Article

Machine Learning Methods to Estimate Productivity of Harvesters: Mechanized Timber Harvesting in Brazil

Rafaele Almeida Munis, Rodrigo Oliveira Almeida, Diego Aparecido Camargo, Richardson Barbosa Gomes da Silva, Jaime Wojciechowski and Danilo Simões

Abstract: The correct capture of forest operations information carried out in forest plantations can help in the management of mechanized harvesting timber. Proper management must be able to dimension resources and tools necessary for the fulfillment of operations and helping in strategic, tactical, and operational planning. In order to facilitate the decision making of forest managers, this work aimed to analyze the performance of machine learning algorithms in estimating the productivity of timber harvesters. As predictors of productivity, we used the availability of hours of machine use, individual mean volumes of trees, and terrain slopes. The dataset was composed of 144,973 records, carried out over a period of 28 months. We tested the predictive performance of 24 machine learning algorithms in default mode. In addition, we tested the performance of blending and stacking joint learning methods. We evaluated the model’s fit using the root mean squared error, mean absolute error, mean absolute percentage error, and determination coefficient. After cleaning the initial database, we used only 1.12% to build the model. Learning by blending ensemble stood out with a determination coefficient of 0.71 and a mean absolute percentage error of 15%. From the use of data from machine learning algorithms, it became possible to predict the productivity of timber harvesters. Testing a variety of machine learning algorithms with different dynamics contributed to the machine learning technique that helped us reach our goal: maximizing the model’s performance by conducting experimentation.

Keywords: individual mean volumes of trees; blending ensemble learning; decision making; terrain slope; forest plantation; stacking ensemble learning

1. Introduction

Management integrates the routine of forest managers responsible for guiding and implementing mechanized logging operations. The optimization of time and biological assets capitalized in planted forests, when exhausted by timber harvesters, affects the success of the operation. Thus, it is necessary to know the variables that influence mechanized timber harvesting, allowing for more effective planning.

Quality indicators, evaluation criteria, and risk analysis techniques enhance the structures that support decision makers. In doing so, data collected in forest inventory, measurements at the stand level, operational forest management, and onboard computers in forest machinery help and allow the management procedures for timber harvesting operations [1–6].

The manipulation and reuse of this information promote its use in the management development itself and helps in the identification of opportunities that can be foreseen. However, a mechanized timber harvesting operation planning scope necessarily requires a quantitative, robust, and reliable quantitative database [7,8].
On a daily basis, a large number of data are generated from the activities machines record that make up different harvesting systems. In Brazil, the mechanized cut-to-length (CTL) system, composed of a harvester and a forwarder, is normally used to supply raw material for forest-based industries. Due to the use of a harvester in the configuration of the CTL system, the trees are cut and sectioned into processed logs. Additionally, the timber without harvest residues, suitable for industrial processes, is arranged for extraction by means of a forwarder [9–16].

The sources and estimates reported by operators give the level of dataset quality. By providing early warnings, machine learning provides qualitative subsidies in detection of associations between variables, in addition to supporting hypotheses. However, machine learning algorithms must be associated with data wrangling methods to highlight outliers and reduce the size and facilitate dataset filtering [17–21].

In addition to data wrangling, the data preparation adapts to different analyses through the identification, extraction, cleansing, transformation, and integration of datasets. Data wrangling allows the user to explore data flow and problem distribution along the dataset to detect error patterns. This technique is conducive to the development of quality in the sequence of operations over time [22–27].

Forest managers, through datasets, are able to extract useful information and make better-informed decisions that include, for example, estimating the productivity of timber harvesters. Development of a model to predict dynamic productivity behavior of machines, under different configurations, can contribute to optimization of the operation [28–31].

Associating the knowledge generated from the application of data wrangling in a dataset with machine learning algorithms can promote intelligent decision making. The union of techniques of machine learning can improve the system performance of cut-to-length timber, forest resources, and services management [32–35].

Some studies investigated and proposed ways to improve this association, e.g., joint machine learning methods use a meta-model to combine predictive results from heterogeneous base models arranged in at least one layer. Thus, the variation in forecast combinations is based on a data validation set [36–39].

Learning methods through blending ensemble learning and stacking ensemble learning can promote more robust and accurate models because they decrease exponentially the error rate of joint learning. The blending ensemble learning method allows recognizing and training second-level algorithms to optimally combine prediction models, while stacking ensemble learning combines values by average or by additional models that add to the partial or final predictions of all individual models [40–46].

Machine learning models are increasingly used to predict attributes because, as they are not subject to traditional statistical assumptions, they can incorporate variables and find nonlinear and complex patterns in data. Machine learning includes a variety of algorithms for learning predictive rules from historical data and building models that can predict future values [47–51].

