A Big Data approach to forestry harvesting productivity

Rossit, Daniel Alejandro[1,2]; Olivera, Alejandro[3]; Viana Céspedes, Víctor *[3]; Broz, Diego[4]

[1] Department of Engineering, Universidad Nacional del Sur (UNS), Av. Alem 1253, Bahía Blanca (B8000CPB), Argentina.

[2] INMABB, Universidad Nacional del Sur (UNS)-CONICET, Av. Alem 1253, Bahía Blanca (B8000CPB), Argentina

[3] Universidad de la República Ruta 5, km 386,5, Tacuarembó (C.P. 45000), Uruguay.

[4] UNaM CONICET Facultad de Ciencias Forestales Bertoni 124, Eldorado (N3382GDD), Misiones, Argentina.

Abstract

Modern industrial technology enables to collect and process large amount of data, providing valuable information for different industry activities. A representative case of this evolution is the Forest industry, since modern forest harvesters are equipped with automatic data collection devices. The collected data can be extracted and communicated to computers using special forestry protocols, as StanForD, where it can be analysed. This skill of modern harvesters allows to study harvest productivity with thousands of records, instead of having a few hundred as it would be possible by recording through traditional methods (visual inspection or filming). However, traditional analytical methods, as linear regression, are not capable to deal with this volume of data (or, at least, does not take full advantage of the data potential), consequently, new approaches must be considered. Our proposal is to address this shortcoming using data mining methods, specially, we consider decision trees and *k*-means algorithms. We study how different variables (DBH, species, shift and operator) affect the productivity of a forest harvester considering real scenario data. The harvest data comes from Eucalyptus spp. plantations in Uruguay where the harvest system implemented is cut-to-length. To analyse the data, firstly, productivity is modelled in a categorical manner considering two different approaches: ranges of equal intervals and ranges calculated using *k*-means clustering algorithm. Then, Decision Trees methods are applied to analyse the influence of the mentioned variables in productivity. The results show that clustering is a proper approach to categorically model scalar productivity and that DBH is the most influential factor in productivity. Moreover, Decision Trees, after setting DBH values, allowed to use new variables to describe productivity, achieving very high levels of accuracy, in many cases greater than 90%.

1. Introduction

Forest industry plays an important role in the economic and social development of South American countries. In Uruguay for instance, this sector has grown considerably in the last few decades. The planted area with pine and eucalyptus used to be less than 100,000 hectares in 1990, and in 2012 the planted area rose to 990,030 hectares (Boscana & Boragno, 2017). The management of forest operations in industrial plantations is an extremely complex decision-making process involving ecological, productive, and economic factors, among others (Broz et al. 2017a). This complexity is increased if the production cycle is considered (i.e. several years from plantation to harvest), since the impact of bad decisions can be exceptionally costly and harmfully for the ecosystem. Thus, having proper and sound support for the decision-making process becomes crucial.

The harvesting operations and transport of round wood are the most expensive activities of the forest value chain (Broz et al., 2016). Accordingly, the planning of these activities requires proficient decision support tools, where the input data for the decision-making process has a critical role (Broz et al., 2018). Modern harvesters and forwarders are often coupled with automatic data recording equipment, which can record several variables during normal operations. When harvester and forwarder work together, they do it in a Cut-to-Length (CTL) configuration or system (Figure 1) (Uusitalo, 2010). CTL system was developed in Scandinavian countries, and has been widely adopted in South America, where the mechanization of forest operations in countries like Uruguay is over 60% (Olivera [2016]; Viana Céspedes [2018]). The harvester fells, processes (delimbs and debarks) and cut the trees into logs in the stand, then, the forwarder extracts the logs to the roadside. The capacity of harvesters and forwarders of data recording is framed under a “de facto” standard called StanForD, which uses a series of specific files (Skogforsk, 2017). This standard enables to collect data from various machines and manufacturing control systems with a twofold purpose, to generate useful information for efficient forest management, and to follow up operations (Skogforsk, 2007). Typically, the data includes records of a working shift and from each single processed tree, these records sum, approximately, 500,000 single datum for machine and for operator in a single shift on an average day. Such sizable data set constitutes a typical big data problem; hence, to transform this data into meaningful information it is necessary to use an adequate approach, as Big Data (BD).

BD is a concept that refers to such large data sets that traditional data processing applications are not enough to deal with them, and to the procedures used to find repetitive patterns within that data. Big data, in early 2000s was characterized by the 3V’s: Volume, Variety and Velocity (Laney 2001).

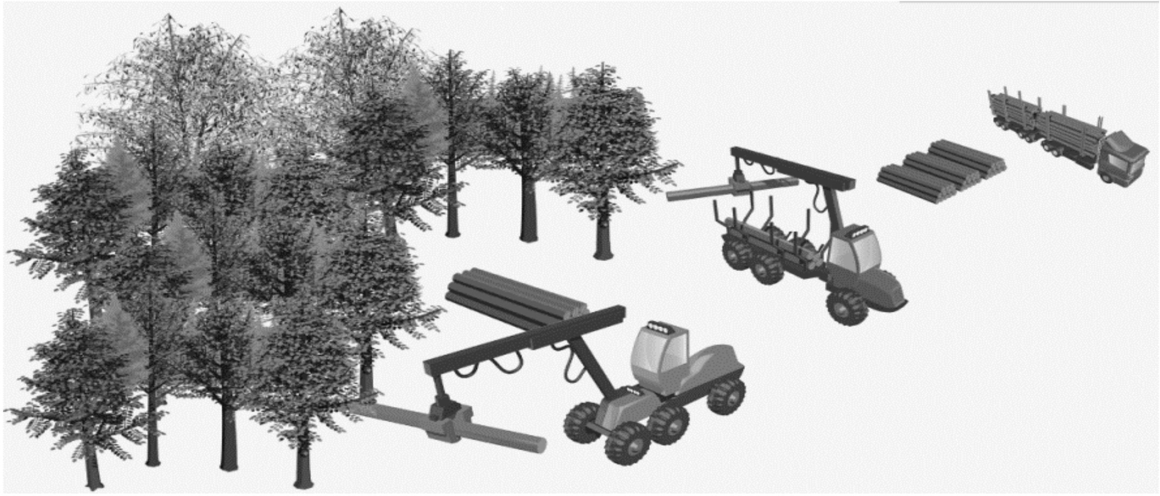

Figure 1. Schematic representation of the Cut-to-length harvesting system showing a harvester a forwarder. Source: http://forestenergy.org

- Volume (V1): the amount of data generated and saved. The size of the data determines the value and potential understanding, and if you can consider them as true big data.
- Velocity (V2): in this context, the speed at which data is generated and processed to meet the demands and challenges of its analysis.
- Variety (V3): the type and nature of the data, this helps people who analyse the data and use the results effectively. Type or nature of data is due to the different sources (e.g. images, videos, remote and field-based sensing data), or to different recording formats, among other aspects.

More recently, other characteristics have been identified on BD datasets, as Veracity (V4) and Valorization (V5) (Chi et al. [2016]; Kamilaris et al. [2017]).

- Veracity (V4): the quality and reliability of the data captured can vary greatly and thus, affect the results of the analysis.
- Valorization (V5): The ability to propagate knowledge and appreciation.

As it is mentioned, our case requires a BD approach since the Volume and Variety of our datasets correspond to BD category. The harvest system considered for this work, generates about 500 thousand records for a single normal operation shift indicating that the Volume is a feature of our system. The other BD feature of the data considered in this work is the Variety, our data type includes: time registers (processing time stamps), metrics registers (trees diameters), categorical registers (tree species, shift class and operator ID) and volumetric registers (harvested wood).

Data mining (DM) techniques arise as a solution to the problem of analysing large amount of data. DM is the process of applying Computer Based Information Systems (CBIS) for discovering knowledge from data (Vlahos et al. 2004). Thus, DM allows discovering patterns and trends, which would be useful to predict the behaviour of a system, and interesting interactions (Ahlemeyer-Stubbe & Coleman [2014]; Traub et al. [2017]). DM has been successfully applied to a wide variety of fields, as industrial processes manufacturing, marketing, sociology, for instance (Liao et al. 2012); however, limited research has been published on its application to forestry.

The most meaningful of this few publications are reviewed here. Mohammadi et al. (2011) compared linear regression and regression tree analyses for forest attribute estimations and their spatial modelling. The results showed that, statistical models of stand volume, tree density, species richness and reciprocal of Simpson Indices using tree regression analysis had higher adjusted compared to linear regression models. Using DM techniques, Özbayoğlu & Bozer (2012) estimated the risk on forest fire and some of the methods analysed were multilayer perceptron, radial basis function networks, support vector machines and fuzzy logic. They used historical forest fire records, which contained parameters like geographical conditions of the existing environment, date and time when the fire broke out, meteorological data such as temperature, humidity and wind speed, and the type and tree stocking. Sanquetta et al. (2013) employed a DM methodology named Instance-based Classification for estimating carbon storage in *Araucaria angustifolia* (Bertol.) Kuntze plantation in Brazil. This technique outperformed the conventional methods. Lokers et al. (2016) examined and analysed three European projects as guidance to describe current possibilities and future challenges for deployment of BD techniques in the field of agro-environmental research, facilitating decision support at the level of societal challenges. They recommended the use of BD approaches to analyse data from various sources, e.g. harvesting, production, and meteorological records. Likewise, Li et al. (2017) analysed the relevant economic, social and ecological factors of China's forestry resources with a BD perspective. Firstly, they used the method of data envelopment analysis to investigate the forestry resources efficiency; then they analysed time series data using the Malmquist total factor productivity index method to determine which factors restrain forest resources efficiency for the studied region in China. Applying Neural network-based models Hlásny et al. (2017) presented a large-scale evaluation of climate effects on the productivity of three temperate tree species in Central Europe. Using this technique, they determined which among 13 tested climate variables best predicted the tree species-specific site index.

Regarding harvest activities, Olivera et al. (2015) proved the usefulness of integrating Global Navigation Satellite System (GNSS) with forest harvester data to improve forest management. They retrieved data from a GNSS-enabled harvester working in Cut-to-length (CTL) operations in *Eucalyptus spp*. plantations in Uruguay. The dataset obtained comprised over 63,000 cycles of felled and processed trees. With this data, a mixed effects model was fitted to evaluate harvester productivity as a function of stem diameter at breast height

DBH), species, shift, slope and operator. On another study, Olivera (2016) used the same dataset to develop a method to map forest stand productivity based on the actual volume recovered by the harvesting operation. Eriksson & Lindroos (2014) developed models to predict machine productivity in Sweden using linear regression based on ordinary least squares parameter estimation. This study constitutes the productivity models based on the largest sample of harvester and forwarder follow up dataset in CTL final felling and thinning operations considered to date.

Various DM techniques have been applied in research for the agro-environmental sciences, including forestry. Prediction of forest fires, the effect of climatic variable on forest productivity, forests structure analysis and carbon storage are some of the case studies published. Techniques comprise mixed models, artificial neural network, association rules, and regression trees. However, there is still a gap in the literature regarding the use of DM techniques in forest operations, especially on harvesting operations using the automatic data collection system available on modern forest harvesters, this article presents a contribution to fill this gap. The first objective of this study is to evaluate the effect of the variables, diameter at breast height (DBH) as an indicator of piece (tree) size, species, shift (day/night) and operator, on the productivity of a harvester using the DT technique. The second is to compare the result of this analysis with the multiple regression analysis performed by Olivera et al. 2015. The DT technique could allow evaluating the performance of operators in different strata of pieces size (DBH) and its interaction with the other evaluated variables (including categorical ones, as tree species or operation shift). The results can be used as a planning tool to allocate workers according to their best performance in tree size range, shift or species by simple rules. Preliminary results of this work were published in (Broz et al. 2017b) and (Rossit et al. 2017).

2. Methods and materials

2.1. Data source

The studied data is part of the data set used by Olivera (2016). The data set was obtained from a single grip harvester Ponsse Ergo 8W equipped with a harvester head designed to process and debark eucalypts (Ponsse H7euca). The control system was Opti4G 4.715, StanForD (Standard) enabled. This machine harvested *Eucalyptus ssp* plantations in the western of Uruguay in the period between August and October of 2014. For this study, only records of *Eucalyptus grandis* Hill. ex Maiden and *Eucalyptus dunnii* Maiden were used. The operation developed from Monday to Friday in double shifts: the day shift from 7:00 a.m. to 5:30 p.m. and the night shift from 8:30 p.m. to 7:00 p.m. All harvested production was debarked for pulpwood and cut to a standard log length of 6.5 m and a second class of variable length between 4 m and 6.5 m. All diameter measurements and volume estimated presented in this work are debarked. The harvester was operated by two workers, Operator 1 and Operator 2. Operator 1 was more experienced (several years operating harvesters) than Operator 2 (ten months in the task) on operating forest harvesters. For more details, refer to Olivera (2016).

For the analysis of the system 9,941 records of individual trees registered under the StanForD standard as .stm files were used. The .stm files contain compressed data from each individual tree processed, which include DBH (Diameter at Breast Height), volume, and log classification, among others. From .stm files, the following data was considered for each registered tree: tree ID, DBH, merchantable harvested volume, time stamp (year, month, day, hour, minute and second) when the tree was felled, and operator identification. A change attribute (day / night) was assigned to each tree according to the timestamp. The .stm files also contain input information, such as species and sites (denomination). Using the time stamp records, cycle time was calculated for each tree by determining the difference between two consecutive stem's time stamps. Then, productivity in m^3 h^{-1} (the dependent variable of the study) was calculated by dividing the volume by the cycle time.

2.2. DM tools

We used DM tools to analyse the effect of the variables DBH, shift, species and operator on the productivity of a harvester working in *Eucalyptus* ssp plantations in Uruguay. The tools used are Decision Trees and the clustering *k*-means method. These methods are widely used in data mining applications and have proved to be versatile in addressing a variety of problems in diverse contexts (Wu et al. 2008).

2.2.1. Cluster Analysis: k-means

Cluster methods allow the generation of knowledge in an exploratory mode without using, a priori, labels or categories. This enable to explore a data set to evaluate its characteristics and to group data into clusters with similar characteristics. There are several methodologies and formats of clusters, the most widely used for different applications is the *k*-means algorithm (Jain 2010). Since we could not find in literature a similar implementation, we used the *k*-means algorithm to generate categories of the productivity variable.

K-means is a clustering method, which aims to cluster a set of *n* observations in *k* groups, where each observation belongs to the group whose average value is closest to its own value. Given a set of observations ($x_1, x_2, \dots, x_j, \dots, x_n$), where each observation is a real vector of *d* dimensions, *k*-means constructs partitions of the observations in *k* sets ($k \leq n$) to minimize the sum of the squares within each group $S = \{S_1, S_2, \dots, S_i, \dots, S_k\}$,

$$argmin_S \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2$$

Where μ_i is the average of points from S_i.

2.2.2. Decision Trees

Decision Trees (DT) are logical structures with wide use in decision making, prediction and data mining. From a set of observations, these models partition the space and predict the response through a model that is a constant function in each subset of the partitions or leaves. DT are strictly non-parametric and do not require assumptions of distributions of the input data (Kohavi 1996). An important characteristic of DT is its capability to break down a complex decision-making process into a collection of simpler decisions, thus providing a solution that is usually easier to understand (Safavian & Landgrebe 1991). Thereby, with the advent of Big Data Analytics, DT has become the most popular algorithm in the Big Data field as indicates Wu et al. (2008) in their Big Data algorithms ranking (*k*-means is the second one). DT can handle nonlinear relationships between features and classes, allow missing values and are able to handle numerical and categorical entries in a natural way; this last characteristic is a requirement of our problem. Despite DT are efficient methods, they depend significantly on their implementation; the different implementations consist on different settings in the validation method and in the branching procedure. The validation method evaluates the ability of the model to describe correctly values of the dependent variable analysed. While, the branching procedure guides the growth of the tree, it analyses how the dependent variable relates to the independent variables (Berry & Linoff, 2004).

DT is fed, in a first stage, with a training data set *S* of already classified samples s_i. Each sample s_i is, in effect, a *p*-dimensional vector *x*, where the $x_{p,i}$ component represent the value of the attribute *p* of the sample s_i, and the class where s_i falls. At each node, DT algorithm chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision.

2.3. Data analysis

2.3.1. Categorization of the productivity

DT are sensitive to the description and modelling of the dependent variable, in our case, how are defined and related the productivity categories. We use two techniques to categorize the productivity variable. The first technique was to divide the dataset into four categories of equal interval. This is adopted since there are no references of categorization of it for harvester's productivity assessment in Uruguay, and productivity is a continuous variable. The second technique was to implement the unsupervised learning methodology of clusters *k*-means, this method is implemented using the software Statistical Package for the Social Sciences (SPSS).

