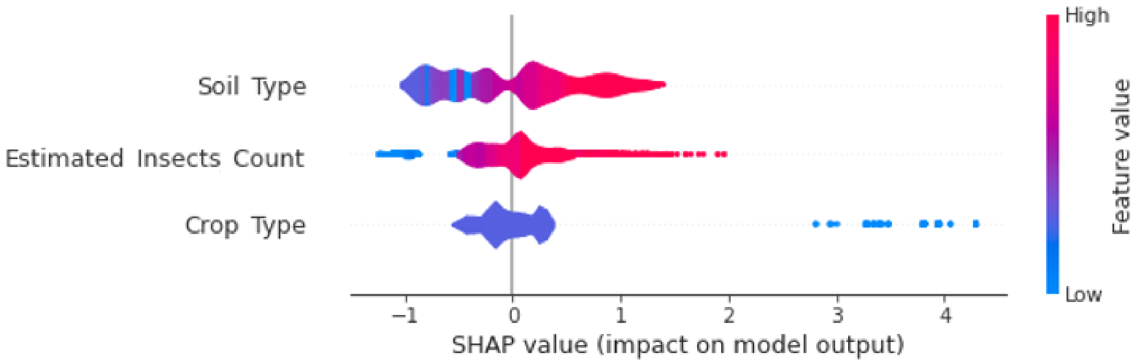


(a) (b)

From Figure 5b, the calibration curve under the Boruta model worsened concerning the fit compared to the model without the Boruta selection feature. The worsening occurred in the intermediate ranges of predicted probability. Despite this worsening, the advantage of the Boruta model is that it is easy to implement at the field level, as it requires only three variables to perform the prediction of crop damage. The SHAP graph for the LightGBM model under the BORUTA algorithm presents the variables' importance configuration. The most influential variable is the soil type and the high value of the soil type category has a strong influence on the classification of damages in the crop. The type of harvest is the variable that has less predictive importance. The collection of points in the figure represents individual data points. Each feature (or predictor) in the column is sorted in descending order of importance, so the SHAP values (on the x-axis) get progressively smaller values in the column, going from highest relevance to soil type to crop type.

To interpret Figure 6 it is necessary to evaluate the estimated number of insects category. The SHAP values correspond to the range of low to positive estimated insect quantity values for the different observations. It is substantial to notice that the low SHAP values of the category of the estimated number of insects have a strong influence on the damage rating in the crop. They are linked to the low (blue coloured) value of the feature. This fact indicates that low estimated insect quantity category values are a crucial feature in predicting crop damage.

Figure 6 - Shap Values for the model under Boruta LightGBM.



Regarding the Boruta model, the percentage of success of the algorithm concerning the values with the highest probability, 20% percentage was 53%. It characterizes that, despite the reduction in AUC, the algorithm has a certain probability in the percentage of success of the algorithm in the 20% most likely values.

In this work, we used through a *Kaggle* database the structure of tabular data in which crop damages are classified as 0 for healthy and 1 for sick. Thus, it differs from most works found in the review study of Liakos et al. (2018), bringing the use of Machine Learning for structured data on disease detection for agriculture in contrast to Deep Learning (DL) techniques.

Sujatha et al. (2021) conducted a study to compare the performance of Machine Learning and Deep Learning techniques in plant disease detection. Yang and Guo (2017) presented Machine Learning in the analysis of resistance genes in plants; and the classification of plant diseases. Ramesh et al. (2018) utilized Random Forest to identify healthy and sick leaves from the datasets created.

Shuriti and Rhavaghent (2019) performed a comparative study on five types of machine learning classification techniques for plant disease recognition in several databases and disease types, Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor, Fuzzy C-Means Classifier and Convolutional Neural Network (CNN).

In the machine learning field, several feature selection methods have been used. Sharma and Misha (2021) compared the performance of 10 classifiers such as *Naive Bayes, logistic regression, k-nearest neighbour (K Nearest Neighbour), vector machine support, decision tree, random forest, artificial neural network, Adaboost, XGBoost and Gradient boosting*. The experimental study applied to a group of variables through three feature selection approaches, such as correlation-based feature selection (CFS), selection based on information gain

(Information Gain attribute evaluation - IG), and the selection of sequential feature selection (SFS) and the combination of voters (voting classifier) for prediction of breast cancer. The authors found that CFS was the technique that showed the best results in conjunction with the combination of classifiers (voting classifier). In this work the utilization of the Boruta method (feature selection) to predict the diseased crop to reduce the number of variables did not present a significant loss in the performance of the algorithm. Lubo-Robles et al. (2020) used the SHAP technique to evaluate the importance of features for seismic data. Their work increased the understanding of how each aspect contributes to the process of distinguishing between Mass Transport Deposits (MTDs) and Salt Seismic facies in the Gulf of Mexico. Similarly, we utilized SHAP values to bring information on how each variable influences the prediction and interpretation of each variable and on the forecast and distinction of crop damage.