Machine learning has become an effective predictive analytics tool for large data volumes, with applications in medicine, finance, and, recently, the forestry sector. Forestry process automation has been proven to increase productivity and work quality. However, the mechanized timber harvesting operation is challenging and complex [52–55].

Traditionally, productivity is estimated through a time study, which consists of analyzing the work elements that make up the timber harvester’s operational cycle, individually, associating the time required and timber volume. However, the requirements of labor and minimum amount of sample needed make it expensive for forest-based industries.

As a result of these complex interactions, the use of efficient analytical tools by forest managers, such as machine learning, is justified to assist them in decision making. This work aimed to analyze the performance of machine learning algorithms in estimating the productivity of timber harvesters.
2. Materials and Methods

2.1. Dataset

We used structured data from the production and operation of mechanized timber harvesting in Eucalyptus- and Pinus-planted forests carried out by cut-to-length systems with harvesters. The planted forests with Eucalyptus had a spacing of $3.3 \times 1.8$ m and mean age of $14 \pm 9.87$ years. The Pinus forest had a spacing of $3.3 \times 1.8$ m and mean age of $22 \pm 9.09$ years. The wood from these forests was used as raw material for the production of pulp and paper.

The average meteorological conditions in the study region, according to the National Institute of Meteorology [56], were a relative humidity of 69.24%, wind speed of $4.29 \text{ m s}^{-1}$, and an air temperature of 289.3 K. The operations took place in Brazil, in a region with a slope gradient from 7.32% to 35.06%. The intervals were categorized by a gentle (3% to 10%), moderate (10% to 32%), and steep slope relief (32% to 56%), according to Speight [57].

This research was based on empirical data and silvicultural inputs; therefore, the data were part of the daily records collected in the field by the onboard computers of timber harvesters. Despite considering all records in the initial analysis, we employed a series of compensatory controls from data wrangling.

In the two 10 h shifts daily, the machine availability, the individual mean volumes of trees, and the terrain slope added up to 144,973 instances incurred in the period of 28 months. These data were categorical, numerical, and ordinal, according to the box plot and distributions of predictor and target variables provided in the Supplementary Material (Figure S1). The bases were labeled, joined, and manipulated using the R programming language [58].

The actual times spent in activities were recorded in the onboard computers of timber harvesters. This way, we estimated productivity from the ratio between the timber volume extracted by a harvester, in cubic meters, and the effective operation time, in seconds [59,60]. The same operator could operate different brands and models of timber harvesters; however, this variable was not added in the construction of the model, due to the difficulty in tracking these data in the database. Altogether, the operating records of 21 harvesters were used (Table 1).

### Table 1. Technical specifications of timber harvesters used in cutting and processing trees.

| Timber Harvesters | Quantities of Timber Harvesters |
|-------------------|-------------------------------|
| **Brand**         | **Model**         | **Power (kW)** | **17** |
| Harvester Ponsse  | ERGO 8W           | 205            |       |
| Harvester John Deere | 1270E         | 170            | 2     |
| Harvester Ponsse  | BEAVER           | 150            | 1     |
| Harvester John Deere | 1270G          | 200            | 1     |

Through the programming language R, when implementing machine learning routines for management planning and detecting data quality, we considered, according to Konstantinou and Paton [61], procedures for transforming, cleaning, and merging different sources. We built a data wrangling routine, in which the outliers and potentially correlated variables were removed.

The instances went through the data wrangling process, which was performed in order to properly transform and gather acquired data. Additionally, through the interquartile range, conceptually defined by the Tukey range [62], we removed outliers and, using Spearman correlation, we verified the correlations between attributes ($p < 0.05$).

Furthermore, data balancing from SMOTE was performed. The SMOTE was adopted because it is a reference algorithm to solve the class disequilibrium learning problem [63]. The SMOTE algorithm has the dynamics of generating new synthetic examples in the neighborhood of small groups of nearby instances, using the $k$-nearest neighbor [64]. The function was implemented from the smotefamily package.
2.2. Different Learning Methods and Algorithm Approaches

Using a single dataset, we compared the predictive performance of 24 machine learning algorithms to estimate the productivity of timber harvesters. These algorithms were based on a decision tree, gradient boosting machine, linear regression, k-nearest neighbors, support vector machine, and artificial neural network.

For determining the best model, we used the metrics: root mean error (RMSE), mean absolute error (MAE), and mean absolute percent (MAPE). We used the determination coefficient ($R^2$) as a final performance measure for each method.

We adopted the gradient (5, 10, 15, 20, and 25) for cross-validation, in which the hyperparameters were automatically optimized. Finally, we implemented stacking ensemble and blending ensemble learning methods. We ordered stacking ensemble learning methods in a hierarchical data structure. On each fold set, we applied k-fold cross-validation.