2.3.2. Decision trees - Experimental design

In order to assess harvest productivity, we design a series of experiments which go from the most general to the most particular insights (Figure 2). Firstly, we studied the productivity of the harvester for the whole dataset and obtained a general overview of the effect of the different variables on productivity considering all the variables (Test 1 of Figure 2). Then, a deeper analysis is performed by segregating the data by the other studied variables; namely operator, shift and species (Test 2 and 3 of Figure 2). This deepening is guided by the usefulness of the analysis for a forest harvest planner. If significant differences are found, such segregation would help the harvest planner to allocate the personnel according to their best performance in DBH strata and work shifts, in an efficient manner. It would also aid to solve different operative scenarios (which stands to be harvested and time horizons or deadlines). For Test 2 and 3 two different experiments are executed: one considers data from both tree species, and the other, discriminates by specie.

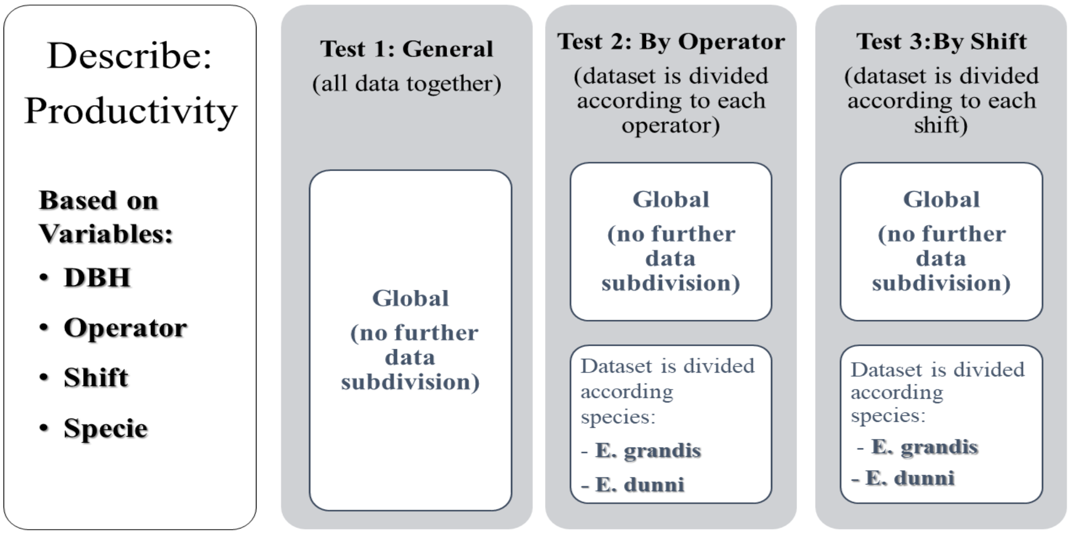
Describe: Productivity
Based on Variables:
• DBH
• Operator
• Shift
• Specie
Test 1: General
(all data together)
Global
(no further data subdivision)
Test 2: By Operator
(dataset is divided according to each operator)
Global
(no further data subdivision)
Dataset is divided according species:
- E. grandis
- E. dunni
Test 3:By Shift
(dataset is divided according to each shift)
Global
(no further data subdivision)
Dataset is divided according species:
- E. grandis
- E. dunni

Figure 2. Schematic representation of the experimental design

2.3.3. Validation

The SPSS software offers three validation options for DT: i) no validation, ii) cross validation and iii) validation by sampling division. The method of no validation implies not making any validation. The cross validation works generating *n* equal partitions, where for each partition *n* generates a prediction model, leaving 1 / n data out of the analysis. Then, it evaluates the prediction model in the 1 / *n* reserved data and elaborates the respective confusion matrix. This is repeated for each set *n*. The final confusion matrix of the model is the average of the *n* confusion matrices obtained. On the other hand, the method that works by sampling division divides the data set into two subsets, one subset of training and the other of evaluation. These subsets are defined based on a percentage of the total data or by choosing variables for each subset. The method works generating a model analysing the training subset, and then, it validates the model in the subset reserved for the evaluation.

The different options of tree growth or branching is another aspect to consider. The branch procedure analyses the predictive capacity of an independent or descriptive variable with respect to the dependent variable. These methods depend strongly on the dependent variable to be studied and on the modelling that is made of it, as well as on the statistical method used. In our case, the variable to be studied (productivity) and its modelling (categorically) are fixed conditions. But the statistical method that supports the branching procedure is not predefined by the problem. The methods offered by SPSS are 4: CHAID, exhaustive CHAID, CRT and QUEST. CHAID method (Chi-squared Automatic Interaction Detection) chooses the independent (predictor) variable that has the strongest interaction with the dependent variable. Categories of each predictor are merged if they are not significantly different with respect to the dependent variable. Exhaustive CHAID method is a modification of CHAID that examines all possible splits for each predictor. CRT method splits the data into segments that are as homogeneous as possible with respect to the dependent variable. QUEST method is fast and avoids other methods' bias in favour of predictors with many categories. We evaluated each of these four methods with their default formulation.

2.4. Categorization of the productivity

For productivity categorization a pre-process was performed in the dataset, basically, values below 6 m^3 h^{-1} were discarded because they were considered outliers. Then, we apply the technique of equal intervals, for this, we divided the range of productivity into four categories of equal range, by intervals of 14 m^3 h^{-1} (Table 1), except for the class with lower values which interval is lower than 14 m^3 h^{-1}. Each category was named like the upper bound of the range.

Table 1. Productivity classes of equal interval.

Category	Range [m^3 h^{-1}]
“<= 12”	[6 – 12]
“<= 26”	(12 – 26]
“<= 40”	(26 – 40]
“> 40”	(40 - …)

Table 2. Productivity classes resulted from clusters analysis for the whole dataset “global”.

Category	Range [m^3/h]
“<= 17”	[6 – 17]
“<= 29.6”	(17 – 29.6]
“<= 43.7”	(29.6 – 43.7]
“> 43.7”	(43.7 - …)

The second technique is based on *k*-means clustering. For this technique, also four categories were generated, to further compare results between both techniques. The *k*-means clusters method allows to generate *k* groups wherein deviations within each group are minimized. Therefore, although the conclusions from the results obtained based on cluster categorizations are more "dependent" on the sample of available data, they allow a deeper analysis of the relationships between the different variables. The results of *k*-means defined the centroids of the clusters (Categories), and the range of values of each cluster (Table 2). Again, the categories of the variable Productivity by clusters are named as the upper bound of the range of the category.

Table 3. Productivity classes resulted from clusters analysis for *E. dunnii*.

Category	Range [m^3/h]
“<= 17.2”	[6 – 17.2]
“<= 29.5”	(17.2 – 29.5]
“<= 43.4”	(29.5 – 43.4]
“> 43.4”	(43.4 - …)

Table 4. Productivity classes resulted from clusters analysis for *E. grandis.*

Category	Range [m^3/h]

“<= 16.9”	[6 – 16.9]
“<= 29.8”	(16.9 – 29.8]
“<= 43.9”	(29.8 – 43.9]
“> 43.9”	(43.9 - …)

Finally, other cluster categories were generated by discriminating the data according to tree species. These categorizations allow evaluating the variations in productivity for each species, *E. dunnii* and *E. grandis* (Tables 3 & 4). For both species, the resulting categories were similar to the global ones generated by clusters.

2.5. Tuning of decision trees

Nextly, the analysis of the different settings configurations for DT is presented. In the implementation of the DT algorithm it was established that the smallest node must contains at least 1% of the data, since obtaining decision rules for sets representing less than 1% is meaningless for the problem addressed. Then, to analyze the setting configuration we run some preliminary DT with different settings options and evaluate which setting perform better for our data set. For this, we used a model similar to the one of Figure 4, where all variables are considered together. The different models generated from the different setting configurations were compared through their global predictive capacity. The variants analysed correspond to different validation options and different tree growth or branching options. The results of these evaluations are in Table 5.

Table 5. Parameter tuning tests.

	Validation method					***Tree growth method***			
	none	***cross***	***sample division***			***CHAID***	***Exhaustive CHAID***	***CRT***	***QUEST***
			50%	60%	65%				
Correct percentage	41.2	**43.2**	43.2	42.4	**43.3**	41.3	**43.3**	42.1	42.6

From Table 5, the validation methods that obtained the best overall prediction values were the cross-validation method and the sampling division method, considering 65% of the data for training. However, the sample validation method showed a strong volatility in the global prediction value, since it is very dependent on how the training and testing subsets are made up. These subsets are generated in a random mode, with the only condition that the number of cases included in each subset respect the percentage indicated. On the other hand, the cross-validation method was much more stable, and it was defined for *n* = 10. Regarding the method of tree growth, the method that obtained the best overall prediction value is the Exhaustive CHAID.

3. Experiments

3.1. Test 1: General

The first experiment describes the behaviour of the productivity variable depending on the main independent variable DBH in mm (Figure 3). Firstly, a DT is presented on Figure 3, where productivity is predicted considering only the DBH variable. In Table 6 the associated confusion matrix is presented.

Table 6: Confusion matrix evaluating the DT of Productivity by DBH presented in Figure 3

Observed	***Forecast***				***Correct %***
	<= 12	***<= 26***	***<=40***	***> 40***	
<= 12	0	1993	36	15	0.00%
<= 26	0	2885	241	173	87.50%
<=40	0	1517	389	297	17.70%
> 40	0	1158	282	455	24.00%
Global %	0.00%	80.00%	10.00%	10.00%	39.50%

Table 6 shows the total prediction level is 39.5%. In addition, for the category "<= 12", the model in Figure 3 fails to predict any case correctly. In Figure 3 there is not node with category "<= 12" highlighted (ref: highlighted categories in each node means the typical category of that node).

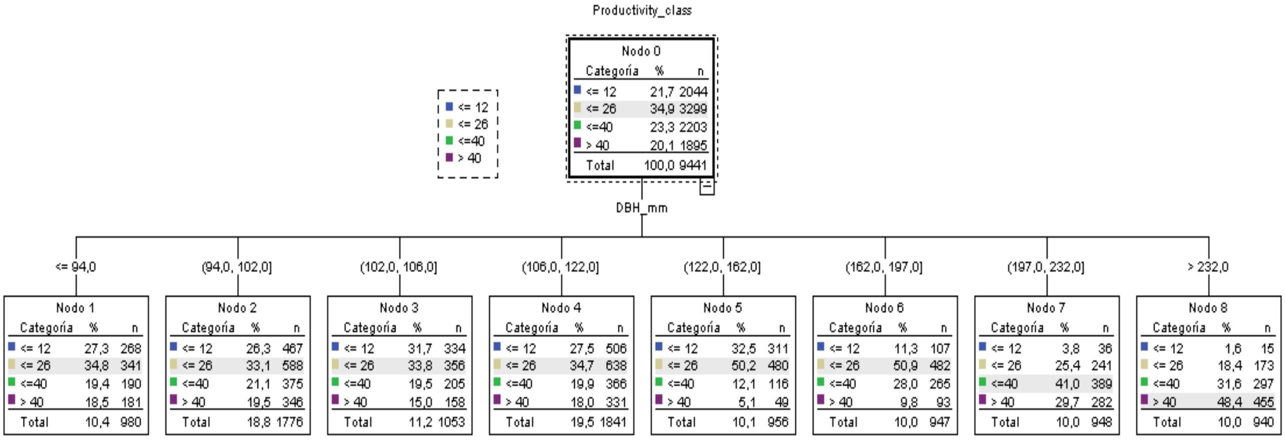

Figure 3. DT of productivity (m^3 h^{-1}) using DBH (mm) as independent variable

The DT in Figure 4 based on the categorization of equal intervals (Table 1) describes the productivity variable through all the independent variables (DBH, Shift, Specie and Operator). It demonstrates that incorporating more variables improves the description of the variable productivity. For example, in node 13 the productivity for DBH between 102 and 106 mm, species *E. dunnii*, is "<= 12" for 91.8% of the cases. Whereas in the model of Figure 3, there is no node dominated by the category "<= 12"; moreover, in Fig. 3 there is no node with higher probability than 51%. Consequently, a contribution of this study is to identify scenarios where productivity can be predicted accurately. These scenarios are described through the values of the independent variables that define the node. In the confusion matrix for this DT (Table 6) it is possible to identify the global value of correct predictions is higher than the one of the model using only DBH (Figure 3), 42.4% Vs 39.5%. In addition, it improves prediction values of categories "<= 12" predicting the 18% of the cases.

A new DT considering the *k*-means based categorization (Table 2) was created (Figure A.1, Appendix) with its corresponding confusion matrix (Table 8). The higher value of the prediction of productivity 44.5% (Table 8) vs 42.4% (Table 7) shows that the implementation of clusters improves the global level of prediction. This improvement can be justified by cluster-based categories tend to be more compact than those of equal intervals.

Table 7. Confusion matrix evaluating the DT of Productivity by DBH, Shift and Operator based on equal intervals classification presented in Figure 4.

Observed	***Forecast***				***Correct %***
	<= 12	***<= 26***	***<=40***	***> 40***	
<= 12	374	1619	36	15	18.3%
<= 26	97	2788	241	173	84.5%
<=40	2	1515	389	297	17.7%
> 40	0	1158	282	455	24.0%
Global %	5.0%	75.0%	10.0%	10.0%	42.4%