CONCLUSION

Several agricultural types of research use ML techniques to detect, identify and predict plant diseases and crop damage. Currently, ML techniques have used the concept of decision trees, and Random Forest and classifiers based on boosting (Catboost, XGBoost, Extra trees and Light GBM) applied to detect diseases in plants, helping the farmer in the automatic detection of types of crop diseases and crop damage. The LightGBM algorithm showed better performance for predicting crop damage with an AUC of 0.81 and 0.76, without using BORUTA, and using BORUTA, respectively. Feature Selection and Feature Importance techniques such as the Boruta and Shapley algorithms, respectively, are still little explored in agricultural research and are of great importance for better applicability of the algorithms.

REFERENCES

Brownlee, J. (2020). *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). DOI: 10.1145/2939672.2939785

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, *145*, 311-318. DOI: 10.1016/j.compag.2018.01.009

Fernandes, F. T., de Oliveira, T. A., Teixeira, C. E., de Moraes Batista, A. F., Dalla Costa, G., & Chiavegatto Filho, A. D. P. (2021). A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil. *Scientific reports*, *11*(1), 1-7. DOI: 10.1038 / s41598-021-82885-y

Géron, A. (2019). *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. Alta Books.

Janatahack: Machine Learning in Agriculture. Analytics Vidhya.2020. Available in: https://datahack.analyticsvidhya.com/contest/janatahack-machine-learning-in-agriculture/

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30, 3146-3154.

Khan, S. A., & Rana, Z. A. (2019, February). Evaluating performance of software defect prediction models using area under precision-Recall curve (AUC-PR). In *2019 2nd International Conference on Advancements in Computational Sciences (ICACS)* (pp. 1-6). IEEE.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *J Stat Softw*, *36*(11), 1-13.

Lubo-Robles, D., Devegowda, D., Jayaram, V., Bedle, H., Marfurt, K. J., & Pranter, M. J. (2020, October). Machine Learning model interpretability using SHAP values: application to a seismic facies classification task. In *SEG International Exposition and Annual Meeting*. OnePetro. DOI: 10.1190/segam2020-3428275.1

Lundberg, S. M., & Lee, S. I. (2017). Consistent feature attribution for tree ensembles. *arXiv preprint arXiv:1706.06060*.

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Moshou, D., Bravo, C., Oberti, R., West, J., Bodria, L., McCartney, A., & Ramon, H. (2005). Plant disease detection based on data fusion of hyper-spectral and multi-spectral fluorescence imaging using Kohonen maps. *Real-Time Imaging*, *11*(2), 75-83. DOI: 10.1016/j.rti.2005.03.003

Moshou, D., Bravo, C., Wahlen, S., West, J., McCartney, A., De Baerdemaeker, J., & Ramon, H. (2006). Simultaneous identification of plant stresses and diseases in arable crops using proximal optical sensing and self-organising maps. *Precision Agriculture*, *7*(3), 149-164. DOI: 10.1007/s11119-006-9002-0

Moshou, D., Bravo, C., West, J., Wahlen, S., McCartney, A., & Ramon, H. (2004). Automatic detection of 'yellow rust'in wheat using reflectance measurements and neural networks. *Computers and electronics in agriculture*, *44*(3), 173-188. DOI: 10.1016/j.compag.2004.04.003

Moshou, D., Pantazi, X. E., Kateris, D., & Gravalos, I. (2014). Water stress detection based on optical multisensor fusion with a least squares support vector machine classifier. *Biosystems Engineering*, *117*, 15-22. DOI: 10.1016/j.biosystemseng.2013.07.008

Pantazi, X. E., Moshou, D., Oberti, R., West, J., Mouazen, A. M., & Bochtis, D. (2017). Detection of biotic and abiotic stresses in crops by using hierarchical self organizing classifiers. *Precision Agriculture*, *18*(3), 383-393. DOI: 10.1007/s11119-017-9507-8

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). CatBoost: unbiased boosting with categorical features. *arXiv preprint arXiv:1706.09516*.
Ramesh, S., Hebbar, R., Niveditha, M., Pooja, R., Shashank, N., & Vinod, P. V. (2018, April). Plant disease detection using machine learning. In *2018 International conference on design innovations for 3Cs compute communicate control (ICDI3C)* (pp. 41-45). IEEE.

Sharma, A., & Mishra, P. K. (2021). Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. *International Journal of Information Technology*, 1-12. DOI: 10.1007/s41870-021-00671-5

Shruthi, U., Nagaveni, V., & Raghavendra, B. K. (2019, March). A review on machine learning classification techniques for plant disease detection. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)* (pp. 281-284). IEEE. DOI: 10.1109 / ICACCS.2019.8728415

Sujatha, R., Chatterjee, J. M., Jhanjhi, N. Z., & Brohi, S. N. (2021). Performance of deep learning vs machine learning in plant leaf disease detection. *Microprocessors and Microsystems*, *80*, 103615. DOI: 10.1016/j.micpro.2020.103615

Yang, X., & Guo, T. (2017). Machine learning in plant disease research. *March*, *31*, 1.
Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013). Advance and prospects of AdaBoost algorithm. Acta Automatica Sinica, 39(6), 745-758. DOI: 10.1016/S1874-1029(13)60052-X

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, *58*, 308-324. DOI: 10.1016/j.trc.2015.02.019