The predictive performance of models in relation to unseen data was maximized by determination coefficient ($R^2$), minimizing test RMSE, which was determined from the random sample mean generated ($n = 80$). We implemented supervised learning regression using the Python programming language PyCaret library [65] to automate machine learning workflow and model development. From data instances, we grouped 90% ($n = 1466$) as a training set and 10% ($n = 163$) as a test set. We tested machine learning algorithms and selected them according to their performance in predicting productivity of forest machines, using universal statistical metrics for evaluating the performance of models [66], such as root mean squared error (RMSE), mean absolute error (MAE), mean absolute percent error (MAPE), and determination coefficient ($R^2$).

Thus, we subjected the algorithms to different machine learning methods. First, we verified the performance of algorithms in a decoupled version, with hyperparameters in default mode. To improve performance, we adjusted the hyperparameters of the selected algorithms. We combined the validated data and formed the meta-feature set, the test data, and the target set. Again, we combined the sets using new meta-resource sets, creating a new meta-training set. The new target sets formed a new meta-test. We generated final predictions by meta-learner level one, from training with the meta-training set.

The combination learning method consisted of combining machine learning algorithms to minimize prediction error rates. For this, we divided the dataset into training and testing, as well as implementing zero-layer algorithms, which generated validation and test sets. We combined respective sets with new meta-training and meta-test sets [67] and generated final predictions by level-one meta-learner, from training with a meta-training set.

3. Results
3.1. Dataset Quality

The manipulation of the dataset with daily records of the mechanized timber harvesting operation resulted in a sample of 144,973 instances. However, because it was consolidated from the unity of different sources, including manual notes, the goodness of the dataset was partially compromised. Thus, we removed duplicate instances and instances with missing information.

With a data wrangling routine, in addition to cleaning filtering and transforming data, we carried out an examination of data quality, excluding outliers and promoting balancing. It is noteworthy that, despite timber harvesters having onboard computers, the data recording process still required manual interactions. Consequently, we implemented this process in 1.12% of the dataset, the models with machine learning algorithms (Table 2).
Table 2. Data wrangling procedures applied to the dataset from the mechanized timber harvesting operation.

| Data Wrangling Procedures     | Input  | Output |
|-------------------------------|--------|--------|
| Duplicate instances          | 144,973| 144,781|
| Removal of missing data      | 144,781| 98,459 |
| Annotation errors            | 98,459 | 1703   |
| Outlier removal               | 1703   | 1629   |

The attributes selected for the model building were individual mean volumes of trees, terrain slope, and availability of hours of machine use. More details about the mean, standard deviation, and median of the dataset from mechanized timber harvesting operation for attributes under study are shown in Table 3.

Table 3. Mean, standard deviation, and median of dataset from mechanized timber harvesting operation, after process of removing outliers of three initial attributes.

| Attribute                          | Minimum | Maximum | Mean  | Standard Deviation | Median | Skewness | Kurtosis |
|-----------------------------------|---------|---------|-------|--------------------|--------|----------|----------|
| Individual Mean Volumes of Trees [m$^3$] | 0.11    | 0.62    | 0.35  | 0.09               | 0.36   | 0.09     | 2.83     |
| Terrain Slope [%]                 | 2.86    | 35.06   | 12.37 | 6.46               | 11.40  | 1.62     | 6.26     |
| Availability of Hours of Machine use [h] | 0.15    | 1.33    | 0.69  | 0.12               | 0.70   | −0.55    | 5.19     |

3.2. Different Learning Methods and Algorithm Approaches

First, we analyzed the predictive performance of 24 algorithms, individually, based on model fit metrics. Of the three trained algorithms, based on the decision tree, the determination coefficient of extra trees stood out, as it was 0.01 higher than the coefficient of determination of random forest and 0.22 higher than that of the decision tree (Table 4).

Table 4. Evaluation metric of models based on decision tree applied to the training set from mechanized timber harvesting operation.

| Model                        | MAE [m$^3$ h$^{-1}$] | RMSE [m$^3$ h$^{-1}$] | R$^2$ | MAPE [%] | Time [s] |
|------------------------------|----------------------|------------------------|-------|----------|----------|
| Extra Trees Regressor        | 3.21                 | 4.68                   | 0.72  | 17       | 0.12     |
| Random Forest Regressor      | 3.36                 | 4.72                   | 0.71  | 18       | 0.13     |
| Decision Tree Regressor      | 4.26                 | 6.22                   | 0.50  | 22       | 0.01     |

MAE: mean absolute error; RMSE: root mean squared error; R$^2$: determination coefficient; MAPE: mean absolute percent error.

When analyzing the four algorithms based on gradient-boosted machines, the best determination coefficient was obtained by CatBoost Regressor, which was 0.04 higher than Gradient Boosting Regressor and 0.20 higher than AdaBoost Regressor (Table 5).