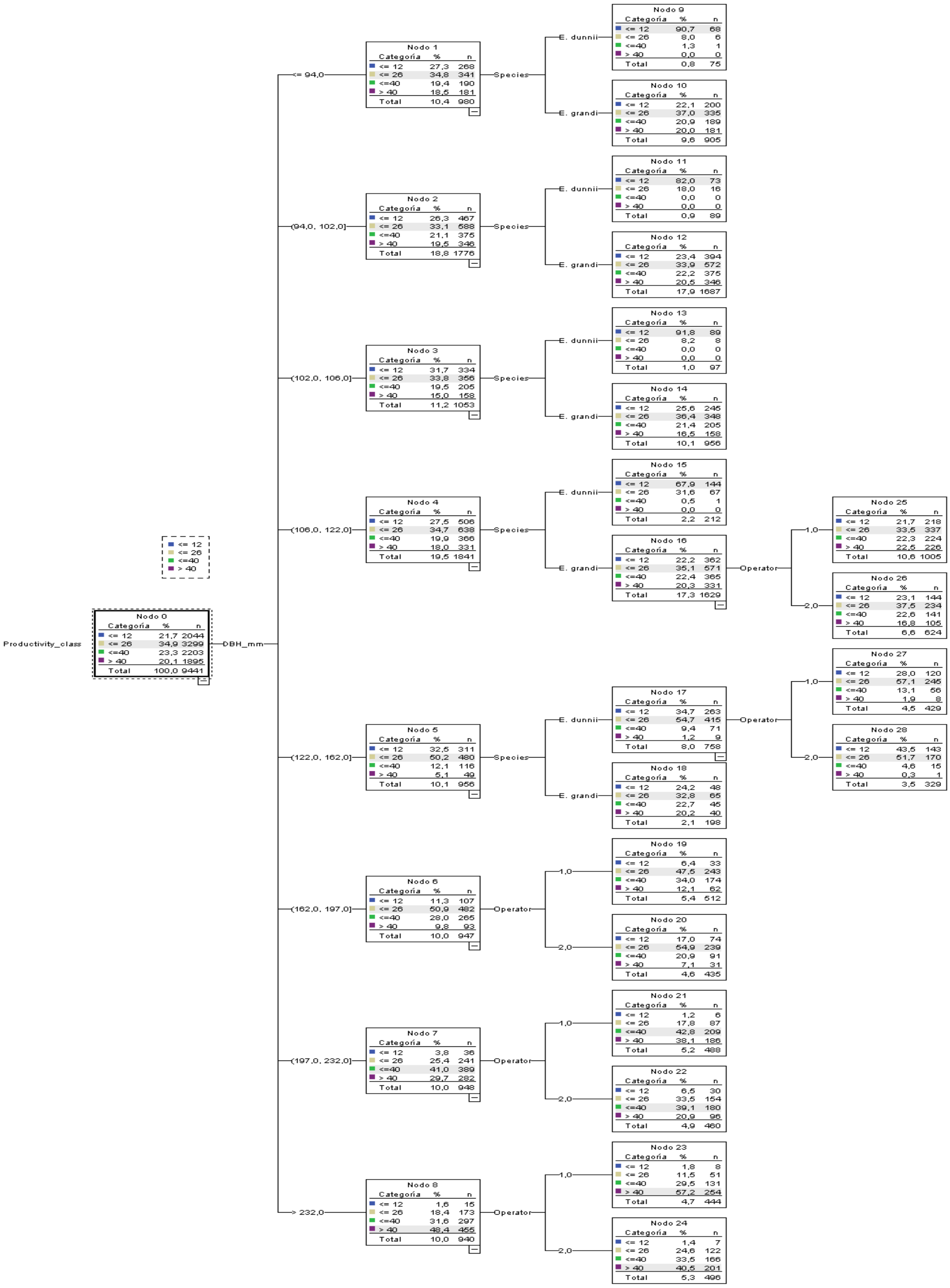
Productivity_class
<= 12
<= 26
<=40
> 40
Nodo 0 | Categoría % n | <= 12 21,7 2044 | <= 26 34,9 3299 | <=40 23,3 2203 | > 40 20,1 1895 | Total 100,0 9441
DBH_mm
<= 94,0
(94,0, 102,0]
(102,0, 106,0]
(106,0, 122,0]
(122,0, 162,0]
(162,0, 197,0]
(197,0, 232,0]
> 232,0
Nodo 1 | <= 12 27,3 268 | <= 26 34,8 341 | <=40 19,4 190 | > 40 18,5 181 | Total 10,4 980
Nodo 2 | <= 12 26,3 467 | <= 26 33,1 588 | <=40 21,1 375 | > 40 19,5 346 | Total 18,8 1776
Nodo 3 | <= 12 31,7 334 | <= 26 33,8 356 | <=40 19,5 205 | > 40 15,0 158 | Total 11,2 1053
Nodo 4 | <= 12 27,5 506 | <= 26 34,7 638 | <=40 19,9 366 | > 40 18,0 331 | Total 19,5 1841
Nodo 5 | <= 12 32,5 311 | <= 26 50,2 480 | <=40 12,1 116 | > 40 5,1 49 | Total 10,1 956
Nodo 6 | <= 12 11,3 107 | <= 26 50,9 482 | <=40 28,0 265 | > 40 9,8 93 | Total 10,0 947
Nodo 7 | <= 12 3,8 36 | <= 26 25,4 241 | <=40 41,0 389 | > 40 29,7 282 | Total 10,0 948
Nodo 8 | <= 12 1,6 15 | <= 26 18,4 173 | <=40 31,6 297 | > 40 48,4 455 | Total 10,0 940
Species
Operator
E. dunnii
E. grandi
1,0
2,0
Nodo 9 | <= 12 90,7 68 | <= 26 8,0 6 | <=40 1,3 1 | > 40 0,0 0 | Total 0,8 75
Nodo 10 | <= 12 22,1 200 | <= 26 37,0 335 | <=40 20,9 189 | > 40 20,0 181 | Total 9,6 905
Nodo 11 | <= 12 82,0 73 | <= 26 18,0 16 | <=40 0,0 0 | > 40 0,0 0 | Total 0,9 89
Nodo 12 | <= 12 23,4 394 | <= 26 33,9 572 | <=40 22,2 375 | > 40 20,5 346 | Total 17,9 1687
Nodo 13 | <= 12 91,8 89 | <= 26 8,2 8 | <=40 0,0 0 | > 40 0,0 0 | Total 1,0 97
Nodo 14 | <= 12 25,6 245 | <= 26 36,4 348 | <=40 21,4 205 | > 40 16,5 158 | Total 10,1 956
Nodo 15 | <= 12 67,9 144 | <= 26 31,6 67 | <=40 0,5 1 | > 40 0,0 0 | Total 2,2 212
Nodo 16 | <= 12 22,2 362 | <= 26 35,1 571 | <=40 22,4 365 | > 40 20,3 331 | Total 17,3 1629
Nodo 17 | <= 12 34,7 263 | <= 26 54,7 415 | <=40 9,4 71 | > 40 1,2 9 | Total 8,0 758
Nodo 18 | <= 12 24,2 48 | <= 26 32,8 65 | <=40 22,7 45 | > 40 20,2 40 | Total 2,1 198
Nodo 19 | <= 12 6,4 33 | <= 26 47,5 243 | <=40 34,0 174 | > 40 12,1 62 | Total 5,4 512
Nodo 20 | <= 12 17,0 74 | <= 26 54,9 239 | <=40 20,9 91 | > 40 7,1 31 | Total 4,6 435
Nodo 21 | <= 12 1,2 6 | <= 26 17,8 87 | <=40 42,8 209 | > 40 38,1 186 | Total 5,2 488
Nodo 22 | <= 12 6,5 30 | <= 26 33,5 154 | <=40 39,1 180 | > 40 20,9 96 | Total 4,9 460
Nodo 23 | <= 12 1,8 8 | <= 26 11,5 51 | <=40 29,5 131 | > 40 57,2 254 | Total 4,7 444
Nodo 24 | <= 12 1,4 7 | <= 26 24,6 122 | <=40 33,5 166 | > 40 40,5 201 | Total 5,3 496
Nodo 25 | <= 12 21,7 218 | <= 26 33,5 337 | <=40 22,3 224 | > 40 22,5 226 | Total 10,6 1005
Nodo 26 | <= 12 23,1 144 | <= 26 37,5 234 | <=40 22,6 141 | > 40 16,8 105 | Total 6,6 624
Nodo 27 | <= 12 28,0 120 | <= 26 57,1 245 | <=40 13,1 56 | > 40 1,9 8 | Total 4,5 429
Nodo 28 | <= 12 43,5 143 | <= 26 51,7 170 | <=40 4,6 15 | > 40 0,3 1 | Total 3,5 329

Figure 4. Global DT based on equal intervals' classification of productivity against all studied variables: DBH, Shit, and Operator.

If the DTs in Fig. A.1 (clusters) and Fig. 4 (equal intervals) are compared, it is clear that the DT based on clusters has a structure quite similar to the DT based on categories of equal intervals, although there are some differences. One of them is that the DT based on clusters uses 22 leaves nodes to describe productivity, while the DT model based on categories of equal intervals uses 18. Other subtler differences, but which mark the impact of categorization by clusters, are for example: the cases of nodes 31 and 32 of the DT based on clusters, and nodes 27 and 28 of the DT based on equal intervals. These pairs of nodes represent the same scenario, defined by DBH between 122 and 162 mm, and individuals of species *E. dunnii*, and where each node corresponds: the first to operator 1 and the second to operator 2 in each DT, respectively. If the nodes of operator 1 in both DTs (nodes 31 and 27, respectively) are considered, the purity of the nodes are quite similar and they are approximately 57%. While the nodes relative to operator 2, nodes 32 and 28 respectively, the purity is quite different, being in the case of clusters greater than 78%, while in the DT based on equal intervals it barely exceeds 50%.

1. **Table 8**. Confusion matrix evaluating the DT of Productivity by DBH, Shift and Operator based on k-means clustering classification presented in Figure A.1.

Observed	***Forecast***				***Correct %***
	<= 17	**<= 29.6**	**<= 43.7**	**> 43.7**	
<= 17	3022	347	58	23	87.6%
<= 29.6	1750	584	269	69	21.9%
<= 43.7	1054	356	378	137	19.6%
> 43.7	780	120	279	215	15.4%
Global %	70.0%	14.9%	10.4%	4.7%	44.5%

3.2. Test 2: By Operator

Nextly, we generate a more detailed analysis on the influence of the operators on productivity. For that we will analyse in isolation the work of operator 1 and 2.

3.2.1. Operator 1

In this section the productivity of operator 1 is assessed. For this, only the records generated by the activity of operator 1 are considered. In turn, given that the records include harvest of individuals of the two species, three separate studies are performed: i) considering the two species together, ii) considering only individuals of *E. grandis* species and iii) considering only individuals of *E. dunnii* species. Also, for each study the different categorizations of the productivity variable are analysed, considering the clusters of Table 2 for case i), categories from Table 4 for the case of *E. grandis* ii), and the categories of Table 3 for *E dunnii* iii). In all cases, categorization by equal interval is also used (Table 1).

Global

As already mentioned, in this case the performance of operator 1 is studied considering the global set of records. For this study, 2 possible categorizations of the productivity variable are considered: by equal intervals and by global clusters.

Table 9 summarizes and compares the results of the DT models based on categorizations by "Equal intervals" and by "Clusters". The columns associated with the results of the categories by Clusters are highlighted in gray. From Table 9 it can be seen that, as in the case of Table 8, clustering achieves an improvement in the overall effectiveness of the models for describing the productivity variable.

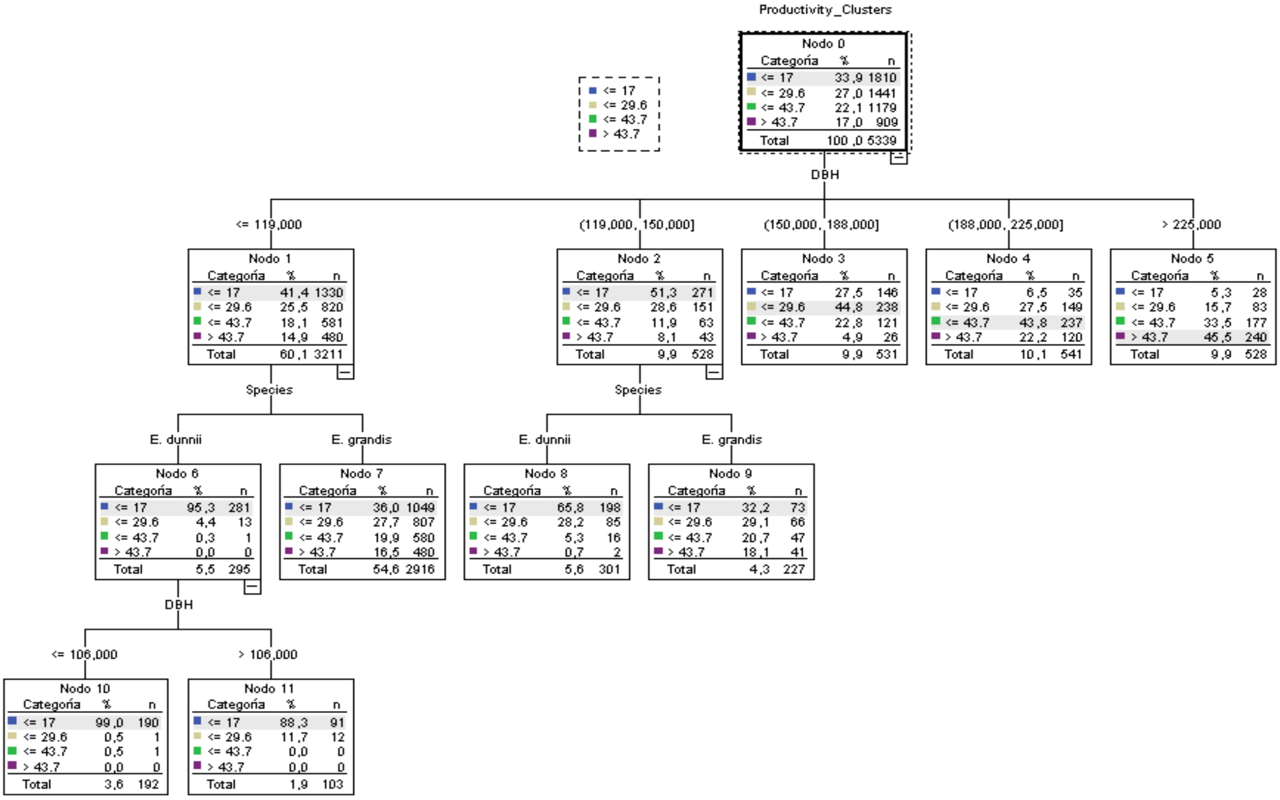

Figure 5. Operator 1 productivity Forecast model considering DBH, shift and specie as independent variables. The productivity classes are based on cluster analysis.

Regarding the DT models, both categorizations generated the same models, that is, the ramifications followed the same pattern and in the same descriptive variables. Therefore, we show only the model that obtained the best global performance according to their confusion matrices, that is, the model based on categories by clusters as shown in Figure 5. Something relevant to delimit the coinciding format of the DTs, is the sensitivity to how the dependent variable is defined and modelled. In this case, the same divisions are generated, and with the same levels, but by modifying the categorization of the dependent variable, the overall performance of the model is improved.

Table 9. Confusion matrix resume for Operator 1 comparing equal intervals and clusters productivity categorization.

Category		**Equal intervals**	**Clusters**
Equal intervals	**Clusters**	**Correct %**	**Correct %**
<= 12	<= 17	20.9%	88.5%
<= 26	<= 29.6	85.9%	16.5%
<=40	<= 43.7	18.1%	20.1%
> 40	> 43.7	24.1%	26.4%
Global %		**42.2%**	**43.9%**

Specifically, from Figure 5, for individuals with DBH greater than 150 mm, operator 1 tends to obtain productivities proportional to the size of the individuals it harvests: as the DBH grows, the node typical category grows, as appears in nodes 3, 4 and 5. Regarding the species, operator 1 is indifferent for the case of large individuals, since the variable species does not improve the purity of nodes 3, 4 and 5. In contrast, for individuals with DBH lower than 150mm, the specie is an influential variable. Mainly, for individuals of specie *E. dunnii*, the typical productivities are "<= 17" but with more than 65% of the cases (nodes 6 and 8), reaching 95% for small individuals (DBH lower than 119 mm). While, for individuals of *E. grandis* species, they have the same typical productivity of "<= 17", but the nodes are not as pure, for example for node 9 the 29% of cases have productivity of "<= 29.6"

E. grandis

In this new case, the productivity of the operator 1 is studied, limiting the analysis to the records of *E. grandis* species. For this purpose, the data is filtered considering only the records of *E. grandis*.

The three models are generated considering the three categorizations established: by equal intervals, by global clusters and by *E. grandis* clusters. The results of the confusion matrices of the three models are presented in Table 10. In this table, the white columns are associated to the equal interval categorization, the clear-coloured columns to global clusters, and the dark-coloured columns to the *E. grandis* clusters. From Table 10, it can be seen that all the models exceed 50% of efficiency, where the best models are the ones based on clusters. Categorization by species-specific clusters slightly improves global clusters categorization. In turn, categorizations by clusters achieve a more even model with respect to the different categories than the model based on equal interval criteria. In the categorization by equal intervals the class with the least number of correct predictions is the "<= 40" with 14.9%, while in the categorizations by clusters has 18.2% ("<= 43.7").

Table 10. Confusion matrix resume for Operator 1 comparing equal intervals, global clusters and *E. grandis* Clusters productivity categorization.

Category			**Equal intervals**	**C_Global**	**C_Grandis**
Equal intervals	**C_Global**	**C_Grandis**	**Correct %**	**Correct %**	**Correct %**
<= 12	<=17	<=16.9	60.40%	77.00%	77.20%
<= 26	<=29.6	<=29.8	58.40%	42.90%	43.50%
<=40	<=43.7	<=43.9	14.90%	18.20%	18.40%
> 40	>43.7	>43.9	70.30%	53.40%	53.60%
Global %			**51.80%**	**51.90%**	**52.20%**

Figure A.2 (Appendix) shows the best of the three models generated, the clusters-based model for the species *E. grandis*. In Figure A.2 it can be seen that operator 1 has a productivity proportional to the size of the individual based on the DBH, as the DBH increases the typical class of productivity increases. A special case are individuals with DBH between 174 and 205 mm (nodes 5 and 6), since for that DBH the shift becomes an influential variable. For example, in nodes 13 and 14, although the typical productivity of both nodes is "<= 29.8", in node 13 (corresponds to the night shift) the distribution of the percentages is concentrated between the classes "<= 16.9 "and" <= 29.8 ", with the 66.7% of the cases. On the other hand, in the day shift, node 14, the typical productivity is also "<= 29.8" but with 44% of the cases, and adding the class "<= 43.9" they group 73% of the cases. In this sense, since the harvested trees are of equal size at least in DBH, it is possible to affirm that operator 1 has a higher processing speed in the day shift than in the night shift.

E. dunnii

Like for the *E. grandis* species, the performance of operator 1 is analyzed with respect to the harvest of *E. dunnii* type trees. Again, the three different categorizations of the productivity variable are used for the DT: by equal intervals, by global clusters, and by *E. dunnii* clusters. The results of the confusion matrices for each categorization are shown in Table 11. Once more, the white columns and clear gray correspond to categorizations by equal intervals and global clusters, respectively. While the dark gray columns correspond to the *E. dunnii* cluster-based categorization.

From Table 11 it is observed that the prediction models for Operator 1 in harvesting of trees of species *E. dunnii* have greater precision than the analogous models for the species *E. grandis*. For the *E. dunnii* the global accuracy exceeds 55% while for *E. grandis* it hovered around 52%. On the other hand, it is possible to identify that categorizations by clusters do not improve the prediction for *E. dunnii* records. For example, in the categorization by equal intervals the model achieves a predictive capacity of 55.7%, while when using categorization by global clusters, it decreases slightly to 55.4% and for the categorization of the species cluster decreases to 55.2%. It is also interesting to note, that the differences between the models do not exceed 0.5%. Considering the case of categorization by equal intervals, the best of the three, it is seen that for all the categories the prediction capacity is notoriously superior than it is to the case of *E. grandis* shown in Table 10. In this case the third class, "<= 40 ", is the least effective predictor and has 45.5% of the correct cases while in Table 10, the worst of the classes does not exceed 20%.