Table 5. Evaluation metric of models based on gradient-boosted machine applied to the training set from mechanized timber harvesting operation.

| Model                        | MAE [m$^3$ h$^{-1}$] | RMSE [m$^3$ h$^{-1}$] | R$^2$ | MAPE [%] | Time [s] |
|------------------------------|----------------------|------------------------|-------|----------|----------|
| CatBoost Regressor           | 3.35                 | 4.55                   | 0.73  | 18       | 1.40     |
| Gradient Boosting Regressor  | 3.68                 | 4.91                   | 0.69  | 20       | 0.04     |
| AdaBoost Regressor           | 4.82                 | 6.04                   | 0.53  | 28       | 0.03     |

MAE: mean absolute error; RMSE: root mean squared error; R$^2$: determination coefficient; MAPE: mean absolute percent error.

The algorithm’s availability based on linear regression contributed to the application of the twelve trained algorithms (Table 6).
Table 6. Evaluation metric of models based on linear regression applied to the training set from mechanized timber harvesting operation.

| Model                          | MAE [m³ h⁻¹] | RMSE [m³ h⁻¹] | R²  | MAPE [%] | Time [s] |
|-------------------------------|--------------|---------------|-----|----------|---------|
| Automatic Relevance Determination | 4.62         | 6.02          | 0.53 | 25       | 0.03    |
| Kernel Ridge                  | 4.65         | 6.04          | 0.53 | 25       | 0.02    |
| Linear Regression             | 4.62         | 6.03          | 0.53 | 25       | 0.30    |
| Huber Regressor               | 4.58         | 6.05          | 0.53 | 25       | 0.02    |
| Ridge Regression              | 4.64         | 6.04          | 0.53 | 25       | 0.20    |
| Bayesian Ridge                | 4.62         | 6.03          | 0.53 | 25       | 0.02    |
| TheilSen Regressor            | 4.72         | 6.19          | 0.51 | 26       | 0.51    |
| Least Angle Regression        | 4.87         | 6.38          | 0.47 | 26       | 0.01    |
| Orthogonal Matching Pursuit   | 5.49         | 6.87          | 0.40 | 29       | 0.01    |
| Lasso Regression              | 6.68         | 8.35          | 0.11 | 40       | 0.28    |
| Elastic Net                   | 6.68         | 8.34          | 0.11 | 40       | 0.02    |
| Lasso Least Angle Regression  | 7.32         | 8.86          | 0.00 | 44       | 0.02    |

MAE: mean absolute error; RMSE: root mean squared error; R²: determination coefficient; MAPE: mean absolute percent error.

It was found that the Automatic Relevance Determination, Kernel Ridge, Linear Regression, Huber Regression, Ridge Regression, and Bayesian Ridge algorithms showed the same determination coefficient, which was 0.02 higher than that of TheilSen Regressor, 0.06 higher than Least Angle Regression, 0.13 higher than Orthogonal Matching Pursuit, and 0.42 higher than Lasso Regression and Elastic Net.

Despite having different dynamics, the best determination coefficient was obtained by the k-neighbors regressor, which was 0.05 higher than the multi-layer perceptron regressor, 0.18 higher than the Random Sample Consensus, and 0.45 higher than Support Vector Regression (Table 7).

Table 7. Evaluation metric of models based on k-nearest neighbor, multi-layer perception regressor, random sample consensus, support vector regression, dummy regressor, and passive-aggressive regressor applied to the training set from mechanized timber harvesting operation.

| Model                          | MAE [m³ h⁻¹] | RMSE [m³ h⁻¹] | R²  | MAPE [%] | Time [s] |
|-------------------------------|--------------|---------------|-----|----------|---------|
| K-Neighbors Regressor          | 4.09         | 5.59          | 0.60 | 24       | 0.02    |
| Multilayer Perceptron Regressor| 4.57         | 5.94          | 0.55 | 25       | 0.40    |
| Random Sample Consensus        | 5.10         | 6.73          | 0.42 | 27       | 0.02    |
| Support Vector Regression      | 6.35         | 8.18          | 0.15 | 37       | 0.03    |
| Dummy Regressor                | 7.32         | 8.86          | 0.00 | 44       | 0.01    |
| Passive-Aggressive Regressor   | 8.25         | 10.02         | -0.49| 52       | 0.02    |

MAE: mean absolute error; RMSE: root mean squared error; R²: determination coefficient; MAPE: mean absolute percent error.

Among applied models that presented better determination coefficients were blending ensemble and stacking ensemble. Next, the algorithms were carried out in default mode, highlighting the Extra Trees Regressor (Table 8).
Table 8. Performance obtained on the training set, by default, stacking ensemble learning and blending ensemble learning model.

| Model                              | Process       | MAE  [m³ h⁻¹] | RMSE [m³ h⁻¹] | R²   | MAPE [%] |
|------------------------------------|---------------|---------------|---------------|------|----------|
| Extra Trees Regressor (A)          | Default       | 3.31          | 4.49          | 0.74 | 18       |
| CatBoost Regressor (B)             | Default       | 3.13          | 4.58          | 0.73 | 17       |
| K-Neighbors Regressor (C)          | Default       | 4.06          | 5.54          | 0.61 | 24       |
| A, B e C                          | Blending ensemble | 3.18        | 4.43          | 0.75 | 18       |
| A, B e C–A                        | Stacking ensemble | 3.21        | 4.43          | 0.75 | 18       |

MAE: mean absolute error; RMSE: root mean squared error; R²: determination coefficient; MAPE: mean absolute percent error.

When analyzing the metrics in the dataset test, the blending ensemble model was confirmed as the best predictor of productivity of timber harvesters (Table 9).

Table 9. Results obtained on the test dataset, by default, stacking ensemble learning and blending ensemble learning model.

| Model                              | Process           | R²      | MAPE [%] |
|------------------------------------|-------------------|---------|----------|
| A, B e C                          | Blending ensemble | 0.77    | 15       |
| A, B e C–A                        | Stacking ensemble | 0.76    | 14       |
| Extra Trees Regressor              | Default           | 0.74    | 14       |
| CatBoost Regressor                 | Default           | 0.74    | 15       |
| K-Neighbors Regressor              | Default           | 0.67    | 19       |

R²: determination coefficient; MAPE: mean absolute percent error.

In addition, as an assessment of overall model performance, we verified the 80 combinations of test set data, with the response. Thus, it was evident that the blending ensemble, followed by the stacking ensemble, produced relatively higher average values of R² (Figure 1) and a lower degree of dispersion.

Figure 2 illustrates the performance of the main algorithms used in model construction to predict productivity, relating to observed values.

Figure 1. Average determination coefficient values of employed models.
When we selected the black box algorithm, the increase in performance compromised the interpretation of the relationships between the predictor variables and the target variable. Complex mathematical functions made it difficult to infer from technical experts. However, by visualizing the distributions of predictor variables in each quartile of the response variable it was possible to infer that higher productivity was associated with greater machine availability and lower slope levels.

Although the algorithms used in the construction of models do not allow interpretability, as evidenced by productivity quartiles of the test set, the density distribution of predictor variables of individual mean volumes of trees, terrain slope, and machine availability was determined (Figure 3).

Figure 2. Comparison of scatter plot values of employed models.
4. Discussion

Incorporating machine learning models into forest operations management routines allows managers to infer tactical and operational adjustments, with agility in decision making and accurate prognosis. In mechanized timber harvesting, the scenario dynamics and external influences that impact activities demand this adaptability together with forecasting capacities.

However, conducting and monitoring the performance of mechanized timber harvesting operations, using analytical tools such as machine learning, is restricted due to the quantity and quality of available data. Liski, et al. [31] and Maktoubian et al. [68], Demirci et al. [69] and Abbasi et al. [70] report that decentralization and lack of data management in forest environments reduce the achievement of significant results.

The harvesters that acted as data sources had embedded technology, with output records of activities of timber cutting and sectioning. The lack of interoperability among electronic devices made communication and data transfer susceptible, compromising the possibility of instant corrections and perceptions of deviations in notes. Furthermore, Buccafurri et al. [71] and Shi et al. [72] point out that the quality of instances generated is part of a cooperative process, which requires participation, and therefore, leveling of all

Figure 3. Density distribution of employed models.

First quartile of productivity

Second quartile of productivity

Third quartile of productivity

Fourth quartile of productivity
those involved. Data management must be aligned with operations’ organization, which makes it the responsibility of forest managers to coordinate these efforts.

The data residual volume, after execution of data wrangling processes, was still sufficient to verify the performance of machine learning algorithms in the productivity modeling of mechanized timber harvesting. Of the algorithm groups applied in the modeling, in the default process, the ones that performed best were those based on the decision tree, gradient-boosted machine, and k-nearest neighbor. Therefore, the best individual performance algorithms were, respectively: extract trees, gradient-boosted, and k-nearest neighbor.

There are many types of decision trees that have as their core the entropy of information. According to An and Zhou [73], in the specific analysis process, the gained information for each attribute is classified and ordered. Among the decision tree algorithms evaluated, the one that presented the best performance was the extremely randomized trees or extra trees algorithm. This algorithm was developed by Geurts et al., [74] and uses the same principle of random forests. However, as supported by Ahmad et al. [75], the extra trees may have differentiated themselves by using the entire training dataset to train each regression tree and not just a bootstrap replica.

Of the algorithms based on gradient-boosted machines, the CatBoost Regressor showed the best fit model to the data. This algorithm developed by Prokhorenkova et al. [76] is an enhancement of gradient boosting, designed to avoid attribute dependency and improve prediction accuracy on small datasets. As it is a non-parametric algorithm that, according to Ortiz-Bejar et al. [77], stores all known observations and uses them in the prediction based on similarity functions, the third-best performance was from the model based on k-nearest neighbor.