Table 11. Confusion matrix resume for Operator 1 comparing equal intervals, global clusters and *E. dunnii* Clusters productivity categorization.

Category			**Equal intervals**	**C_Global**	**C_*E. dunnii***
Equal intervals	**C_Global**	**C_*E. dunnii***	**Correct %**	**Correct %**	**Correct %**
<= 12	<= 17	<= 17.2	73.1%	73.3%	72.4%
<= 26	<= 29.6	<= 29.5	54.0%	36.0%	35.7%
<=40	<= 43.7	<= 43.4	45.5%	52.2%	52.1%
> 40	> 43.7	> 43.4	55.1%	56.4%	56.2%
Global %			**55.7%**	**55.4%**	**55.2%**

With respect to the prediction models, we present in Figure A.3 (Appendix) the model based on the categorization of the equal intervals as the one that best describes the productivity of operator 1 in individuals of *E. dunnii* species. In this DT, the operator tends to have matched productivities according to DBH, as the DBH increases, the nodes tend to be dominated by higher productivities categories. However, this relationship is clearer seen in extreme cases, that is, for low values of DBH the dominance of low productivity categories is marked in those nodes. The same for the opposite case, high DBH imply a clear dominance of the nodes by high productivities. In the middle area of the DBH, [201, 216] mm, in node 7, we see that the dominant productivity is "<= 40" with 41.8% of the records. However, if the shift is incorporated as a descriptive variable, operator 1 achieves typical productions of "> 40" in 52.9% of the cases, and in the night shift "<= 40" in 43.8%.

3.2.2. Operator 2

In this case we will analyse the performance of the operator 2. As for operator 1, the records of the 2 species are initially worked together, to then disaggregate their performance for each species.

Global

Next, the results of the productivity of operator 2 are presented considering the global set of records. For this study, 2 possible categorizations of the productivity variable were considered: by equal intervals and by global clusters. In Table 12 the results of the confusion matrices are presented for each categorization of the productivity variable, the results of the categories by clusters are highlighted in gray.

Table 12. Confusion matrix resume for Operator 2 comparing equal intervals and clusters productivity categorization.

Category		**Equal intervals**	**Clusters**
Equal intervals	**Clusters**	**Correct %**	**Correct %**
<= 12	<= 17	18.3%	88.1%
<= 26	<= 29.6	84.5%	27.8%
<=40	<= 43.7	17.7%	17.5%
> 40	> 43.7	24.0%	0.0%
Global %		**42.4%**	**45.7%**

From Table 12 the cluster categorization manages to improve the overall accuracy of the prediction model (45.7% versus 42.4%). However, a striking case is that the model based on categories by clusters, fails to predict any correct case for high productivities. While the model based on equal intervals' categorization manages to predict at least 24% of cases of high productivity. For this reason, both models are presented on figures, since one achieves better prediction of high productivity categories (based on equal interval criteria, Figure A.4), and the other, better global efficiency (based on clusters, Figure 6).

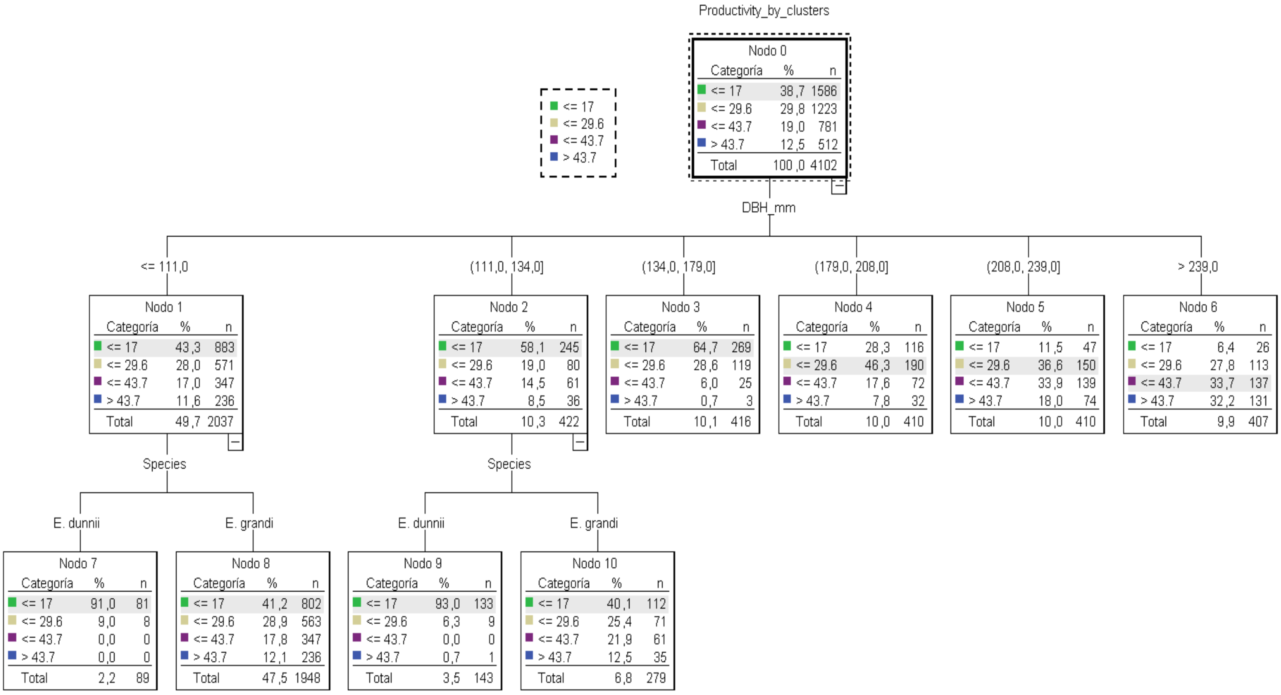

Figure 6. Operator 2 productivity Forecast model considering DBH, shift and specie as independent variables. The productivity classes are based on cluster analysis.

When comparing both DTs it is possible to see that the one based on categories by equal intervals (Figure A.4, Appendix) generates a more detailed division of the productivities according to DBH, generating 8 nodes against 6 generated by the DT that uses clusters categorization (Figure 6). In turn, the model with categorization by equal intervals, shows a high incidence of the species variable, which branches the 5 nodes with DBH less than 162 mm, while in the cluster model the species variable branches only 2 nodes with lower DBH to 134. In both models, it is observed that the operator 2 for trees with low DBH, has a very precise productivity for the species *E. dunnii*. This is evident in the degree of purity of nodes 9, 11, 13, 15 and 17 of Figure A.4, with purities greater than 50%, and in many cases higher than 90%. For the model of Figure 6, the same behavior is observed, purity greater than 90% for terminal nodes of the species *E. dunnii*. Although in the first model (Figure 6) the shift variable is not determinant to predict productivity, in the second (Figure A.4) it collaborates in the description, mainly, for high DBH values (greater than 162 mm).

E. grandis

In this section the performance of operator 2 for individuals of the species *E. grandis* is detailed. Again, as in the case of operator 1, the 3 categorizations of the productivity variable are used: based on equal intervals, based on global cluster and based on *E. grandis* clusters.

The results of the confusion matrices are shown in Table 13. It can be observed that all the prediction models achieve a predictive capacity of around 50%, where the ones based on cluster categorizations are higher. Among the categories based on global clusters and on *E. grandis* clusters no differences in overall effectiveness are observed, and between the categories the differences are very subtle, less than 1%. As for the format of the DT, those based on clusters have the same branch structure and the DT based on equal intervals' categories, it is also quite similar. Figure 7 shows the model based on global clusters.

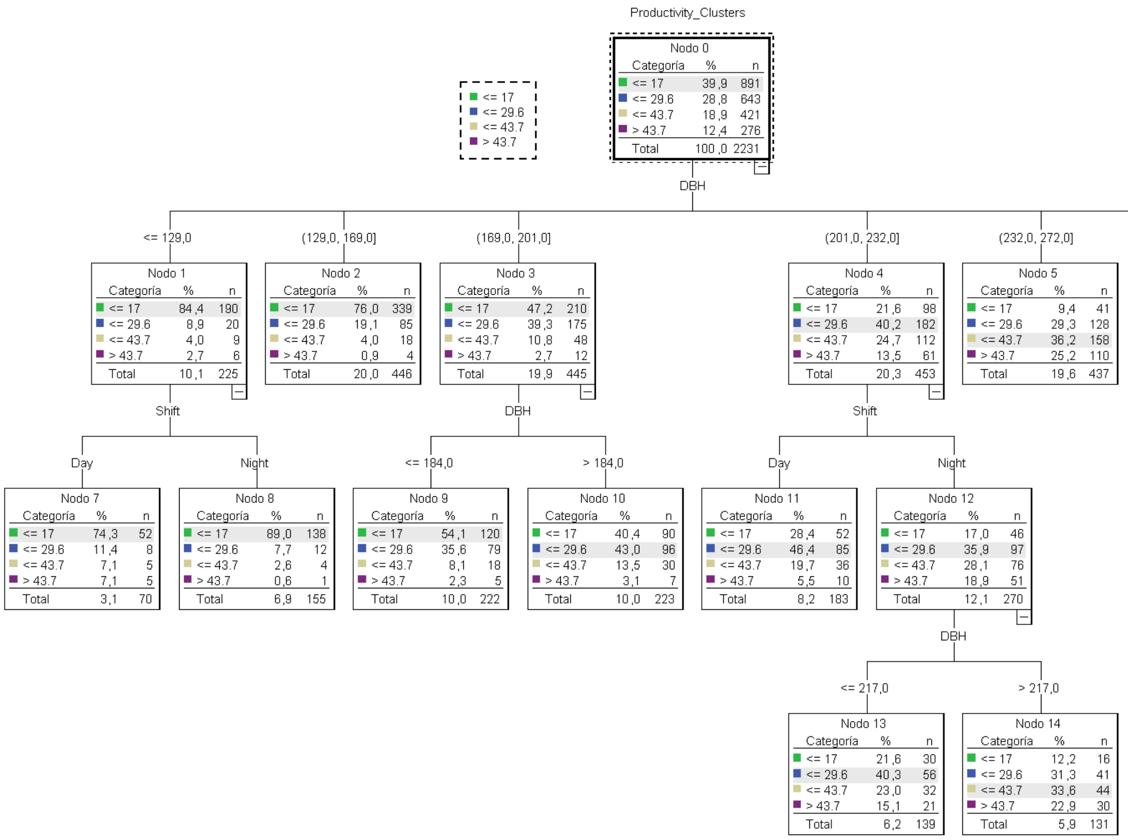

Figure 7. Operator 2 productivity Forecast model for *E. grandis* specie considering DBH and shift as independent variables. The productivity classes are based on global clusters.

What can be seen in Figure 7 about operator 2 for the species *E. grandis* and individuals with DBH less than 201 mm, is that it tends to have low typical productivities with a considerable purity in the nodes. Meanwhile for large individuals, with DBH greater than 201 mm, it tends to have high typical productivities, but the purity of the nodes drops considerably. Regarding the influence of the shift, it is not such as decisive on productivity.

Table 13. Confusion matrix resume for Operator 2 comparing equal intervals, global clusters and *E. grandis* Clusters productivity categorization.

Category			**Equal intervals**	**C_Global**	**C_Grandis**
Equal intervals	**C_Global**	**C_Grandis**	**Correct %**	**Correct %**	**Correct %**
<= 12	<= 17	<= 16.9	70.7%	72.8%	73.0%
<= 26	<= 29.6	<= 29.8	48.2%	36.9%	36.7%
<=40	<= 43.7	<= 43.9	45.7%	48.0%	48.6%
> 40	> 43.7	> 43.9	29.3%	30.1%	30.4%
Global %			**50.0%**	**52.5%**	**52.5%**

E. dunnii

In this new section we analyse the performance of operator 2 considering only individuals of the species *E. dunnii*. Prediction models are generated for each of the categorizations: equal intervals, global clusters and clusters considering only *E. dunnii*.

Table 14 shows the results of the confusion matrices using each possible categorization. The columns are associated by colours, the white columns correspond to categorization by equal intervals, clear-coloured by global cluster and dark-coloured by clusters of *E. dunnii*. From Table 14, is observed that the productivity prediction models of operator 2 for *E. dunnii* individuals are around 50% of accuracy. The categorizations by cluster manage to improve this accuracy in values of 1% and 1.4% for global clusters and clusters of the species *E. dunnii*, respectively. However, the clusters-based models fail to predict productivities between 29.6 m^3/h and 43.7 m^3/h for global clusters, and between, 29.5 m^3/h and 43.4 m^3/h for *E. dunnii* clusters. Given that the differences between global efficiencies for different models do not exceed 2%, we will choose the model based on categorizations by equal intervals as the one that best describes the performance of operator 2 with individuals of the species *E. dunnii*. The model is presented in Figure A.5 (Appendix).

Table 14. Confusion matrix resume for Operator 2 comparing equal intervals, global clusters and *E. dunnii* Clusters productivity categorization.

Category			**Equal intervals**	**C_Global**	**C_*E. dunnii***
Equal intervals	**C_Global**	**C_*E. dunnii***	**Correct %**	**Correct %**	**Correct %**
<= 12	**<= 17**	**<= 17.2**	38.4%	66.6%	66.5%
<= 26	**<= 29.6**	**<= 29.5**	56.6%	74.7%	75.0%
<=40	**<= 43.7**	**<= 43.4**	33.6%	0.0%	0.0%
> 40	**> 43.7**	**> 43.4**	66.6%	29.4%	29.6%
Global %			**49.2%**	**50.4%**	**50.6%**

In Figure A.5 (Appendix), from node 0 it can be seen that productivities tend to concentrate on productivities between higher than 12 m^3/h and lower or equal than 40 m^3/h, by 60%. A striking aspect is that the only variable that allows predicting productivity is the variable DBH. This implies an indifference of the operator 2 in its operation to the different shifts. Regarding the productivity according to the DBH, is evidenced the relation: the higher the DBH, the higher the typical productivity.

3.4. Test 3: By Shift

In the study of Test 3, the influence of the different shifts on productivity is analyzed. To improve the understanding of this operative aspect we will analyze each shift in isolation.

1.1.1. Night Shift

In this section the prediction models considering only the night shift is analyzed. As in the case of operators, the study is divided into three cases: I) considering the two species together, II) considering only *E. grandis* records and III) considering only *E. dunnii* records. In each case the different categorizations of the productivity variable are considered.

Global

In this first study, both species are considered together, thereby the intervening categorizations are: based on equal intervals and based on global clusters.

Table 15. Confusion matrix resume for Night Shift comparing equal intervals and clusters productivity categorization.

Category		**Equal intervals**	**Clusters**
Equal intervals	**Clusters**	**Correct %**	**Correct %**
<= 12	<= 17	28.9%	90.9%
<= 26	<= 29.6	82.1%	17.8%
<=40	<= 43.7	21.1%	19.8%
> 40	> 43.7	23.2%	26.1%
Global %		**43.8%**	**45.4%**

Table 15 presents the results of the confusion matrices considering only the night shift and categorizations based on equal intervals and global clusters. The results of global clusters are in the gray columns. Table 15 shows that the overall accuracy for each prediction model has values similar to the rest of the results that consider both species together (Tables 2, 3, 7 and 12), and these are between 40% and 45% efficiency. In Table 15, it is shown that cluster categorization improves the global level of prediction, and, punctually, a significant fact is that for the smallest category of the clusters ("<= 17"), the model manages to predict 90.9% of the values correctly.

Figure A.6 (Appendix) shows the model based on categorization by clusters. In it, it can be observed that in several cases the species and operator variables are influential regarding the level of productivity in the night shift. In the case of large individuals, DBH greater than 199 mm, it is manifest that operator 1 achieves superior productivities to those of operator 2. An example of this difference are nodes 15 and 16, node 15 corresponds to operator 1 and shows that it manages to obtain productivities superior to 29.6 m^3/h in 72% of the cases, being the category "<= 43.7" the most probable with 46.5% of the cases. While operator 2, achieves productivities higher than 29.6 m^3/h in 48% of cases, and the most likely productivity is "<= 29.6" with 38% of cases. Regarding the species, it is identified that the species *E. dunnii*, at least for DBHs smaller than 167 mm, tends to have more distributed productivities within the same range of DBH. This is evidenced in the purity of nodes 7, 9 and 11, reaching a purity of 99% for node 7.

E. grandis

In this case we will consider the harvest in the night shift considering only individuals of the species *E. grandis*. Therefore, equal intervals based, global clusters and *E. grandis* clusters are considered as possible categorizations.