As a way of enhancing the prediction, tests were carried out with the blending ensemble and stacking ensemble learning methods, using combined learning. These learnings were combined from the three algorithms, in default mode, which presented the best performances. The predictions obtained by both methods were superior to those obtained by algorithms in default mode. Jong et al. [78] and Jordan and Mitchell [79] point out that, in general, combined learning methods increase the performance of models built with machine learning.

Associating the blending ensemble use with the possibility of pre-determining productivity, based on attributes of individual mean volumes of trees, terrain slope, and availability of hours of machine use, promotes dynamism in managers’ planning, especially in operational planning, which requires quick responses in adverse operating conditions. This corroborates the limitations of traditional estimating method productivity through the study of times.

In addition, the comparison through values of employed models’ scatter diagrams demonstrated the effects of predictor variables on productivity. In upper quartiles, in operating conditions with lower slopes and longer availability harvesters, their effects increase considerably the productivity.

The building of models involving machine learning algorithms, in addition to providing prediction of harvester productivity in the mechanized timber harvesting operation, allowed us to look at the bases that guide strategic decisions of operations in planted forests. This opportunity has shown that, despite the quality, suitable data promote knowledge extraction, mainly from attributes not correlated with productivity.

5. Conclusions

From the use of adjusted data of machine learning algorithms, it is possible to predict the productivity of timber harvesters.

Among the attributes that compose datasets of mechanized timber harvesting activities, the individual mean volumes of trees, terrain slope, and machine availability are the main factors that impact harvester productivity estimation.

Testing a variety of machine learning algorithms with different dynamics contributed to the development of a machine learning technique that enabled what it proposes, i.e.,
experimentation and good performance of the models. Thus, the choice for blending ensemble learning was guided by the comparison of model fit statistical metrics. Among the learning methods by blending ensemble, stacking ensemble, and algorithms, in default mode, the blending ensemble had a determination coefficient of 0.71 and a mean absolute percent error of 15%.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/f13071068/s1. Figure S1: Raincloud plot with box plot and the distributions of predictor variables and target (A) individual mean volumes of trees; (B) terrain slope; (C) machine availability; and (D) productivity.

**Author Contributions:** Conceptualization, R.A.M., D.S. and J.W.; methodology, R.A.M., D.S., R.O.A. and D.A.C.; software, R.A.M. and R.O.A.; validation, R.A.M., R.O.A., D.A.C., R.B.G.d.S., J.W. and D.S.; formal analysis, R.A.M., R.O.A. and D.A.C.; investigation, R.A.M., D.A.C. and R.B.G.d.S.; data curation, R.A.M. and R.O.A.; writing—original draft preparation, R.A.M., D.A.C. and R.B.G.d.S.; writing—review and editing, R.A.M., R.O.A., D.A.C., R.B.G.d.S., J.W. and D.S.; visualization, supervision, D.S. and J.W.; project administration, D.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All data are provided in the main manuscript. Contact the corresponding author if further explanation is required.

**Acknowledgments:** Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Pollard, S.T.; Brookes, A.; Earl, N.; Lowe, J.; Kearney, T.; Nathanail, C.P. Integrating decision tools for the sustainable management of land contamination. *Sci. Total Environ.* 2004, 325, 15–28. [CrossRef]
2. Bai, C.; Sarkis, J. Integrating and extending data and decision tools for sustainable third-party reverse logistics provider selection. *Comput. Oper. Res.* 2019, 110, 188–207. [CrossRef]
3. Welch, H.; Brodie, S.; Jacob, M.G.; Bograd, S.J.; Hazen, E.L. Decision-support tools for dynamic management. *Conserv. Biol.* 2020, 34, 589–599. [CrossRef]
4. McIntire, R.E.; Tomppo, E.O. Remote sensing support for national forest inventories. *Remote Sens. Environ.* 2007, 110, 412–419. [CrossRef]
5. Liang, X.; Kankare, V.; Hyypää, J.; Wang, Y.; Kukko, A.; Haggrén, H.; Yu, X.; Kaartinen, H.; Jaakkola, A.; Guan, F.; et al. Terrestrial laser scanning in forest inventories. *ISPRS J. Photogramm. Remote Sens.* 2016, 115, 63–77. [CrossRef]
6. Eyvindson, K.; Saad, R.; Eriksson, L.O. Incorporating stand level risk management options into forest decision support systems. *For. Syst.* 2017, 26, e013. [CrossRef]
7. Wagner, J.E. Misinterpreting the internal rate of return in sustainable forest management planning and economic analysis. *J. Sustain. For.* 2012, 31, 239–266. [CrossRef]
8. Karttunen, K.; Laitila, J. Forest management regime options for integrated small-diameter wood harvesting and supply chain from young Scots pine (*Pinus sylvestris* L.) stands. *Int. J. For. Eng.* 2015, 26, 124–138. [CrossRef]
9. Camargo, D.A.; Munis, R.A.; Simões, D. Investigation of exposure to occupational noise among forestry machine operators: A case study in Brazil. *Forests* 2021, 12, 299. [CrossRef]
10. Fernandez-Lacruz, R.; Edlund, M.; Bergström, D.; Lindroos, O. Productivity and profitability of harvesting overgrown roadside verges—A Swedish case study. *Int. J. For. Eng.* 2021, 32, 19–28. [CrossRef]
11. Visser, R.; Berkett, H. Effect of terrain steepness on machine slope when harvesting. *Int. J. For. Eng.* 2015, 26, 1–9. [CrossRef]
12. Sherwin, L.M.; Owende, P.M.O.; Kanali, C.L.; Lyons, J.; Ward, S.M. Influence of tyre inflation pressure on whole-body vibrations transmitted to the operator in a cut-to-length timber harvester. *Appl. Ergon.* 2004, 35, 253–261. [CrossRef]
13. Ovaskainen, H.; Heikkilä, M. Visuospatial cognitive abilities in cut-to-length single-grip timber harvester work. *Int. J. Ind. Ergon.* 2007, 37, 771–780. [CrossRef]
14. Walsh, D.; Strandgard, M. ScienceDirect Productivity and cost of harvesting a stemwood biomass product from integrated cut-to-length harvest operations in Australian Pinus radiata plantations. *Biomass Bioenergy* 2014, 738, 93–102. [CrossRef]
15. Hera, P.L.; Morales, D.O.; Mendoza-Trejo, O. A study case of Dynamic Motion Primitives as a motion planning method to automate the work of forestry engine. Comput. Electron. Agric. 2021, 183, 106037. [CrossRef]

16. Huang, X. Application analysis of AI reasoning engine in microblog culture industry. Pers. Ubiquitous Comput. 2020, 24, 393–403. [CrossRef]

17. Cho, G.; Park, H.-M.; Jung, W.-M.; Cha, W.-S.; Lee, D.; Chae, Y. Identification of candidate medicinal herbs for skincare via data mining of the classic Donguibogam text on Korean medicine. Integr. Med. Res. 2020, 9, 100436. [CrossRef] [PubMed]

18. Rodrigues de Holanda Maia, M.; Plastino, A.; Penna, P.H.V. MineReduce: An approach based on data mining for problem size reduction. Comput. Oper. Res. 2020, 122, 104995. [CrossRef]

19. Xu, Z.; Cheng, X.; Wang, K.; Yang, S. Analysis of the environmental trend of network finance and its influence on traditional commercial banks. J. Comput. Appl. Math. 2020, 379, 112907. [CrossRef]

20. da Silva, A.K.V.; Borges, M.V.V.; Batista, T.S.; da Junior, C.A.S.; Furuya, D.E.G.; Osco, L.P.; Teodoró, L.P.R.; Baio, F.H.R.; Ramos, A.P.M.; Gonçalves, W.N.; et al. Predicting eucalyptus diameter at breast height and total height with uav-based spectral indices and machine learning. Forests 2021, 12, 680. [CrossRef]

21. Giannetti, F.; Pecchi, M.; Travaglini, D.; Francini, S.; D’Amico, G.; Vangi, E.; Cocozza, C.; Chirici, G. Estimating via windstorm damaged forest area in Italy using world view-2 imagery and continuous change detection algorithms. Forests 2021, 12, 680. [CrossRef]

22. Kandel, S.; Heer, J.; Plaisant, C.; Kennedy, J.; Van Ham, F.; Brodbeck, D.; Buono, P. Research directions in data wrangling: Visualizations and transformations for usable and credible data. Inf. Vis. 2011, 10, 271–288. [CrossRef]

23. Endel, F.; Piringer, H. Data Wrangling: Making data usable again. IFAC-PapersOnLine 2015, 28, 111–112. [CrossRef]

24. Furche, T.; Gottlob, G.; Libkin, L.; Orsi, G.; Paton, N.W. Data wrangling for big data: Challenges and opportunities. Adv. Database Technol.-EDBT 2016, 2016, 473–478. [CrossRef]

25. Bellomarini, L.; Fayzrakhmanov, R.R.; Gottlob, G.; Kravchenko, A.; Reissfelder, S.; Sallinger, E.; Sherkhonov, E.; Vahdati, S.; et al. Data science with Vadalas: Knowledge Graphs with machine learning and reasoning in practice. Futur. Gener. Comput. Syst. 2022, 129, 407–422. [CrossRef]