Table 16. Confusion matrix resume for Night Shift comparing equal intervals, global clusters and *E. grandis* Clusters productivity categorization.

Category			**Equal intervals**	**C_Global**	**C_Grandis**
Equal intervals	**C_Global**	**C_Grandis**	**Correct %**	**Correct %**	**Correct %**
<= 12	<= 17	<= 16.9	64.8%	84.5%	74.0%
<= 26	<= 29.6	<= 29.8	58.7%	26.4%	41.6%
<=40	<= 43.7	<= 43.9	35.9%	38.2%	38.7%
> 40	> 43.7	> 43.9	26.2%	28.5%	28.8%
Global %			**49.3%**	**51.7%**	**51.8%**

Table 16 presents the results of the confusion matrices for the models based on the three possible categorizations of productivity. The columns are associated by colours, white considers categorization by equal intervals, light gray by global clusters, and gray-dark by *E. grandis* clusters. The three models have efficiencies of around 50%, where models based on categories by clusters are better. In turn, the model that uses *E. grandis* clusters is the best in the overall value, and in terms relative to the categories, it also achieves a greater proportion of efficiency among the different categories. The best predicted category is the lowest productivity category “<= 16.9” with an efficiency of 74%.

Figure 8 shows the model based on categories by *E. grandis* clusters, because it best describes the productivity in the night shift considering individuals of *E. grandis*. In the model it can be observed that for small individuals, productivity tends to be consistently low. Proof of this is the purity of nodes 7 and 8, where productivity "<= 16.9" exceeds 80% of cases. On the other hand, the operator factor is significant for individuals with a diameter between 146 and 179 mm, where operator 2 is regular in its productivity, sustaining low productivity categories "<= 16.9", with purities at nodes 10 and 14 of 76% and 57%, respectively. In contrast, operator 1, for the same cases, also has a typical productivity of "<= 16.9", however, it does it in a smaller percentage. This implies that, for those diameters, operator 1 achieves higher productivities implying a greater speed in the processing of individuals.

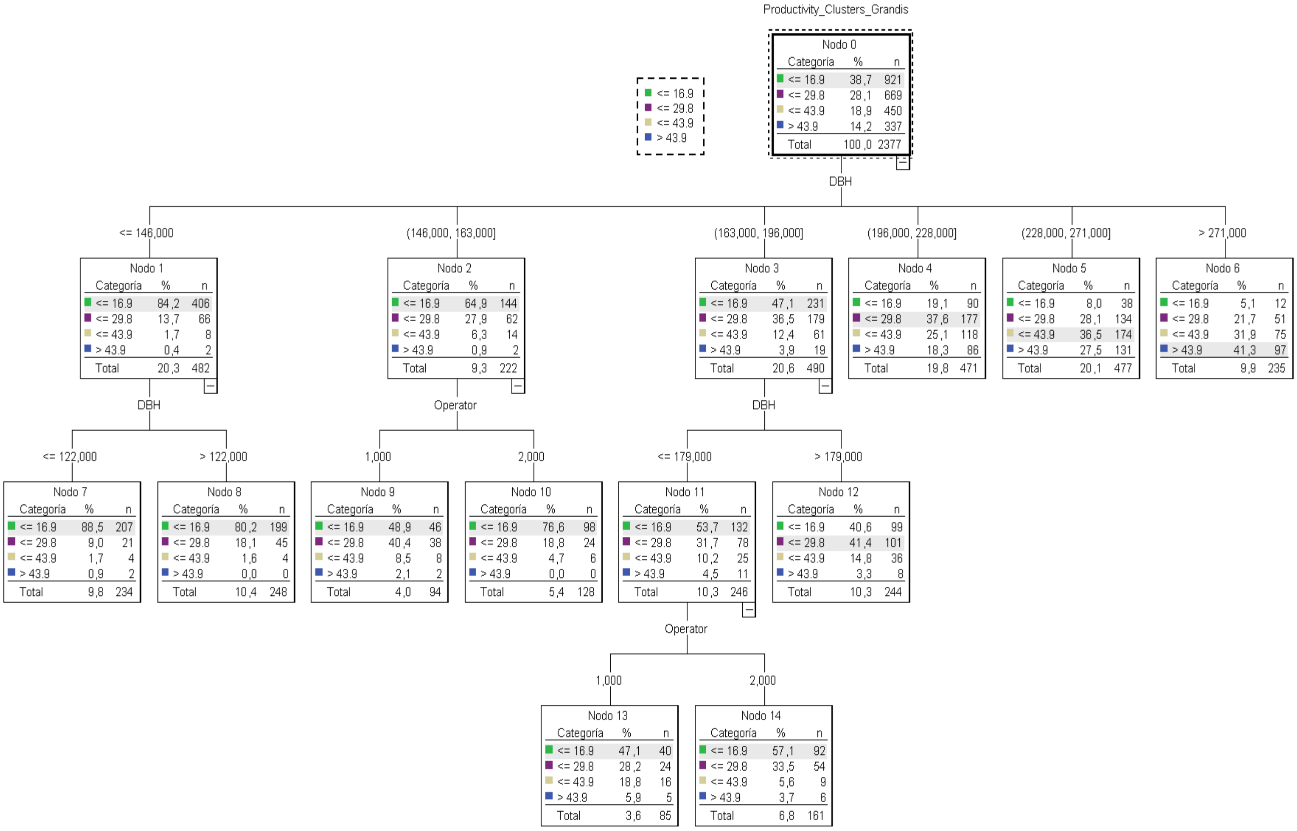

Figure 8. Night Shift Productivity Forecast model for *E. grandis* specie considering DBH and operator as independent variables. The productivity classes are based on *E. grandis* clusters.

E. dunnii

In this case, the harvest of individuals of *E. dunnii* in the night shift is analyzed. For this, the three options of the productivity variable categorization are considered: by equal intervals, global clusters and *E. Dunni* clusters. The results of the three models are presented in Table 17, where again the columns corresponding to each categorization of productivity are associated by colours.

Table 17. Confusion matrix resume for Night Shift comparing equal intervals, global clusters and *E. dunnii* Clusters productivity categorization.

Category			**Equal intervals**	**C_Global**	**C_*E. dunnii***
Equal intervals	**C_Global**	**C_*E. dunnii***	**Correct %**	**Correct %**	**Correct %**
<= 12	<= 17	<= 17.2	71.7%	71.9%	71.4%
<= 26	<= 29.6	<= 29.5	54.1%	23.9%	24.0%
<=40	<= 43.7	<= 43.4	50.0%	64.3%	60.0%
> 40	> 43.7	> 43.4	43.7%	49.0%	55.0%
Global %			**54.1%**	**53.4%**	**53.4%**

From Table 17 the general level of prediction is above 50%. With respect to the three categorizations, it can be observed that for the night shift and with individuals of the species *E. dunnii*, the one that best predicts is the categorization by equal intervals. This is also the most stable categorization regarding the different categories, since the category with the lowest prediction efficiency is "> 40" with 43.7%. Cluster categorizations have similar overall performance, but their prediction level is more disproportionate between categories.

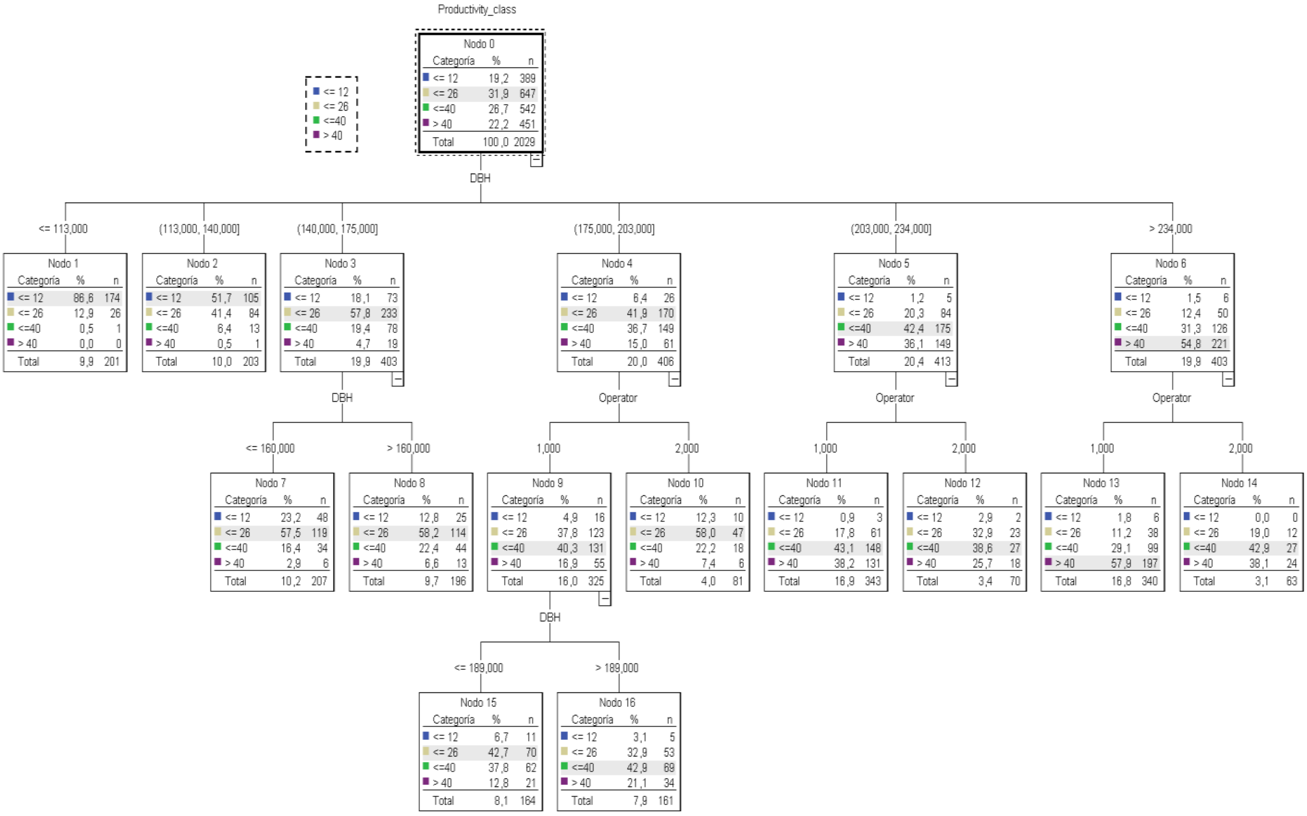

Figure 9. Night Shift Productivity Forecast model for *E. dunnii* specie considering DBH and operator as independent variables. The productivity classes are based on equal intervals.

The prediction model shown in Figure 9 is the most effective one, which is the model based on the equal interval criteria. In the model it can be seen in node 0 that the relations between the productivities is fairly even, ranging from 19.2% for "<= 12" to 31.9% for "<= 26". As a general comment it is possible to identify that the productivity in the night shift of individuals with DBH greater than 175 mm tends to be quite uneven among operators. For example, in trees with DBH between 175 and 203 mm, operator 1 achieves productivities superior to 26 m^3/h in 57.2% of the cases (adding the two superior categories). On the other hand, operator 2 achieves this in 29.6% of cases. A particular case is for individuals with DBH greater than 234mm, for these cases both operators achieve productivities superiors to 26 m^3/h for 80% of the cases, only that operator 1 in almost 58% of the cases achieves productivity "> 40" while that the operator 2 does it only in 38.1% of the cases.

1.1.2. Day Shift

In contrast to the previous study, in this case only harvest records made in day shift are considered. As in the previous studies, first a global study including both species is carried out, and then the study for each species is particularized.

Global

In this first study of the day shift, the records of both species are considered and as possible categorizations of the productivity variable, we have those based on equal interval criteria and on global clusters. Table 18 presents the results of the confusion matrices for the models based on both categorizations. The columns associated with each model are distinguished by colours.

When analysing the efficiencies of the prediction models in Table 18, it is observed that the global level is 41%, being a little better for the case of using categorizations by global clusters. However, this categorization has a very different precision among the different categories, reaching 90% accuracy for the category "<= 17" and falling to 5.6% for "> 43.7". In this sense, the model based on categories by equal intervals is slightly better, although the global level of prediction decreases a little, between the categories it has a better precision.

Table 18. Confusion matrix resume for Day Shift comparing equal intervals and clusters productivity categorization.

Category		**Equal intervals**	**Clusters**
Equal intervals	**Clusters**	**Correct %**	**Correct %**
<= 12	<= 17	13.4%	90.2%
<= 26	<= 29.6	83.4%	17.3%
<=40	<= 43.7	13.3%	17.8%
> 40	> 43.7	26.8%	5.6%
Global %		**41.3%**	**41.6%**

The model shown in Figure 10 is the one obtained by considering the productivity variable by the equal interval categories. In the same it is observed that the operator has a strong influence on the prediction, and this influence depends on the size of DBH. The higher DBH, the greater the differences between operators. For DBH greater than 195 mm, operator 1 achieves productivities clearly superior to operator 2. For example, for DBH between 195 and 233 mm, operator 2 has a typical productivity of "<= 40" and in 60% manages to overcome the 26 m^3/h. And the operator 1, has a typical productivity "> 40" and in 83% of the cases it exceeds 26 m^3/h.

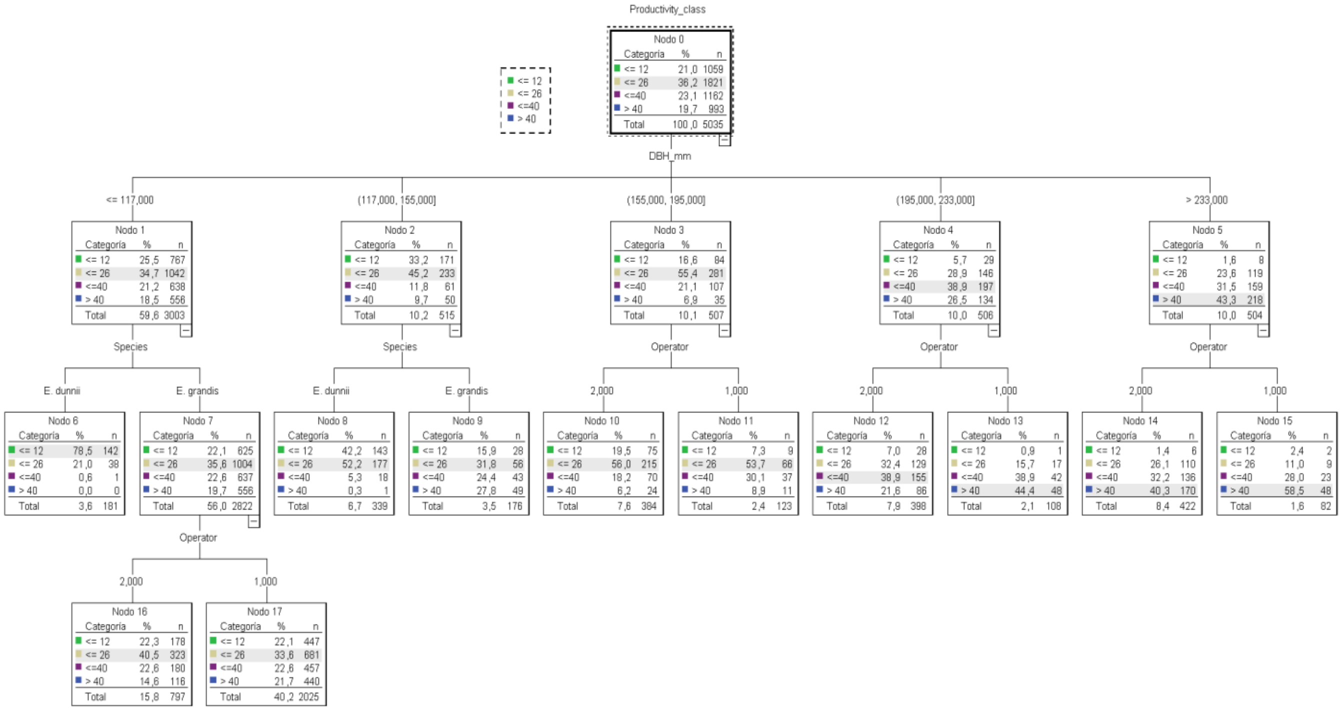

Figure 10. Day Shift Productivity Forecast model considering DBH, specie and operator as independent variables. The productivity classes are based on equal interval criterion.

E. Grandis

In this case the impact of the day shift when harvesting individuals of *E. grandis* is studied. The categorizations considered are based on equal intervals, global clusters and *E. grandis* clusters. Table 19 shows the results of the models based on the three possible categorizations. The columns that present the results of each categorization are distinguished by colours.

From the precision data of the prediction models in Table 19, it is observed that clustering allows to slightly improve the overall predictive capacity of the model, and among cluster categorizations, the clusters of the species have the best result. However, the most even performance in predicting the different categories is the one based on equal interval criteria. Therefore, the latter is chosen to show and analyze the study of the impact on productivity of the day shift in individuals of the species *E. grandis*.

Table 19. Confusion matrix resume for Day Shift comparing equal intervals, global clusters and *E. grandis* Clusters productivity categorization.

Category			**Equal intervals**	**C_Global**	**C_Grandis**
Equal intervals	**C_Global**	**C_Grandis**	**Correct %**	**Correct %**	**Correct %**
<= 12	<= 17	<= 16.9	62.5%	66.3%	66.5%
<= 26	<= 29.6	<= 29.8	49.9%	62.9%	63.3%
<=40	<= 43.7	<= 43.9	33.5%	19.7%	19.9%
> 40	> 43.7	> 43.9	60.3%	41.1%	41.4%
Global %			**51.0%**	**52.2%**	**52.6%**

The prediction model shown in Figure A.7 (Appendix) is obtained by modelling the productivity variable through the categories defined by equal intervals. The other two models have a very similar structure and coincide in the branching of the first level according to DBH and in the same intervals. Specifically, from the model presented in Figure A.7, it can be observed that in the day shift the productivity of individuals of *E. grandis* species is particular for each operator. All DBH intervals are branched according to the operator. As in the previous cases, for the same type of individual to be harvested, the operator 1 tends to achieve better productivities and these improvements are accentuated as the individual has a larger DBH. As is the case of individuals with DBH between 193 and 224 mm, nodes 17 and 18, in which the operator 1 achieves productivities exceeding 26 m^3/h in 62% of cases, and operator 2 only in 32%, that is, it practically doubles in productivity. For individuals with larger DBH it is maintained that operator 1 achieves higher productivities than operator 2.

E. dunnii

In this last study, the impact of the diurnal shift in individuals of the *E. dunnii* is analyzed. The categorizations considered are based on equal intervals, global clusters and *E. dunnii* clusters. Table 20 shows the results of the models based on the three possible categorizations. The columns that present the results of each categorization are associated by colours.

From Table 20, it is observed that the models generated have an efficiency in the prediction of daytime productivity of the species *E. dunnii* of approximately 50%, being slightly better for clusters categorizations. Nonetheless, when considering the ability to describe the different categories, the model based on equal interval criteria performs better than the cluster-based models. Therefore, to describe the productivity in individuals of *E. dunnii* in the day shift, the model based on categories by equal intervals is used.

Table 20. Confusion matrix resume for Day Shift comparing equal intervals, global clusters and *E. dunnii* Clusters productivity categorization.

Category			**Equal intervals**	**C_Global**	**C_*E. dunnii***
Equal intervals	**C_Global**	**C_*E. dunnii***	**Correct %**	**Correct %**	**Correct %**
<= 12	<= 17	<= 17.2	38.9%	77.6%	77.6%
<= 26	<= 29.6	<= 29.5	59.4%	48.8%	49.2%
<=40	<= 43.7	<= 43.4	36.1%	12.4%	12.1%
> 40	> 43.7	> 43.4	63.4%	52.6%	52.5%
Global %			**50.6%**	**51.7%**	**52.1%**

Figure 11 shows the model based on equal interval criteria. In this model it can be observed that the productivities tend to be well proportional to the DBH, as the DBH increases, the typical productivity of the different nodes also increases. On the other hand, there is a clear differentiation between operators' productivity for individuals with DBH greater than 164mm. In all these cases operator 1 has greater productivity than operator 2, in some cases directly the typical productivity of operator 1 is higher than that of operator 2, as in nodes 10 and 11. In other cases, although they have the same typical productivity, the distribution of cases per node is different, as in nodes 12 and 13.

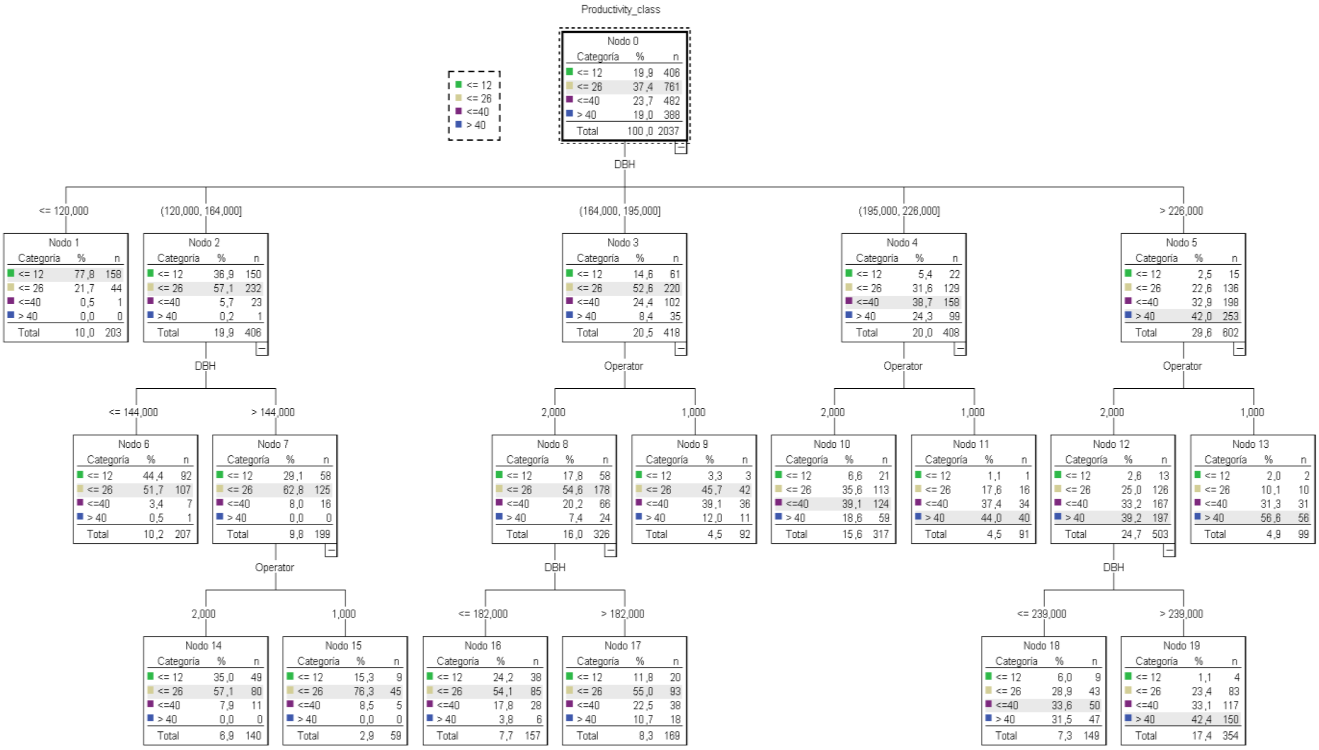

Figure 11. Day Shift Productivity Forecast model for *E. dunnii* specie considering DBH and operator as independent variables. The productivity classes are based on equal interval criterion

4. Discussion

In this section, both, the implementation of the methodologies and the results obtained are discussed. First, it is discussed whether the methodologies are adequate and effective for the proposed study and, then, the most important results of the study of productivity in CTL forest harvesting systems are discussed.

4.1. Discussion of the Methodologies used

The Big Data approach used consists on the implementation of DT and classification by clusters. Regarding DT, it turned out to be a useful method for a study of this type. Since the study sought to describe the influence factors of productivity in forest harvest, and these factors had different natures: scalar, nominal or categorical. In this sense, DT allow an integral analysis incorporating and combining the influence of different independent variables on a dependent. This allowed to describe scenarios or situations in which productivity can be described or predicted with a high level of precision, as was the case of high purity nodes throughout the study. By analyzing all the results from a more generic perspective it is possible to establish that the general prediction capacity was approximately 50%. In this sense, our results agree with others in the literature such as (Erikson & Lindroos 2014). Erikson & Lindroos (2014), used a large amount of data collected in Sweden's forest systems, where they are identified by statistical models at 55%. However, these authors focused their study on the impact of the characteristics of the machinery on productivity, such as the load capacity of the machine or the type of head, and in our case, we analyze factors that are more related to the management of operations. In order to establish some precise criterion on whether this value is high or low, it would be necessary to contrast it with other values of the literature, which is not possible due to the lack of this type of study as mentioned in the introduction.

Regarding the use of classification by clusters, it can be mentioned that, in general terms, it collaborates to improve the performance of DT. This improvement validates, at least for this study, *k*-means as a proper method for modelling productivity classes. Even more, *k*-means has the great advantage of enabling the automation of this categorization process if necessary. However, further studies should be considered for confirming this primary result. In general terms, the models based on clusters tended to be a little more precise, without this gain in precision being overwhelming. When considering our classification in the range of productivity, it can be noted that our productivity levels are similar to others found in the literature. Gerasimov et al. (2012a) and Gerasimov et al. (2012b) analyze the productivity of the CTL system in Russian forests, and find that, together with the size of the log, the species of the harvested tree are the main factors that describe the level of productivity. In their studies they determined that on average the productivities vary between 16 and 49.5 m^3 h^{-1}, ranges similar to those found in the analyzed systems of Uruguay.

4.2. Results Discussion

The first significant conclusion about the results is the strong dependence of productivity on the variable DBH. This dependency is expected, since it is a direct estimator of the amount of wood that each harvested individual contains. Results that are very frequent in the literature of the subject (Strandgard et al. [2013]; Gerasimov et al. [2012a]; Tolan & Visser [2015]; Strandgard et al. [2017]). However, by being able to analyse productivity by DT, relationships of other variables could be discovered from different DBH ranges. In this sense, the descriptive capacity of other variables was improved with respect to productivity. For example, in Figure A.1 nodes 17, 29 and 30, allow to identify relationships between productivity and tree species with a very high precision (higher than 90% in node 17) when limiting the study to individuals with DBH between 111 and 122. Which in turn manages to improve up to 98% efficiency in node 30, by linking the operator variable.

Regarding the analysis of the operators, the general conclusion is that operator 1 has higher productivities than operator 2, and that this superiority is more noticeable for individuals with higher DBH. When considering the influence of the shifts, the operator 1 manages to greatly exceed the productivity of operator 2 in the night shift. Analyzing punctually the operator 1, it was noted that its productivity in the day shift was clearly superior to that of the night shift for individuals with intermediate DBH (Figure A.3 nodes 11, 12, 13 and 14, and Figure 6 nodes 11 and 12). The reason for this must be that during the day, operator 1 has a higher processing speed than at night. On the other hand, operator 2 has more similar productivities during both shifts. The notorious differences in productivity rate between operators match to similar studies in literature, where the experience of the operator is a determining factor in the level of productivity (Purfürst [2010]; Hiesl & Benjamin [2013]). Moreover, Purfürst (2010) carried out an in-depth study on the impact of training on the productive capacity of workers, indicating that at the end of the training process they double their initial production capacity, this trend is evident in our case.

Considering the shift studies, it was discovered that in the night shift for the species *E. dunnii* and DBH greater than 175 mm, the operator 1 produces much more than the operator 2 (Figure 10 nodes 9, 10, 11, 12, 13 and 14). While in the day shift the differences between operators were noted in the species *E. grandis,* operator 1 achieved better productivity for virtually all DBH ranges. Although the study of shifts is not one of the classic aspects when analyzing productivity in the literature, recently (Häggström & Lindroos 2016) proposed a framework to carry out future studies from which it is possible to analyze new aspects in the productivity of the forest harvest. In that framework it is inferred that the shift scheduling is a factor that should not be ignored. Our results support this inference.

Table 21. Summary of rules with purity level over 80%.

Test	***Conditions***	***Rule***	***Purity %***	***Figure***
1	Prod. Class: equal interval	DBH ≤ 94 and Specie *E. dunni*: Productivity "≤ 12"	90.7%	4
		DBH (94, 102] and Specie *E. dunni*: Productivity "≤ 12"	82%	
		DBH (102, 106] and Specie *E. dunni*: Productivity "≤ 12"	91.8%	
	Prod. Class: Global cluster	DBH ≤ 94 and Specie *E. dunni*: Productivity "≤ 17"	96%	A.1
		DBH (102, 106] and Specie *E. dunni*: Productivity "≤ 17"	97.9%	
		DBH (106, 111] and Specie *E. dunni*: Productivity "≤ 17"	91.4%	
		DBH (111, 122] and Specie *E. dunni*: Productivity "≤ 17"	93.5%	
2: Op. 1	Prod. Class: Global cluster	DBH ≤ 119 and Specie *E. dunni*: Productivity "≤17"	95%	5
	Prod. Class: *E.grandis* cluster, only *E. grandis* Records	DBH ≤ 116: Productivity "<= 16.9"	90%	A.2
	Prod. Class: equal interval, only *E. dunni* Records	DBH ≤ 109: Productivity "≤ 12"	86.6%	A.3
2: Op. 2	Prod. Class: Global cluster	DBH ≤ 111, and Specie *E. dunni*: Productivity "≤ 17"	91%	6
		DBH (111, 134] and Specie *E. dunni*: Productivity "≤ 17"	93%	
	Prod. Class: equal interval	DBH ≤ 94, and Specie *E. dunni*: Productivity "≤ 12"	90.7%	A.4
		DBH (94, 102] and Specie *E. dunni*: Productivity "≤ 12"	82%	
		DBH (102, 106] and Specie *E. dunni*: Productivity "≤ 12"	91.8%	
	Prod. Class: Global cluster, only *E. grandis* Records	DBH ≤ 129: Productivity "≤ 17" (Including Shift variable, DBH ≤ 129 and Shift Night, the purity rises to 89%)	84%	7
3: Shift Night	Prod. Class: Global cluster	DBH ≤ 102 and Specie *E. dunni*: Productivity "≤ 17"	99%	A.6
		DBH (102, 129] and Specie *E. dunni*: Productivity "≤ 17"	90.5%	
	Prod. Class: *E.grandis* cluster, only *E. grandis* Records	DBH ≤ 146: Productivity "≤ 16.9"	84.2%	8
	Prod. Class: equal interval, only *E. dunni* Records	DBH ≤ 113: Productivity "≤ 12"	86.6%	9

In Table 21 all the rules with a precision over 80% are presented. This rules are related to which Test they belong, the conditions of the experiments and in which Figure they are. As a general comment from Table 21 is possible to identify that the harvest system considered for this study tends to have a very regular productivity for small trees, as most of the DBH involved in these rules are lower than 130 mm. This shows that the studied CTL system produce in a regular rate for trees that are not large. This can be justified by the fact that harvesting and managing small trees can be a more regular operation, than dealing with big and heavy trees. Another consideration from these rules, is that Operator do not seem to be an important factor, both operators can have good productivity (regarding Table 21 data). About the Shift, a similar appreciation can be done, however for the case of the Rule of Figure 7 (i.e. considering only operator 2 and *E. grandis* records), shift tends to have impact on productivity. The specie influence productivity, at least for our case of study, since in every rule of Table 21 it appears as a descriptor variable (in the rule enunciation or in the data filtering).

When considering the general forestry system to which this harvest system belongs, we can say that the main production line tends to consider small trees, since most of the harvested wood is used for pulp production. In this scenario, our approach contributes to orientate the forest planner in how to assign the resources for the harvesting, and on which factors it should pay attention and on which it could be indifferent.

On the other hand, it is interesting to comment on the enormous potential of having automatic data collection, since as studied Strandgard et al. (2013) and Brewer et al. (2018), these data allow us to analyze (contemplating the necessary precautions) the harvest systems in terms of productivity and operational aspects in a more flexible and accessible way than conducting studies through video or timing. At the same time, when the data collection is automatic, the "Hawthorne effect" is avoided (which implies the change of attitudes of a person by the mere fact of being observed by another person). Even, these data collected in the StanForD format allow to establish a common base for the comparison of the studies based on them, making comparable the case studies from Austraila (Strandgard et al., 2013), Russia (Gerasimov et al., 2012), Uruguay (Olivera et al., 2016) or South Africa (Brewer et al., 2018), to mention a few examples.

5. Conclusions

In this work, a big data-based approach is proposed to analyze and generate valuable information from the large amount of data stored by forest harvesting teams. In this sense it was possible to verify that the decision trees and the k-means algorithm are suitable methods for this type of problems. On the other hand, it was possible to study the relationship of different factors with the productivity of the harvest. Of these factors, the DBH stands out as the most significant, but the species and the operator also showed significant influences.

In turn, the advantages of DT were used to study particular situations defined when setting values for each variable. This makes possible to identify important prediction rules for harvest planners who must face specific and concrete situations. In many of those cases, the predictive capacity was very good. However, there were other cases in which no. This may be due to other factors that influence the analysis that are not contemplated in the formation of the data set.

Therefore, as a future line of research, it is proposed to evaluate other forest production environments in order to analyze if the predictive capacity of these methods can be improved.

References

Ahlemeyer-Stubbe, A., Coleman, S. (2014). *A practical guide to data mining for business and industry*, John Wiley & Sons.

Berry, M. J., & Linoff, G. (1997). *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc..

Brewer, J., Talbot, B., Belbo, H., Ackerman, P., & Ackerman, S. (2018). A comparison of two methods of data collection for modelling productivity of harvesters: manual time study and follow-up study using on-board-computer stem records. *Annals of Forest Research*, *61*(1), 109-124.

Broz, D., Durand, G., Rossit, D., Tohmé, F., y Frutos, M. (2016). Strategic planning in a forest supply chain: a multigoal and multiproduct approach. *Canadian Journal of Forest Research*, 47(999), 297-307.

Broz, D., Milanesi, G., Rossit, D. A., Rossit, D. G., & Tohmé, F. (2017a). Forest management decision making based on a real options approach: An application to a case in northeastern Argentina. *Forestry Studies*, *67*(1), 97-108.

Broz, D., Olivera, A., Viana Céspedes, V. & Rossit, D.A. (2017b). *Review of Data mining applications in forestry sector*. I International Conference on Agro BigData and Decision Support Systems in Agriculture, 143-145. Montevideo, Uruguay.

Broz, D., Rossit D. A., Rossit, D. G., & Cavallin, C. (2018). The Argentinian forest sector: opportunities and challenges in supply chain management. *Uncertain Supply Chain Management*, 6(4), 375-392.

Chi, M., Plaza, A., Benediktsson, J. A., Sun, Z., Shen, J., & Zhu, Y. (2016). Big data for remote sensing: Challenges and opportunities. *Proceedings of the IEEE*, *104*(11), 2207-2219.

Eriksson, M., & Lindroos, O. (2014). Productivity of harvesters and forwarders in CTL operations in northern Sweden based on large follow-up datasets. *International Journal of Forest Engineering*, *25*(3), 179-200.

Gerasimov, Y., Senkin, V., & Väätäinen, K. (2012a). Productivity of single-grip harvesters in clear-cutting operations in the northern European part of Russia. *European Journal of Forest Research*, *131*(3), 647-654.

Gerasimov, Y., Seliverstov, A., & Syunev, V. (2012b). Industrial round-wood damage and operational efficiency losses associated with the maintenance of a single-grip harvester head model: a case study in Russia. *Forests*, *3*(4), 864-880.

Häggström, C., & Lindroos, O. (2016). Human, technology, organization and environment–a human factors perspective on performance in forest harvesting. *International Journal of Forest Engineering*, *27*(2), 67-78.

Hiesl, P., & Benjamin, J. G. (2013). Applicability of international harvesting equipment productivity studies in Maine, USA: A literature review. *Forests*, *4*(4), 898-921.

Hlásny, T., Trombik, J., Bošeľa, M., Merganič, J., Marušák, R., Šebeň, V., Štěpánek, P., Kubišta, J., Trnka, M. (2017). Climatic drivers of forest productivity in Central Europe. *Agricultural and Forest Meteorology*, 234, 258-273.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), 651-666.

Kamilaris, A., Kartakoullis, A., & Prenafeta-Boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, *143*, 23-37.

Kohavi, R. (1996, August). Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In *KDD* (Vol. 96, pp. 202-207).

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6(70).

Li, L., Hao, T., Chi, T. (2017). Evaluation on China's forestry resources efficiency based on big data. *Journal of Cleaner Production*, 142, 513-523.

Liao, S. H., Chu, P. H., Hsiao, P. Y. (2012). Data mining techniques and applications–A decade review from

2000 to 2011. *Expert systems with applications*, 39(12), 11303-11311.

Lokers, R., Knapen, R., Janssen, S., van Randen, Y., Jansen, J. (2016). Analysis of Big Data technologies for use in agro-environmental science. *Environmental Modelling & Software*, 84, 494-504.

Mac Donagh, P. M., Hildt, E., Friedl, R. A., Zaderenko, C., Alegranza, D. A. (2013). Influencia de la intensidad de raleos en la performance de un harvester de ruedas en el noreste Argentino. *Floresta*, 43(4), 653-662.

Mohammadi, J., Shataee, S., Babanezhad, M. (2011). Estimation of forest stand volume, tree density and biodiversity using Landsat ETM+ Data, comparison of linear and regression tree analyses. *Procedia Environmental Sciences*, 7, 299-304.

Olivera Farias, A. (2016). *Exploring opportunities for the integration of GNSS with forest harvester data to improve forest management*. PhD thesis, University of Canterbury, New Zealand.

Olivera, A., Visser, R., Acuna, M., & Morgenroth, J. (2016). Automatic GNSS-enabled harvester data collection as a tool to evaluate factors affecting harvester productivity in a Eucalyptus spp. harvesting operation in Uruguay. *International journal of forest engineering*, *27*(1), 15-28.

Özbayoğlu, A. M., Bozer, R. (2012). Estimation of the burned area in forest fires using computational intelligence techniques. *Procedia Computer Science*, 12, 282-287.

Purfürst, F. T. (2010). Learning curves of harvester operators. *Croatian Journal of Forest Engineering: Journal for Theory and Application of Forestry Engineering*, *31*(2), 89-97.

Rönnqvist, M., D'Amours, S., Weintraub, A., Jofre, A., Gunn, E., Haight, R. G., ... & Romero, C. (2015). Operations Research challenges in forestry: 33 open problems. *Annals of Operations Research*, 232(1), 11-40.

Rossit, D.A., Olivera, A., Viana Céspedes, V. & Broz, D.(2017). *Application of data mining to forest operations planning*. I International Conference on Agro BigData and Decision Support Systems in Agriculture, 147-149. Montevideo, Uruguay.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660-674.

Sanquetta, C. R., Wojciechowski, J., Dalla Corte, A. P., Rodrigues, A. L., Maas, G. C. B. (2013). On the use of data mining for estimating carbon storage in the trees. *Carbon balance and management*, 8(1), 1-9.

Skogforsk (2007). Standard for Forest Data and communications. In:
https://www.skogforsk.se/contentassets/b063db555a664ff8b515ce121f4a42d1/stanford_maindoc_070327.pdf

Skogforsk. (2017). StanForD. Retrieved October 19, 2017, from
https://www.skogforsk.se/english/projects/stanford/

Strandgard, M., Walsh, D., & Acuna, M. (2013). Estimating harvester productivity in Pinus radiata plantations using StanForD stem files. *Scandinavian journal of forest research*, *28*(1), 73-80.

Strandgard, M., Mitchell, R., & Acuna, M. (2017). Time consumption and productivity of a forwarder operating on a slope in a cut-to-length harvest system in a Pinus radiata D. Don pine plantation. *Journal of Forest Science*, *63*(7), 324-330.

Tolan, A., & Visser, R. (2015). The effect of the number of log sorts on mechanized log processing productivity and value recovery. *International Journal of Forest Engineering*, *26*(1), 36-47.

Traub, B., Meile, R., Speich, S., & Rösler, E. (2017). The data storage and analysis system of the Swiss National Forest Inventory. *Computers and Electronics in Agriculture*, *132*, 97-107.

Uusitalo, J. (2010). Timber Harvesting. In Introduction to forest operations and technology (pp. 66–124). Hameenlinna: JVP Forest Systems Oy.

Viana Céspedes, V. (2018). Optimización en la planificación de servicios de cosecha forestal. Ms. Thesis.

Vlahos, G. E., Ferratt, T. W., Knoepfle, G. (2004). The use of computer-based information systems by German managers to support decision making. *Information & Management*, 41(6), 763-779.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.

Appendix

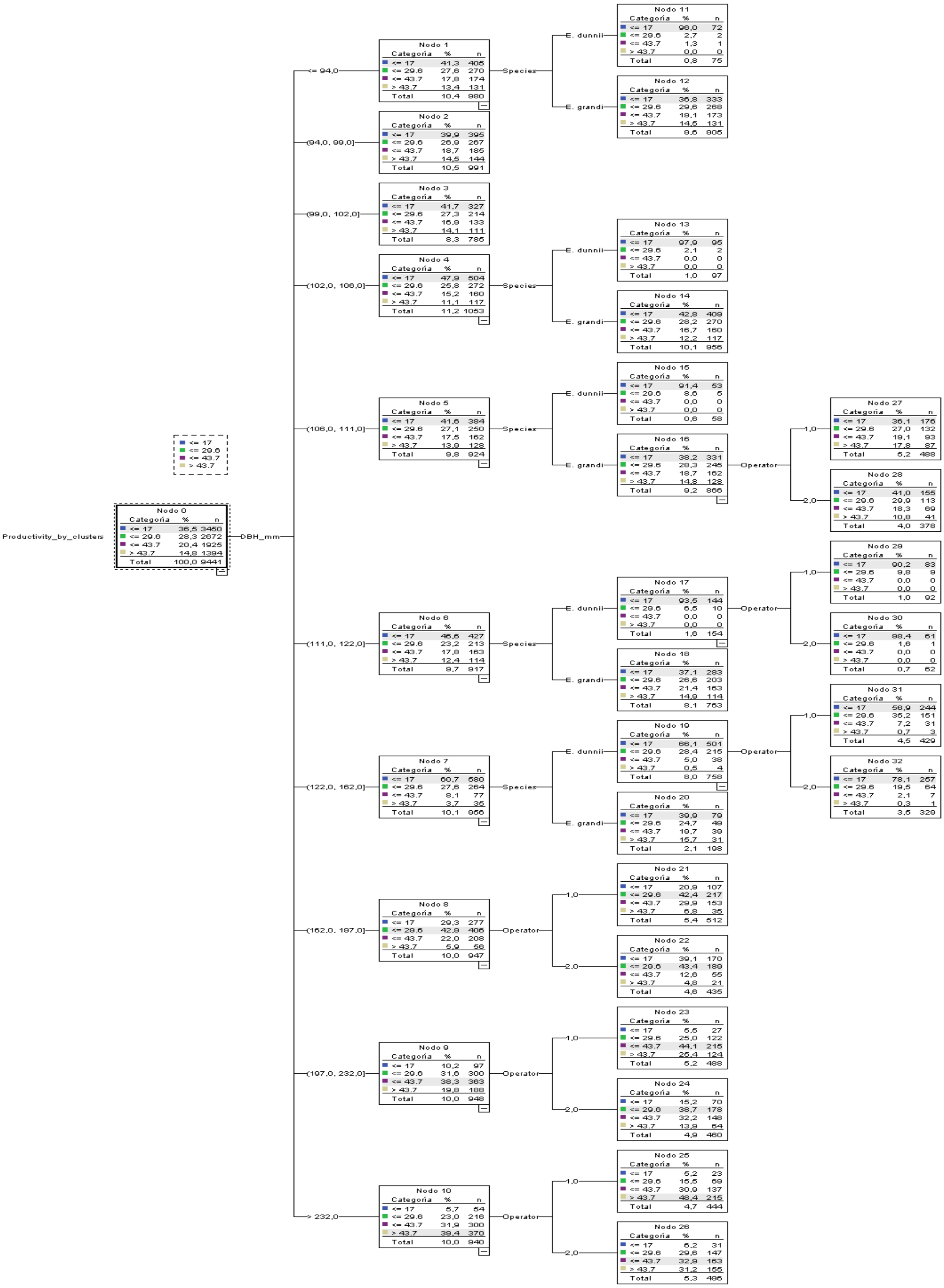

Figure 5. Productivity Forecast model considering DBH, operator, shift and specie as independent variables. The productivity classes are based on cluster analysis.

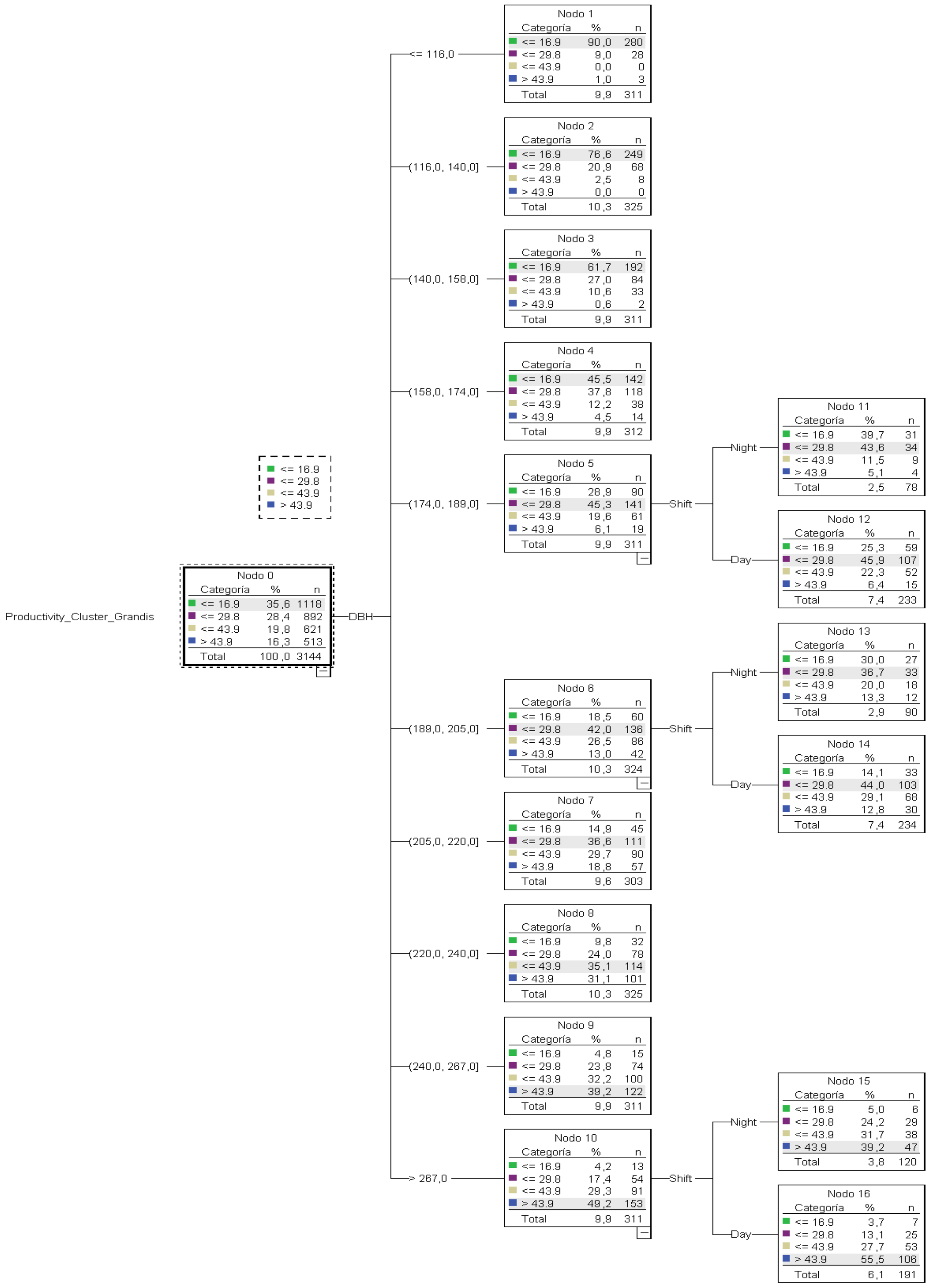

Figure 7. Operator 1 productivity Forecast model fory *E. grandis* specie considering DBH and shift as independent variables. The productivity classes are based on cluster analysis for *E. grandis* registers.

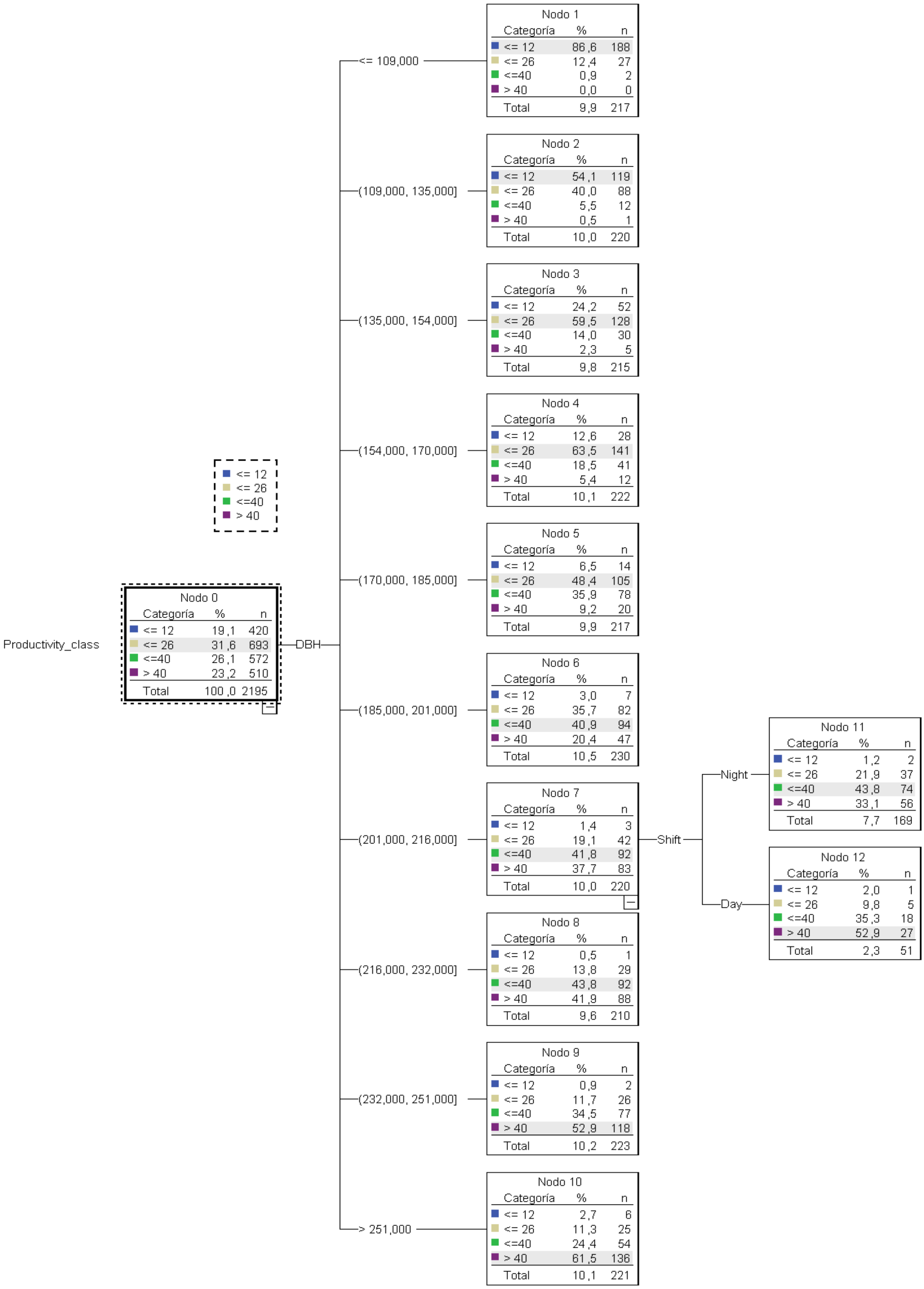

Figure 8. Operator 1 productivity Forecast model for *E. dunnii* specie considering DBH and shift as independent variables. The productivity classes are based on equal intervals' criterion.

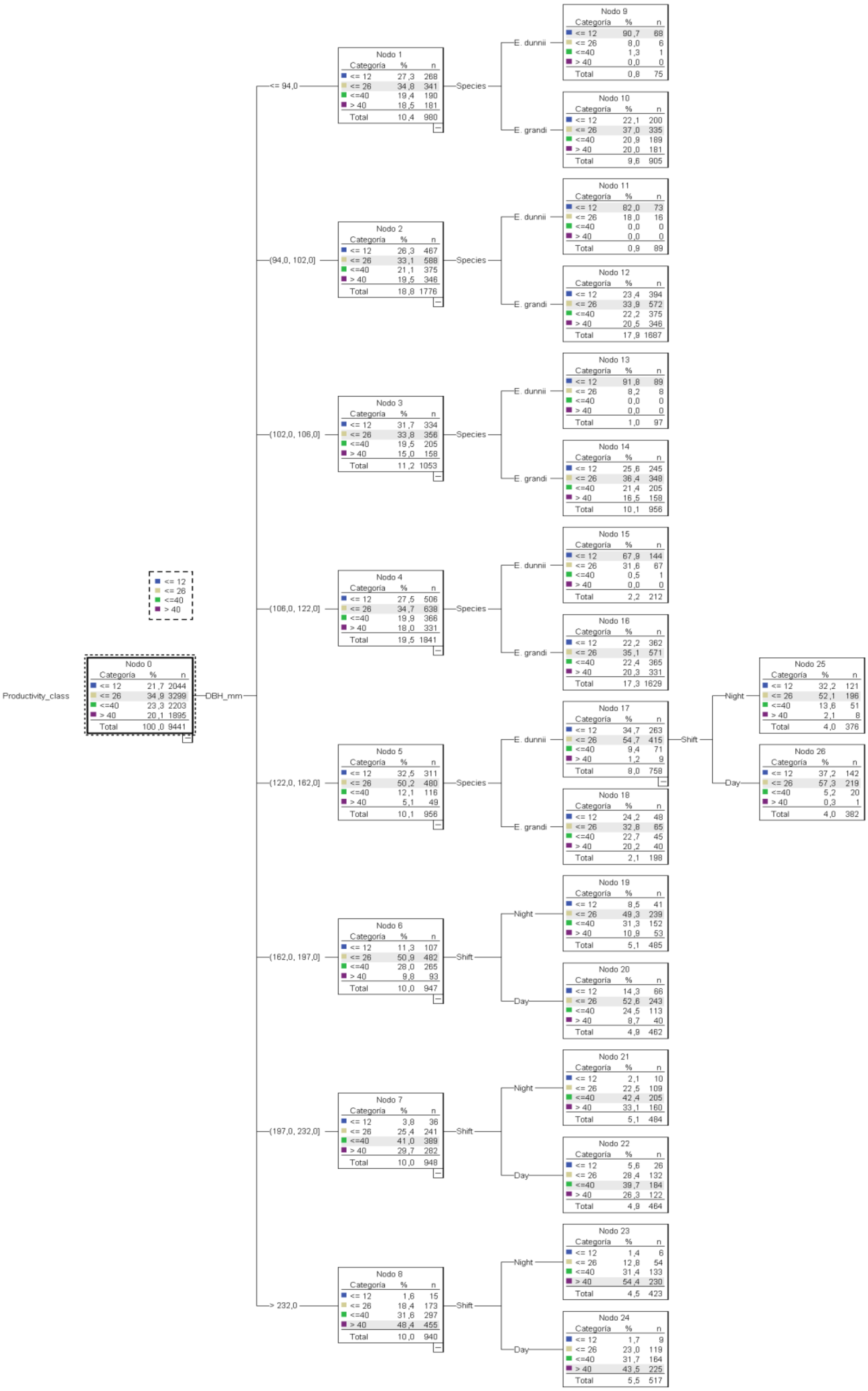

Figure 10. Operator 2 productivity Forecast model considering DBH, shift and specie as independent variables. The productivity classes are based on equal intervals' criterion.

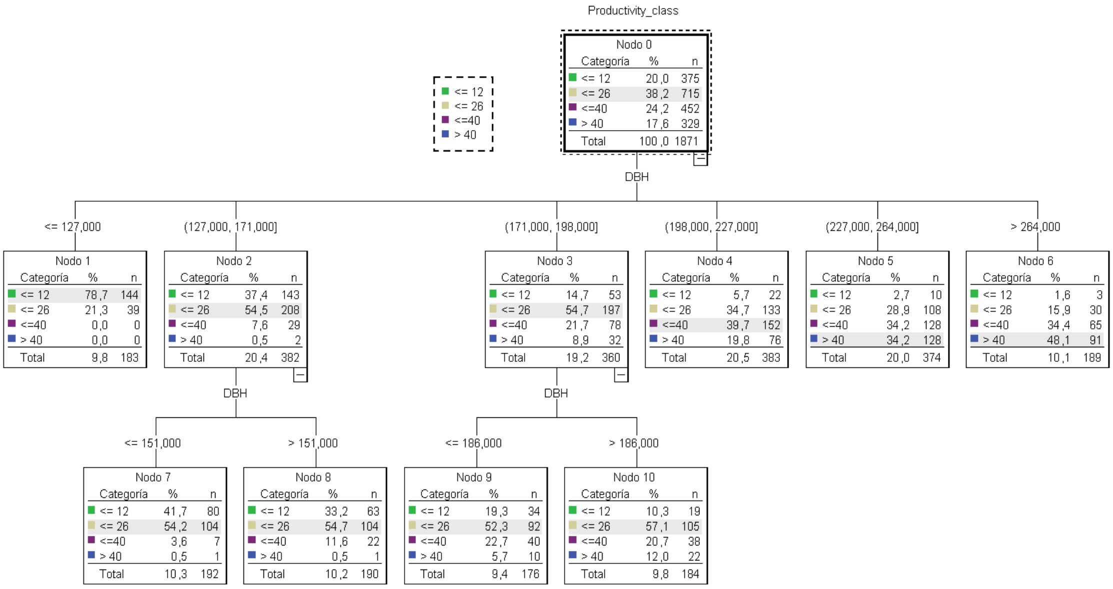

Figure 12. Operator 2 productivity Forecast model for *E. dunnii* specie considering DBH and shift as independent variables. The productivity classes are based on equal interval's criterion.

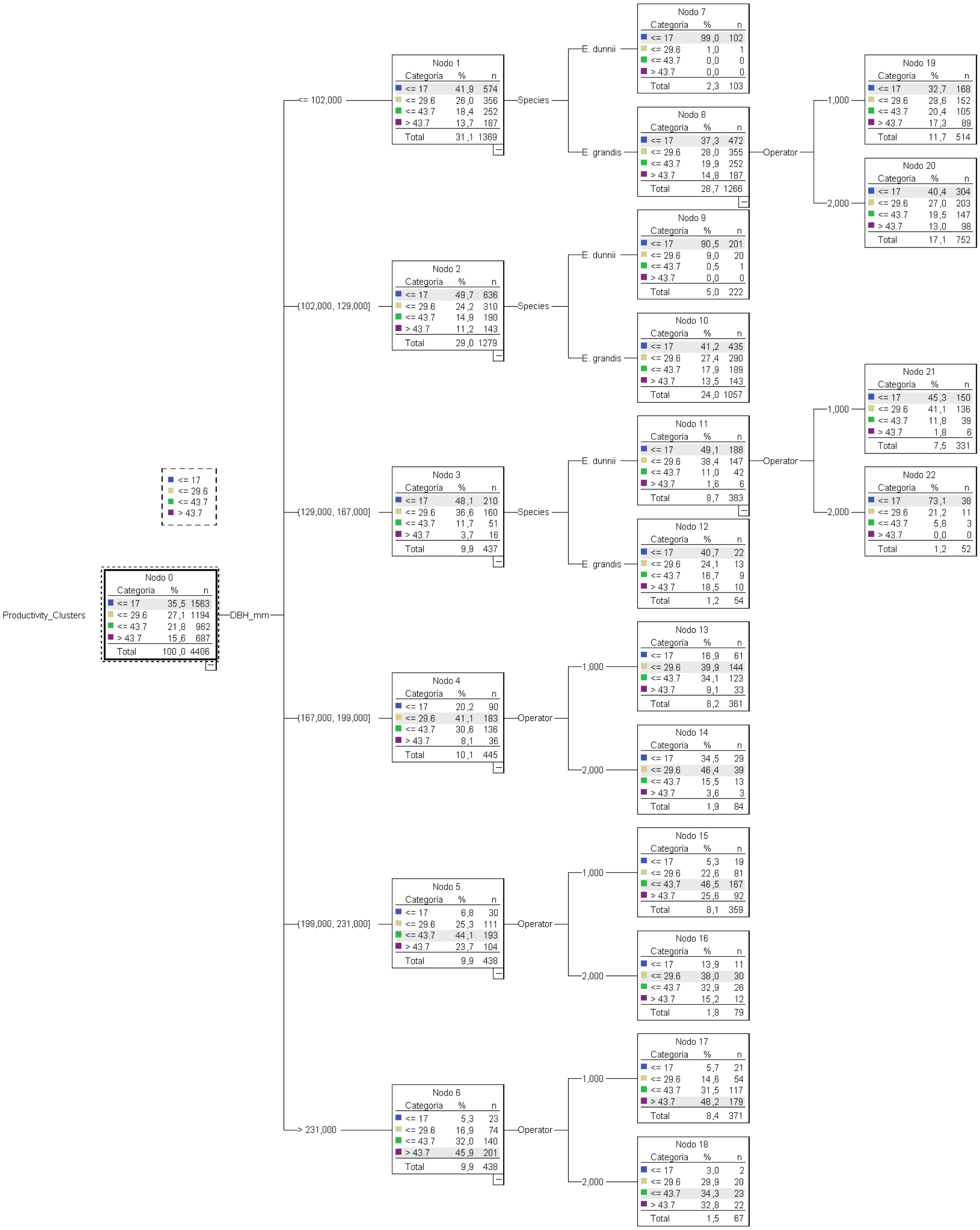

Figure 13. Night Shift Productivity Forecast model considering DBH, specie and operator as independent variables. The productivity classes are based on cluster analysis for global registers.

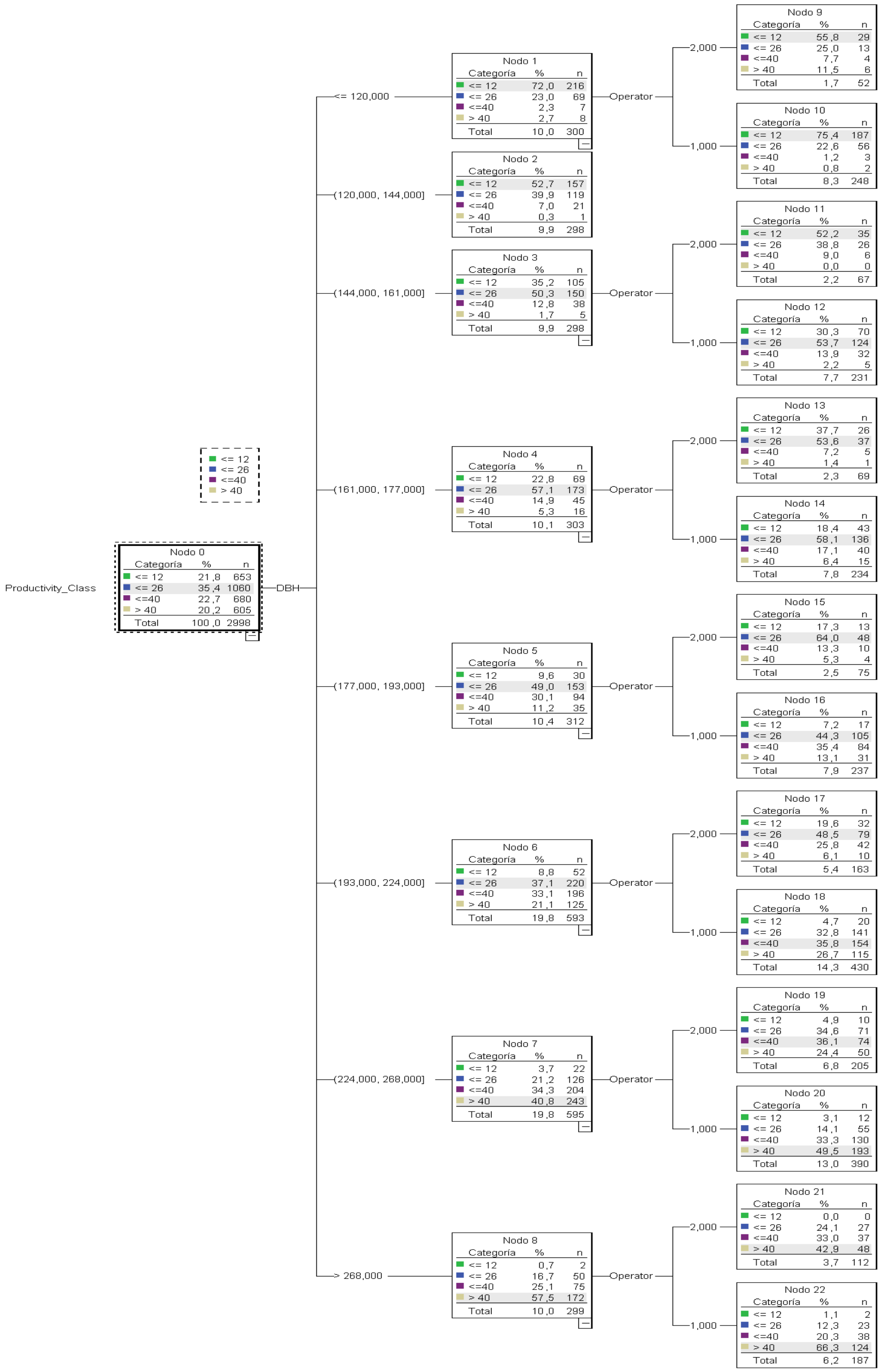

Figure 17. Day Shift Productivity Forecast model for *E. grandis* specie considering DBH and operator as independent variables. The productivity classes are based on equal interval's criterion.