26. Bors, C.; Gschwandtner, T.; Miksch, S. Capturing and visualizing provenance from data wrangling. IEEE Comput. Graph. Appl. 2019, 39, 61–75. [CrossRef]

27. Yang, J.; Li, R.; Chen, L.; Hu, Y.; Dou, Z. Research on equipment corrosion diagnosis method and prediction model driven by data. Process Saf. Environ. Prot. 2022, 158, 418–431. [CrossRef]

28. De Jaegher, B.; Larumbe, E.; De Schepper, W.; Verliefde, A.; Nopens, I. Colloidal fouling in electrodialysis: A neural differential equations model. Sep. Purif. Technol. 2020, 249, 116939. [CrossRef]

29. Kudýba, S. A hybrid analytic approach for understanding patient demand for mental health services. Netw. Model. Anal. Health Inform. Bioinforma. 2018, 7, 3. [CrossRef]

30. Liu, J.; Kadziriski, M.; Liao, X.; Mao, X.; Wang, Y. A preference learning framework for multiple criteria sorting with diverse additive value models and valued assignment examples. Eur. J. Oper. Res. 2020, 286, 963–985. [CrossRef]

31. Liski, E.; Jounela, P.; Korpunen, H.; Sosa, A.; Lindroos, O.; Jylhä, P. Modeling the productivity of mechanized CTL harvesting with statistical machine learning methods. Int. J. For. Eng. 2020, 31, 253–262. [CrossRef]

32. Sunhare, P.; Chowdhary, R.R.; Chattopadhyay, M.K. Internet of things and data mining: An application oriented survey. J. King. Faisal Univ.-Comput. Inf. Sci. 2020, 34, 3590–3591. [CrossRef]

33. Luo, M.; Wang, Y.; Xie, Y.; Zhou, L.; Qiao, J.; Qiu, S.; Sun, Y. Combination of feature selection and catboost for prediction: The first application to the estimation of aboveground biomass. Forests 2021, 12, 216. [CrossRef]

34. Kusiak, A. Data mining: Manufacturing and service applications. Int. J. Prod. Res. 2006, 44, 4175–4191. [CrossRef]

35. Maxwell, A.E.; Warner, T.A.; Strager, M.P.; Conley, J.F.; Sharp, A.L. Assessing machine-learning algorithms and image- and lidar-derived variables for GEOBIA classification of mining and mine reclamation. Int. J. Remote Sens. 2015, 36, 954–978. [CrossRef]

36. Taylor, P.; Griffiths, N.; Bhalerao, A.; Anand, S.; Popham, T.; Xu, Z.; Gelencser, A. Data Mining for Vehicle Telemetry. Appl. Artif. Intell. 2016, 30, 233–256. [CrossRef]

37. Chatzimparmpas, A.; Martins, R.M.; Kucher, K.; Kerren, A. Empirical Study: Visual Analytics for Comparing Stacking to Blending Ensemble Learning. In Proceedings of the 2021 23rd International Conference on Control Systems and Computer Science (CSCS), Bucharest, Romania, 26–28 May 2021; pp. 1–8. [CrossRef]

38. Hansraja, A.; Adeliyi, T.T.; Wing, J. Detection of Online Fake News Using Blending Ensemble Learning. Sci. Program. 2021, 2021, 3434458. [CrossRef]

39. Wu, T.; Zhang, W.; Jiao, X.; Guo, W.; Alhaj Hamoud, Y. Evaluation of stacking and blending ensemble learning methods for estimating daily reference evapotranspiration. Comput. Electron. Agric. 2021, 184, 106039. [CrossRef]

40. Divina, F.; Gilson, A.; Gomez-Vela, F.; Torres, M.G.; Torres, J.F. Stacking ensemble learning for short-term electricity consumption forecasting. Energies 2018, 11, 949. [CrossRef]
73. An, Y.; Zhou, H. Short term effect evaluation model of rural energy construction revitalization based on ID3 decision tree algorithm. *Energy Rep.* **2022**, *8*, 1004–1012. [CrossRef]

74. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]

75. Ahmad, M.W.; Reynolds, J.; Rezgui, Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *J. Clean. Prod.* **2018**, *203*, 810–821. [CrossRef]

76. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. Catboost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *2018*, 6638–6648.

77. Ortiz-Bejar, J.; Graff, M.; Tellez, E.S.; Ortiz-Bejar, J.; Jacobo, J.C. \( \kappa \)-Nearest neighbor regressors optimized by using random search. In *Proceedings of the 2018 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, Ixtapa, Mexico, 14–16 November 2018. [CrossRef]

78. Jong, N.; Krumelich, J.S.M.; Versteegen, D.M.L. To what extent can PBL principles be applied in blended learning: Lessons learned from health master programs. *Med. Teach.* **2017**, *39*, 203–211. [CrossRef]

79. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef]