Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

# Application of the J48 Decision Tree Algorithm in the Mass Valuation of Real Estate

**EDWIN ROBERT PEREZ** ( edwinperezc@gmail.com )
Universidad Distrital Francisco Jose de Caldas https://orcid.org/0000-0002-3916-7305

**NELSON OBREGON**
Pontificia Universidad Javeriana

**Adriana Albancando**
Universidad Distrital Francisco Jose de Caldas

---

**Research**

**Keywords:** Machine learning, decision trees, Cross validation, Percentage Split, J48 algorithm, Linear Regression, BIG DATA, mass valuation, real estate

**Posted Date:** February 15th, 2020

**DOI:** https://doi.org/10.21203/rs.2.23397/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

# Application of the J48 Decision Tree Algorithm in the Mass Valuation of Real Estate

**Edwin Robert Perez Carvajal [1,2], Nelson Obregon Neira [3], Adriana Albancando Robles [4]**

[1] Engineering Faculty, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia; edwinperezc@gmail.com
[2] Engineering Faculty, Pontificia Universidad Javeriana, Bogotá D.C., Colombia; erperezc@correo.udistrital.edu.co
[3] Engineering Faculty, Pontificia Universidad Javeriana, Bogotá D.C., Colombia; nobregon@javeriana.edu.co
[4] Engineering Faculty, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia. (engineering student); aalbancandor@correo.udistrital.edu.co

**Abstract:** This article presents the results of the research work of the GIGA research group of the Distrital University in Bogota Colombia in the framework of the advanced doctoral research in the engineering doctorate of the Javeriana University and shows the development followed and the corresponding analysis determine the effectiveness of the M5P algorithm within the decision tree approach for the management of BIG DATA and Machine Learning for the determination of massive property valuation. The analysis was carried out with data provided by the Special Administrative Unit of the Distrital Cadaster (UAECD) and the Technical Cadastral Observatory corresponding to three sectors of the city of Bogotá D.C. The results of the J48-based model were compared by means of a statistical analysis with the traditionally used linear regression method, obtaining satisfactory results with errors well below the linear regressions in the training, validation and forecast stages.

**Keywords:** Machine learning, decision trees, Cross validation, Percentage Split, J48 algorithm, Linear Regression, BIG DATA, mass valuation, real estate.

## 1. Introduction

The massive valuation of real estate is really important for the execution of both public and private works that generate high impact on the transformation of the territory. Traditionally for the development of such projects, methods have been used that try to approximate the real values (commercial values) that serve as the basis for decision making. Traditionally, RL Linear Regression models have been used in Colombia, which have been considered reliable although other approaches such as those related to Artificial Intelligence defined as the science of building machines that do things that, if humans did, would require intelligence (Cazorla 2003).

With the search for efficiency, it is currently common to see that the main concerns of contemporary models focus on simplifying processes and improving results and accuracies, aspects in which artificial intelligence is neat, because the training of machines through different elements and data for a specific process show great flexibility of these tools to adjust to new data and circumstances, showing greater versatility in "learning" to obtain greater precision and better classification processes (Murphy 1997).

. Within machine learning approaches there are multiple algorithms with different characteristics, strengths and weaknesses. One of these widely accepted in the academic community is known as decision trees that generally use "yes - then" rules and generally used to make classifications based on data provided (Mitchell 1997). This article was developed within the GIGA group of the Francisco José de Caldas Distrital University in Bogotá Colombia with the collaboration of the student of Cadastral Engineering and Geodesta Adriana Albancando Robles in the framework of the doctoral research of the engineer Edwin Perez Carvajal.

## 2. Materials and Methods

The data used for this article correspond to three sectors of the city of Bogotá D.C in the Republic of Colombia called Zonal Planning Units (UPZ by its acronym in Spanish): Garces Navas (73), Arborizadora (65) and Calandaima (79).

## 2.1 Spatial Framework

### 2.1.1 Garces Navas (UPZ 73)

It is located west of the city and has 557.43 hectares, it shows a wide variety of land uses, such as housing, commercial premises, and mixed uses (housing, commerce, equipment) (SDP 2011) and has cadastral sectors presented in Figure 1.

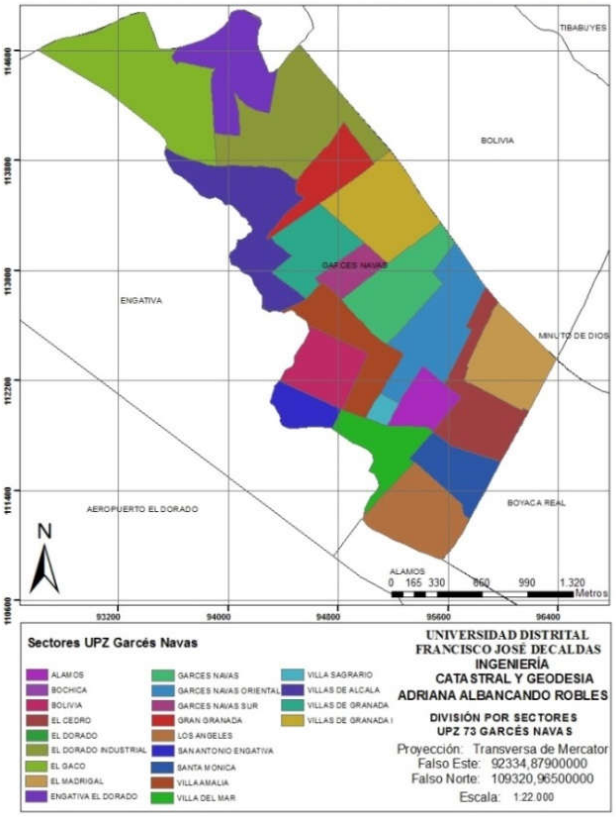



**Figure 1.** Garcés Navas UPZ Zoning

### 2.1.2 Arborizadora (UPZ 65)

The city of Bogotá is located to the south west, has an area of 326.97 hectares and contains the cadastral sectors that are visualized in Figure 2.

The most representative land uses of this UPZ are: housing, commerce, commercial premises, department stores and supermarkets, Industry and area for mixed uses (housing, commerce, equipment) (Alcaldía Mayor de Bogotá D.C 2008)

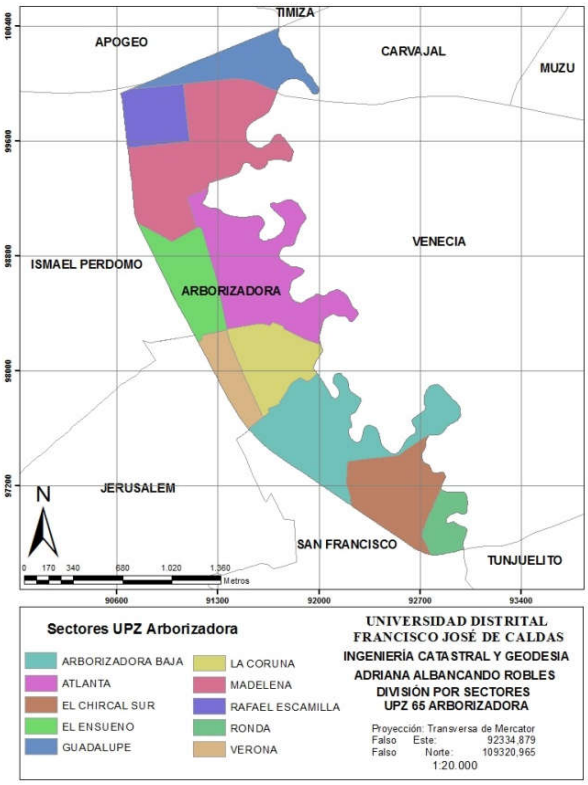



**Figure 2.** Arborizadora UPZ Zoning

### 2.1.3 Calandaima (UPZ 79)

Located to the south west of the city of Bogotá, it has an area of 319 hectares and contains the sectors presented in Figure 3.

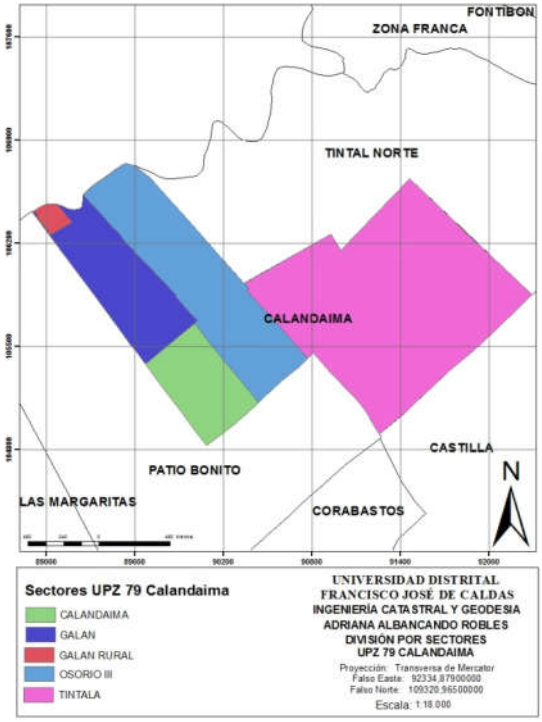



**Figure 3.** Calandaima UPZ Zoning

The UPZ Calandaima is in the development stage for what is considered a multi-activity sector (residential, commercial, industrial and institutional urban developments) (Alcaldía Mayor de Bogotá 1993).

## 2.2 Theoretical Framework

### 2.2.1 Artificial Intelligence

The term Artificial Intelligence (AI) is used to identify a field of computational science and engineering that deals with the creation of artifacts that have the faculties to know, understand and solve problems if following pre-established theories or theoretical models.

In general, AI corresponds to data-guided models and is classified in areas such as: Natural language treatment, Automatic reasoning - Expert systems, Machine or machine learning, Knowledge representation and Artificial and robotic vision (Pino 2001).

### 2.2.2 Machine Learning

With machine learning, it is sought that the device is capable of inferring, automatically and autonomously, a large set of results through the knowledge acquired in training (McCarthy 1958). Figure 4 illustrates the process contained in two stages, the first training, starting with the human component, who create and use the algorithm that will subsequently run the machine, then the sample data for processing with which It will produce the learning and finally the model is obtained. The second stage, the validation stage is made up of four parts: (1) it consists of providing a new set of data by the user, (2) the data supplied is processed by the machine in the model, (3) the Results information, (4) the results are validated to be used, subsequently as a basis for forecasts (Garcia 2012).

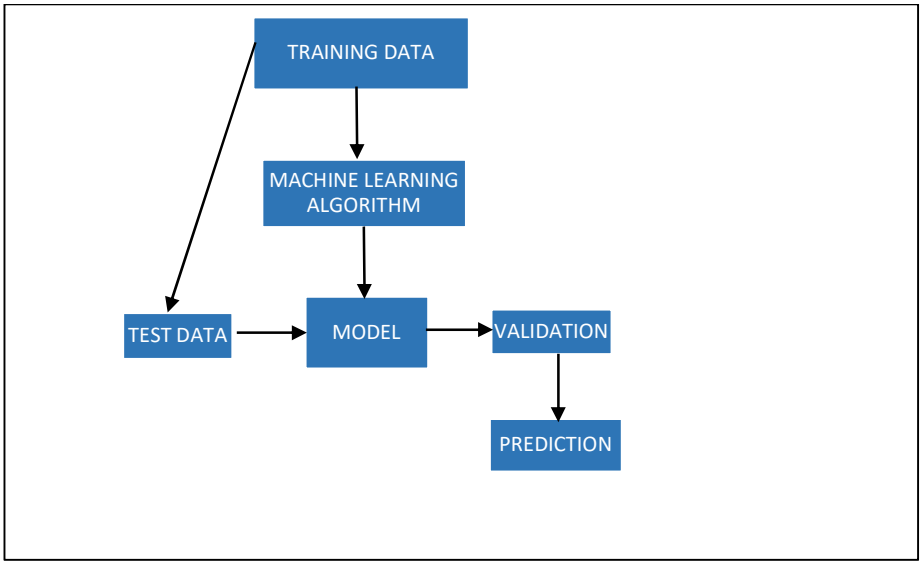



**Figure 4.** Learning Machine Process

### 2.2.3 Decision trees

The decision tree technique corresponds to a classification of inductive inference methods, where general information is derived from particular information (Mitchell 1997).

This technique groups rules in an organized way, in a hierarchical structure, to reach the final solution following the conditions established by each rule from the root to the leaves.

The decision trees are composed of: the root that is the upper part of the tree where the attribute with which the classification begins is located, the branches that are located in the intermediate part of the tree and that allow the root to connect with the nodes , the internal nodes that are located inside the tree, where are the attributes that generate the classification and the final nodes or leaves that are at the ends of the tree and represent the rules of the final classification (Vizcaino 2008).

In Figure 5, the scoring attribute is located at the root of the tree and with it the algorithm starts the classification, in the subsequent branches that arise from there is the rule: "it is less than or equal to 30.5"; if the condition is met, the following attribute (age) is verified, if it is less than or equal to 16.5, rule 1 applies.

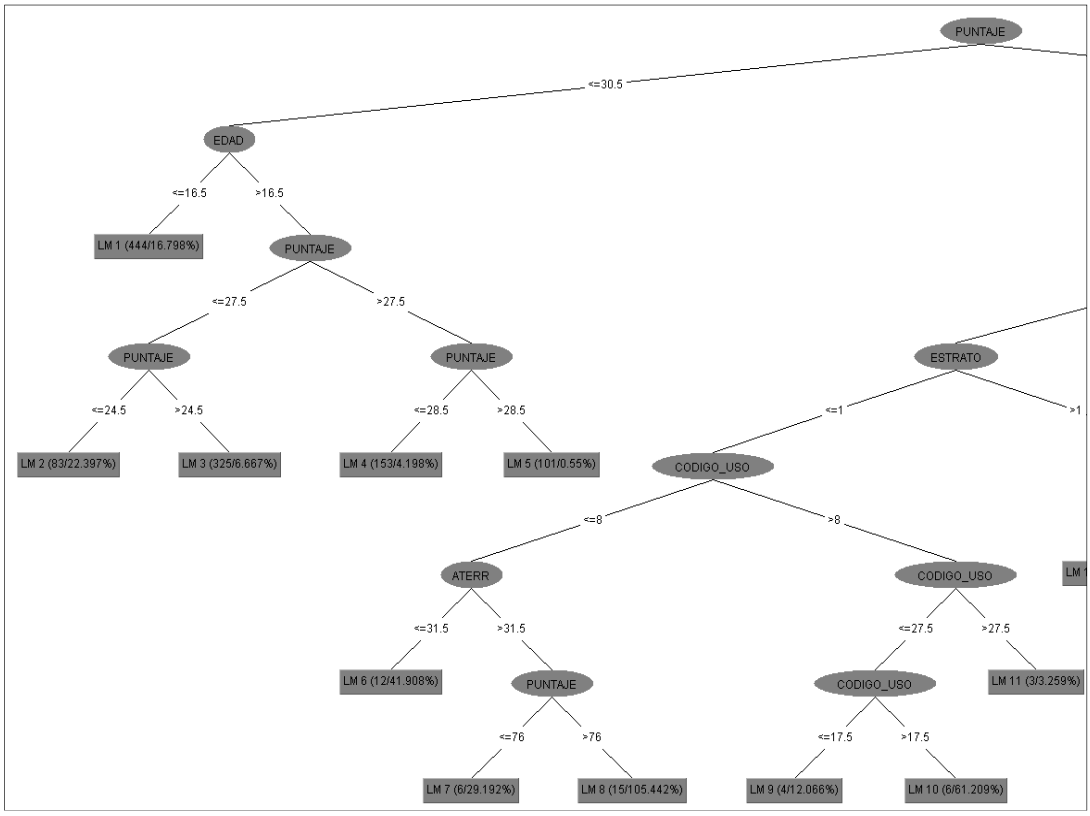



**Figure 5.** Example of decision tree generated in WEKA

### 2.2.4 Information Entropy

The information entropy or Shannon Entropy measures the uncertainty of the information from the data provided as seen in equation 1. (Sancho 2016).

$$Entropy(S) = \ -p_{+}log_{2}p_{+} - p_{-}log_{2}p_{-}, \tag{1}$$

Where:

$p_{+} =$ Average positive examples in *S*.

$p_{-} =$ Average negative examples in *S*.

When the examples can be classified in the same class, the entropy is zero, therefore the absolute certainty, but when the number of negative examples is equal to the positive ones, the uncertainty is total and the entropy is 1, therefore for Variable data sets, the entropy will be between 0 and 1.

Given that an attribute can take different values, entropy is defined by equation 2, where $p_i$ is the proportion of S that belongs to class *i*.

$$Entropy \ \ (S) = \ \sum_{i=1}^{c} -p_i \ \ log_2 p_i, \tag{2}$$

### 2.2.5 Information Gain

In the generation of decision trees, we begin by selecting the attribute that should go to the root of the tree, taking into account that this attribute must correspond to the one that provides the most information according to equation 3 (McCarthy 1958).

$$Gain\ \ (S,A) = Entropy\ \ (S) - \sum_{v\ \in values\ \ (A)} \frac{|S_v|}{|S|} Entropy\ \ (S_v), \tag{3}$$

Where:

$S =$ Example set
$A =$ Set of possible values for attribute A
$S_v =$ Subset of $S$ for which attribute A has a value $v$ (Equation 4)

$$S_v = \{s\ \ \in S\ \ |A_{(S)} = v|\}, \tag{4}$$

### 2.2.6 J48 Algorithm

The J48 algorithm is a variation of the Quinlan C4.5 algorithm in Java that the Weka software has. The C4.5 technique is an algorithm developed by Ross Quinlan as an improvement of the ID3 algorithm, so its developments are similar.

In order to create a decision tree by means of the C4.5 algorithm, first a set of training data is determined, these are divided into subsets that will be evaluated through the gain of information to determine the attribute with the highest gain that is taken as a classification parameter to occupy the root node. To continue with the classification, the algorithm uses two tools: “info” and “gain”. With “gain”, the information that each branch contributes to the process is calculated and with the “gain” tool it calculates the overall improvement generated by the rule. In this way the path and structure of the tree is established, taking as a starting point the results of the previous cycle, calculating the precision of the model according to the totality of the data and obtaining in the output a categorical variable (Vizcaino 2008).

Among the improvements that the J48 algorithm incorporates into the ID3 algorithm can be mentioned:

1. J48 can handle continuous and discrete attributes: To work in the process with continuous attributes, the algorithm divides the values of the attributes among those that are greater, less than or equal to the limit generated.

2. J48 handles the data from the set of examples with incomplete information: All attributes are included even though they do not have the complete information, omitting it only for the calculation of entropy and information gain.

3. J48 eliminates branches that do not provide information to the model: The pruning or Pruning process is implemented in two moments, while the tree grows (pre pruning) or when it is already complete (post pruning). Once the tree is created, the algorithm is returned to find the branches that do not provide enough information in the process (pruning), to replace them with end nodes or leaf nodes. Among the methods used to determine the sub-trees to be pruned are:

a. Cross validation, where training data is reserved (validation set - tuning set) to evaluate the usefulness of sub-trees.

b. Statistical tests, used in training sets to determine information that can be eliminated.

c. Minimum Description Length - MDL, which allows to determine if the hypothesis of the whole tree is more complex than that of the tree resulting from the cut.

J48 avoids data overfitting: Unlike ID3, algorithm C4.5 performs a search of hypotheses or set of decision trees to adjust training data taking into account inductive bias, referred to the Ockham's razor principle, preferring short trees to larger ones, because shorter trees will have more information near the root, generalize better and contain less irrelevant attributes (Mitchell 1997).

### 2.2.7 Validation

For the validation of the experiments presented in this article the following techniques were used:

1. Cross Validation: The data set is arranged in *n* partitions or folds, then a classifier with *n-k* sub-sets is constructed to build both the training data set and the validation data set. The process is repeated until all data have been used for both training and validation, as shown in the example in Figure 6 for a set of 20 data and *n = 5* (Corso 2009).

2. Percentage Split: It allows choosing the percentages of data for training and validation of the model (Garcia 2013).

3. Supplied test set: Two different data sets are organized one for training and the other for model validation (Hernandez 2006).

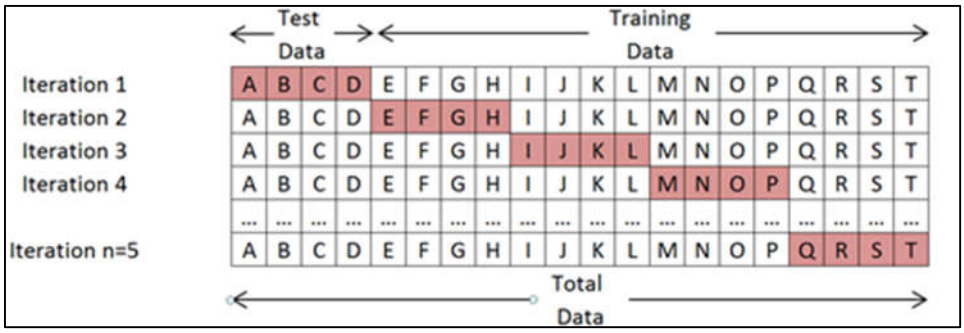



**Figure 6.** Example of cross validation
Source WEKA

### 2.2.8 Error Measurement

To measure the performance and accuracy of the J48 algorithm, the relative error indicators illustrated in equations 5 and 6 were used, because they allow comparisons between different models with different error measures (Mood 1974).

Relative Absolute Error RAE:

$$RAE = \frac{\sum_{i=1}^{N}|\hat{\theta}_i - \theta_i|}{\sum_{i=1}^{N}|\bar{\theta}_i - \theta_i|}, \tag{5}$$

Root Relative Squared Error RRSE:

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{\sum_{i=1}^{N}(\bar{\theta}_i - \theta_i)^2}}, \tag{6}$$

### 2.2.9 WEKA

WEKA (Waikato Environment for Knowledge Analysis) is a free software from the University of Waikato in New Zealand. In the Explorer application you can have tools such as Classify where regression and classification algorithms are available, Cluster that allows you to find different data grouping techniques (Morate 2000).

Three data sets were organized for properties under the Horizontal Property (PH) regulation and three data sets for properties outside the Horizontal Property (No PH) regulation as can be seen in Figure 7. For the final tests, a set with 20 randomly selected data was arranged. The training and validation data set corresponds to 95% of the total data while for the forecast data set 5% was retained.

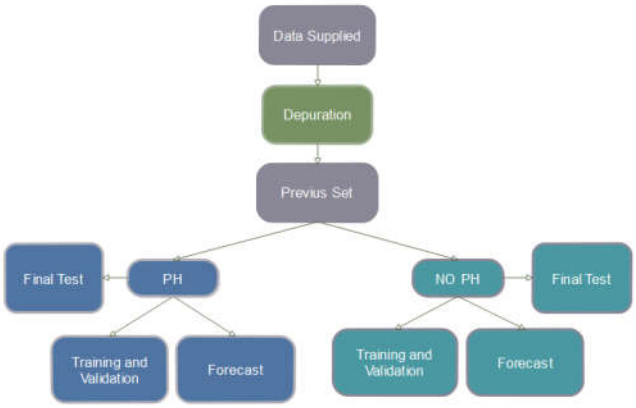



**Figure 7.** Organization of data for experimentation.

The same protocol for data enlistment was followed for the 3 UPZ finally obtaining a configuration like the one shown in Table 1.

**Table 1.** Number of data for experimentation by UPZ

| Experiment | UPZ | | |
|---|---|---|---|
| | **Garces N.** | **Arborizadora** | **Calandaima** |
| PH Training and Validation | 14456 | 9438 | 28824 |
| PH Forecast | 761 | 497 | 1517 |
| NO PH Training and Validation | 17875 | 6423 | 2636 |
| NO PH Forecast | 941 | 338 | 139 |

#### 2.2.10 Mass Valuation

It consists of the activities and processes necessary to carry out real estate valuations in bulk based on the inference and extrapolation of information from individual properties, based on individual valuations.

In order to comply with this process in Colombia, four stages are contemplated: Property identification, determination of homogeneous physical and geo-economic zones, determination of unit values for different types of buildings and finally the liquidation of appraisals (IGAC 2008).

## 3. Results

### 3.1 Data Organization

For the experiment, the data provided by the UAECD was processed, generating 6 data sets for each UPZ. As illustrated in Figure 7, the data were divided into properties PH regulation and into properties NPH.

Subsequently, the attributes were renamed to ensure compliance with the requirements of the algorithm and the WEKA software. The attributes finally used in the experiments are shown in Table 2.

**Table 2.** Attributes for the Experiments

| UPZ | Garcés N. | Arborizadora | Calandaima |
|---|---|---|---|
| PH | Zone<br>Age<br>Qualification<br>Stratum<br>Economic Activity<br>Treatment<br>Built area<br>Building value (m²) | Zone<br>Use<br>Age<br>Qualification<br>Stratum<br>Economic activity<br>Treatment<br>Built area<br>Building value m² | Zone<br>Use<br>Floors<br>Age<br>Qualification<br>Stratum<br>Economic activity<br>Treatment<br>Built area<br>Building value m² |
| N PH | Zone<br>Floors<br>Age<br>Qualification<br>Stratum<br>Economic Activity<br>Treatment<br>Land area<br>Land Value m²<br>Building area<br>Building value m² | Zone<br>Use<br>Age<br>Qualification<br>Stratum<br>Economic activity<br>Treatment<br>Built area<br>Building value m² | Zone<br>Use<br>Floors<br>Age<br>Qualification<br>Stratum<br>Economic activity<br>Treatment<br>Land area<br>Land Value m²<br>Built area<br>Building value m² |

### 3.2 Training and Validation Processes

Table 3 shows the different parameters and indicators analyzed in each experiment. Figure 8 illustrates how the experiment was designed; the data was run in the WEKA software, with the J48 algorithm. Once the final decision tree was generated, it executed its validation by means of the Cross validation techniques (for 10 and 20 repetitions) and Percentage Split (for division percentages of 66 - 33 and 80 - 20).

**Table 3.** Parameters and Validation indicators for J48

| | Obtained Parameters |
|---|---|
| **Algorithm J48** | Instances<br>Attribute Amount<br>Instances classified correctly<br>Instances classified incorrectly<br>Kappa statistic<br>Mean of the absolute error<br>Root of the mean square error<br>Relative absolute error<br>Root of relative quadratic error<br>Unclassified instances |

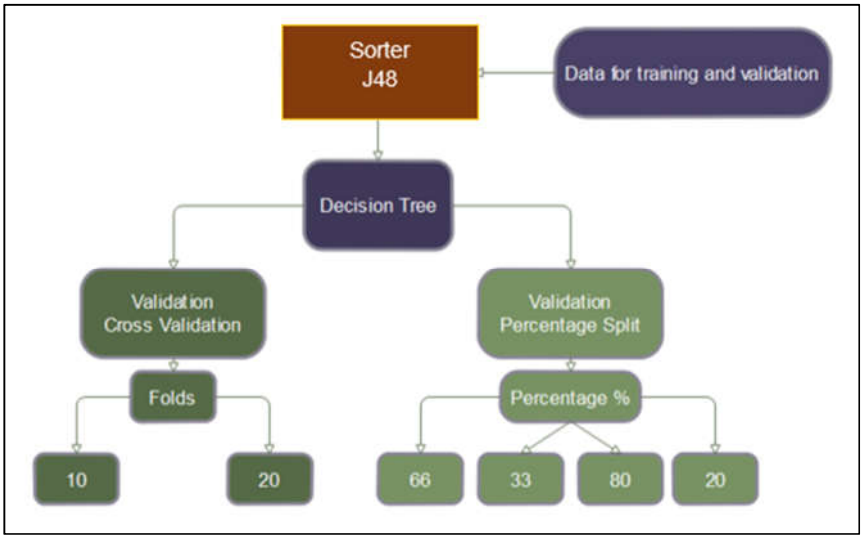



**Figure 8.** Training and Validation process

### 3.3 Forecast Stage

To evaluate the forecasting capacity of the J48 algorithm, the Supplied Test Set method was used, which gives the possibility of validating the model with data different from those used in the training of the decision tree, for which the process shown in Figure 9.

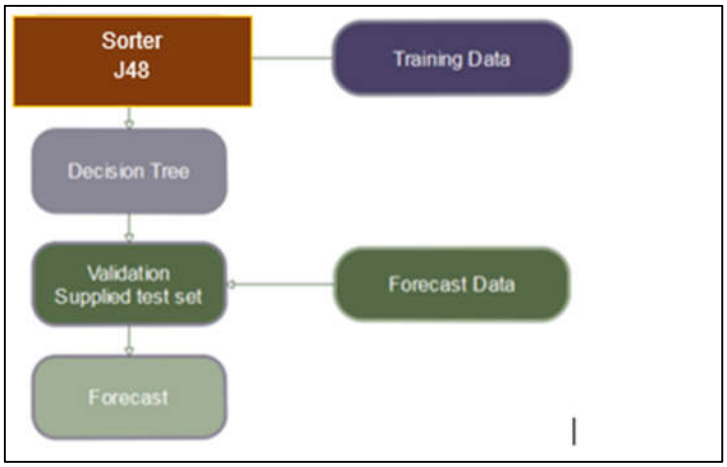



**Figure 9.** Forecast process

### 3.4 Linear Regression (LR)

In order to analyze the accuracy of the results obtained with J48, a traditional Linear Regression model was used, fed with the same validation and forecast training data and procedures: Cross Validation, Percentage Split and Supplied Test Set. The list of parameters used is presented in Table 4.

**Table 4**. Parameters and validation indicators for Linear Regression

| | **Obtained Parameters** |
|---|---|
| | Instances |
| | Attribute Amount |
| **Linear** | RAE |
| **Regression** | RSSE |
| | Correlation coefficient |
| | Mean of the absolute error MAE ($) |
| | Root of the mean square error RMSE ($) |

When comparing the models by analyzing the resulting indicators and parameters, such as: the correctly classified instances, the kappa statistics and the mean of the absolute error for the case of method J48 and the correlation coefficient, the mean of the absolute error and the root of the mean square error for linear regression, allowed to determine the "selections 1 and 2" that represent the best models obtained for each algorithm.

### 3.5 Analysis of results

Once the different results of each experiment were obtained, their comparison and evaluation were tabulated, this process is shown in Figure 10. As can be seen, 3 models for J48 and 3 models for lR corresponding to each validation process were selected: Cross validation, Percentage Split and Forecast.

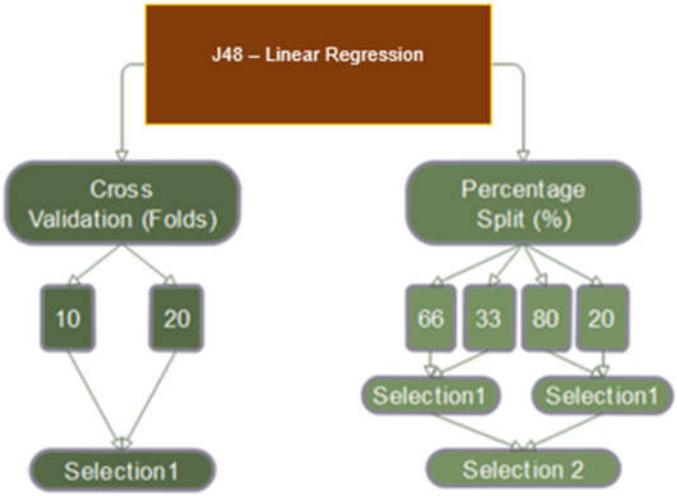



**Figure 10.** Model selection for J48 and for LR.

As an example, Tables 5 and 6 allow comparing the results obtained for selections 1 and 2 for both PH and NPH of the UPZ Garcés Navas. The results of the UPZ Arborizadora and Calandaima were also organized.

As a final step, the relative absolute errors and the roots of the quadratic errors of the models obtained with J48 and LR were compared. The results are presented in tables 7, 8, 9, 10, 11 and 12.

**Table 5**. Results of experiments for UPZ Garcés Navas (PH)

| | TRAINING - VALIDATION | | | | FORECAST | | TRAINING - VALIDATION | | FORECAST |
|---|---|---|---|---|---|---|---|---|---|
| **EXPERIMENT** | J48_CROSS_20 | | J48_SPLIT_80 | | J48_SUPPLIED | | LR_CROSS_20 | LR_SPLIT_80 | LR_SUPPLIED |
| **REGULATORY** | PH | | PH | | PH | | PH | PH | PH |
| **QUANTITY OF DATA** | 14456 | | 14456 | | 761 | | 14456 | 14456 | 761 |
| **SORTER** | J48 | | J48 | | J48 | | LR | LR | LR |
| **TEST OPTIONS** | | | **P. SPLIT** | | **SUPPLIED TEST SET** | | | **P. SPLIT** | **SUPLIED TEST SET** |
| **TEST** | FOLDS | 20 | % | FOLDS | % | 14456 | 761 | | |
| **N. OF SHEETS** | 631 | | 631 | | 631 | | | | |
| **TREE SIZE** | 713 | | 713 | | | | | | |
| **INSTANCES** | 14456 | | 14456 | 2891 | 14456 | | 14456 | 761 | |
| **ATTRIBUTES** | 8 | | 8 | | 8 | | 8 | 8 | 8 |
| **CORRECT C.** | 11990 | 82,9413% | 2406 | | | | | | |
| **INCORRECT C.** | 2466 | 17,0587% | 485 | | | | | | |
| **KAPPA ST.** | 0,8022 | | 0,8051 | | 0,8152 | | | | |
| **MAE** | 0,0161 | | 0,016 | | 0,0152 | | | | |
| **RMSE** | 0,0913 | | 0,0904 | | 0,0875 | | | | |
| **RAE** | 27,0197% | | 26,8669% | | 25,543% | | 79,8469% | 81,515% | 86,0071% |
| **RRSE** | 52,8943% | | 52,4048% | | 50,6862% | | 78,3007% | 79,0384% | 81,1613% |
| **UNCLASSIFIED I.** | | | | | | | | | |
| **NUMBER OF RULES** | | | | | | | | | |
| **CORRELATION C.** | | | | | | | 0,622 | 0,6146 | 0,6002 |
| **MAE ($)** | | | | | | | 221802,7789 | 218043,9783 | 226823,3249 |
| **RMSE ($)** | | | | | | | 459841,9517 | 427704,5767 | 439934,2052 |
| **SELECTION 1** | X | | X | | | | X | X | |
| **SELECTION 2** | | | X | | | | | X | |
| **FINAL SELECTION** | | | | | | | | | |

**Table 6.** Results of the UPZ Garcés Navas (NPH) Experiments

| | TRAINING - VALIDATION | | | | FORECAST | | TRAINING - VALIDATION | | FORECAST |
|---|---|---|---|---|---|---|---|---|---|
| **EXPERIMENT** | J48_CROSS_20 | | J48_SPLIT_80 | | J48_SUPPLIED | | RL_CROSS_20 | RL_SPLIT_66 | RL_SUPPLIED |
| **REGULATORY** | NO_PH | | NO_PH | | NO_PH | | NO_PH | NO_PH | NO_PH |
| **QUANTITY OF DATA** | 17875 | | 17875 | | 17875 | | 941 | 17875 | 17875 |
| **SORTER** | J48 | | J48 | | J48 | | REGRESION LINEAL | REGRESION LINEAL | REGRESION LINEAL |
| **TEST OPTIONS** | **CROSS VALIDATION** | | **PERCENTAGE SPLIT** | | **SUPPLIED TEST SET** | | **CROSS VALIDATION** | **PERCENTAGE SPLIT** | **SUPPLIED TEST SET** |
| **TEST** | FOLDS | 20 | % | FOLDS | 20 | % | FOLDS | 20 | % |
| **N. OF SHEETS** | 1775 | | 1775 | | 1775 | | | | |
| **TREE SIZE** | 1904 | | 1904 | | 1904 | | | | |
| **INSTANCES** | 17875 | | 17875 | 3575 | 17875 | | 17875 | 6077 | 17875 |
| **ATTRIBUTES** | 11 | | 11 | | 11 | | 11 | 11 | 11 |
| **CORRECT C.** | 15166 | 84,844 % | 3035 | | | | 15166 | 84,8448% | 3035 |
| **INCORRECT C.** | 2709 | 15,155 % | 540 | | | | 2709 | 15,1552% | 540 |
| **KAPPA ST.** | 0,7457 | | 0,7483 | | 0,7577 | | | | |
| **MAE** | 0,0415 | | 0,0418 | | 0,0411 | | | | |
| **RMSE** | 0,1512 | | 0,1514 | | 0,1488 | | | | |
| **RAE** | 33,9418% | | 34,0816% | | 33,9441% | | 51,5979% | 51,9574% | 52,2463% |
| **RRSE** | 61,1518% | | 61,0425% | | 60,7461% | | 62,317% | 62,8131% | 62,6413% |
| **UNCLASSIFIED I.** | | | | | | | | | |
| **NUMBER OF RULES** | | | | | | | | | |
| **CORRELATION C.** | | | | | | | 0,7821 | 0,7782 | 0,7795 |
| **MAE ($)** | | | | | | | 40755,3571 | 40833,5809 | 40663,8247 |
| **RMSE ($)** | | | | | | | 63687,9119 | 64200,8245 | 62540,4607 |
| **SELECTION 1** | X | | X | | | | X | X | |
| **SELECTION 2** | | | X | | | | X | | |
| **FINAL SELECTION** | | | | | | | | | |

**Table 7.** Error Comparison Garcés Navas PH (J48 vs. LR)

| EXPERIMENT | RAE (%) | RRSE (%) |
|---|---|---|
| TRA- VAL (J48 CROSS 20) | 27.0197 | 52.8943 |
| TRA- VAL (LR CROSS 20) | 79.8469 | 78.3007 |
| TRA- VAL (J48 SPLIT 80) | 26.8669 | 52.4048 |
| TRA- VAL (LR SPLIT 80) | 81.515 | 79.0384 |
| FORECAST (J48 SUPPLIED) | 25.543 | 50.6862 |
| FORECAST (LR SUPPLIED) | 86.0071 | 81.1613 |

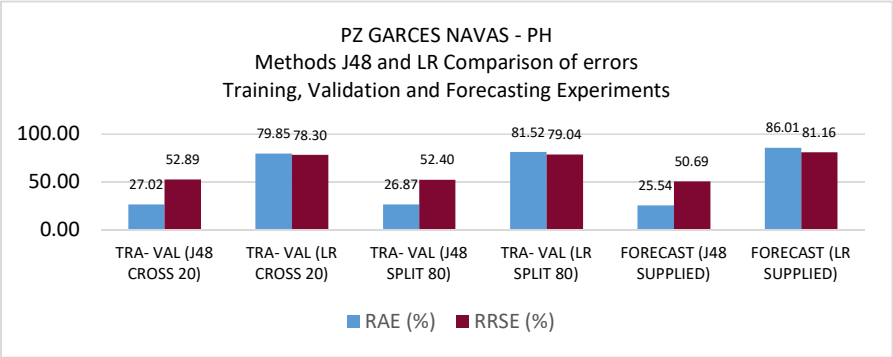



**Table 8.** Error Comparison Garcés Navas NPH (J48 vs. RL)

| EXPERIMENT | RAE (%) | RRSE (%) |
|---|---|---|
| TRA- VAL (J48 CROSS 20) | 33.94 | 61.15 |
| TRA- VAL (LR CROSS 20) | 51.60 | 62.32 |
| TRA- VAL (J48 SPLIT 80) | 34.08 | 61.04 |
| TRA- VAL (LR SPLIT 66) | 51.96 | 62.81 |
| FORECAST (J48 SUPPLIED) | 33.94 | 60.75 |
| FORECAST (LR SUPPLIED) | 52.25 | 62.64 |

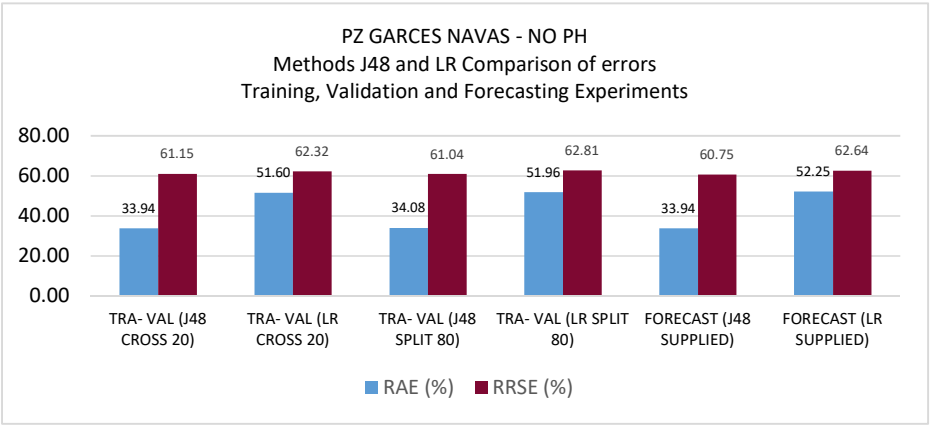


**Table 9.** Error Comparison Arborizadora PH (J48 vs. LR)

| EXPERIMENT | RAE (%) | RRSE (%) |
|---|---|---|
| TRA- VAL (J48 CROSS 20) | 31.98 | 57.03 |
| TRA- VAL (LR CROSS 20) | 81.55 | 77.10 |
| TRA- VAL (J48 SPLIT 80) | 31.45 | 56.65 |
| TRA- VAL (LR SPLIT 66) | 80.59 | 76.22 |
| FORECAST (J48 SUPPLIED) | 30.74 | 55.29 |
| FORECAST (LR SUPPLIED) | 82.74 | 77.49 |

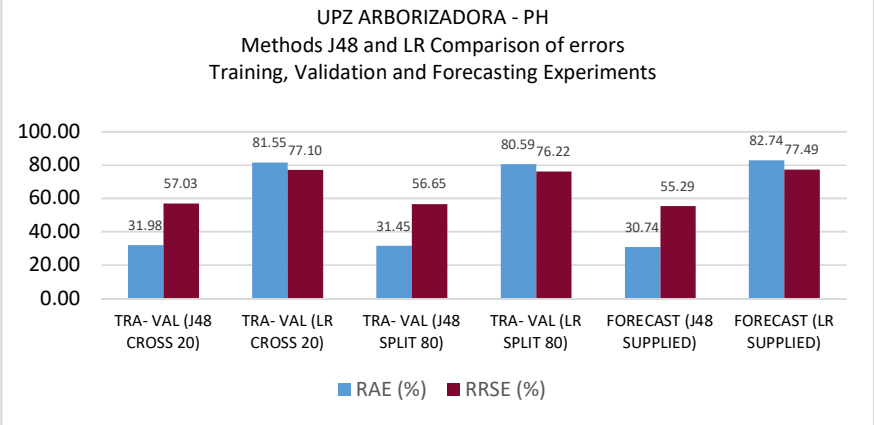



**Table 10.** Error Comparison Arborizadora NPH (J48 vs. LR)

| EXPERIMENT | RAE (%) | RRSE (%) |
|---|---|---|
| TRA- VAL (J48 CROSS 20) | 45.16 | 69.95 |
| TRA- VAL (LR CROSS 20) | 67.44 | 72.29 |
| TRA- VAL (J48 SPLIT 80) | 46.35 | 72.11 |
| TRA- VAL (LR SPLIT 66) | 67.01 | 72.03 |
| FORECAST (J48 SUPPLIED) | 44.27 | 70.36 |
| FORECAST (LR SUPPLIED) | 71.58 | 97.41 |

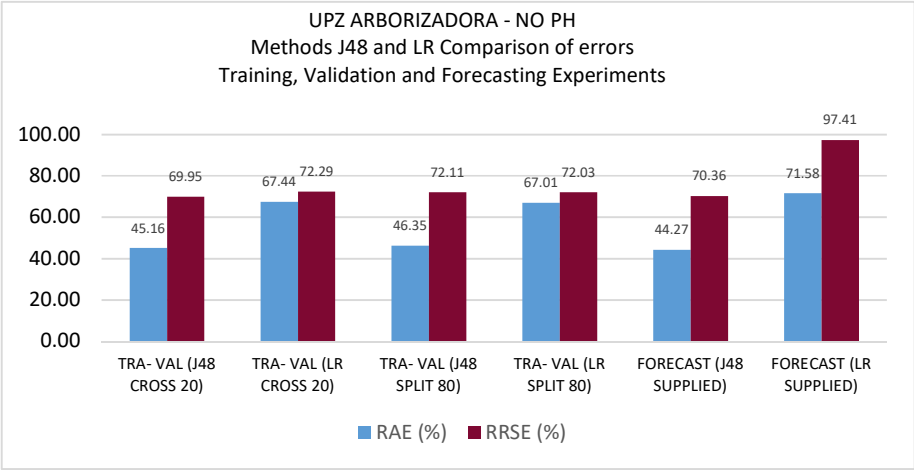


**Table 11.** Error Comparison Calandaima PH (J48 vs. LR)

| EXPERIMENT | RAE (%) | RRSE (%) |
|---|---|---|
| TRA- VAL (J48 CROSS 10) | 30.36 | 55.21 |
| TRA- VAL (RL CROSS 10) | 76.96 | 54.16 |
| TRA- VAL (J48 SPLIT 80) | 29.98 | 55.06 |
| TRA- VAL (RL SPLIT 66) | 76.35 | 53.26 |
| FORECAST (J48 SUPPLIED) | 30.38 | 55.33 |
| FORECAST (RL SUPPLIED) | 78.64 | 58.60 |

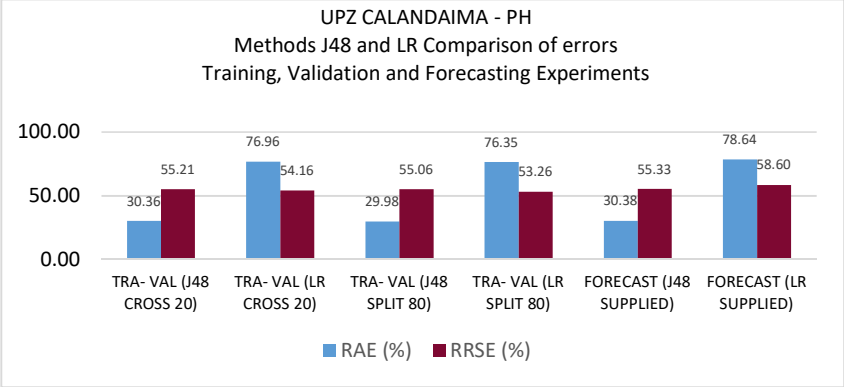
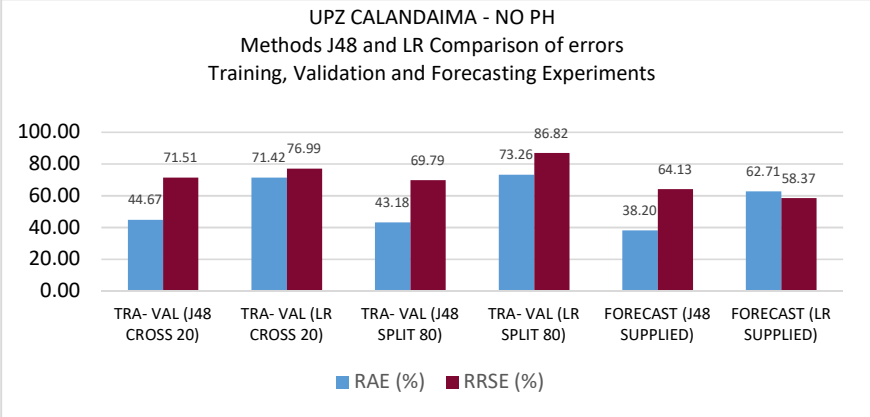



**Table 12.** Error Comparison Calandaima NPH (J48 vs. LR)

| EXPERIMENT | RAE (%) | RRSE (%) |
|---|---|---|
| TRA- VAL (J48 CROSS 10) | 44.67 | 71.51 |
| TRA- VAL (RL CROSS 10) | 71.42 | 76.99 |
| TRA- VAL (J48 SPLIT 80) | 43.18 | 69.79 |
| TRA- VAL (RL SPLIT 66) | 73.26 | 86.82 |
| FORECAST (J48 SUPPLIED) | 38.20 | 64.13 |
| FORECAST (RL SUPPLIED) | 62.71 | 58.37 |



## 4. Discussion

From the resulting graphs it is possible to observe:

- In all cases the relative absolute error (SAR) values obtained in all experiments were lower with the J48 algorithm than in RL between 15 and 50 percentage points.
- Regarding the root of the relative quadratic error, (RRSE) was lower for all experiments in the UPZs Garces Navas and Arborizadora PH and NPH.
- For the UPZ Calandaima PH, the results of RRSE while for validation were better for RL between 1% and 2%, however, for prognosis J48 behaved better by approximately 3%.
- For the UPZ Calandaima No PH, the RRSE of J48 were lower for J48 than for RL in validation, but for forecast RL obtained lower values of RRSE.

## 5. Conclusions

The results of this paper allow us to conclude that with the six databases of the three UPZs and the corresponding subdivisions of PH and Non-PH, with sufficient amounts of data, the J48 decision tree algorithm is a reliable and available procedure to be used. in the calculation of mass valuation with different purposes.

Perhaps the costliest procedure in terms of time was the purification and ordering of the data, which can be solved by standardizing processes and defining modeling protocols; in contrast, the amount of time spent processing data and obtaining results is minimal.

Consequently, the simple processing procedure and the fast and easy classification protocol applied by J48, is the most outstanding quality that this algorithm presents both in training and in prognosis even surpassing traditional methodologies such as RL in precision and efficiency.

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

**List of abbreviations:**

**GIGA:** Research group of the Distrital University

**UAECD:** Special Administrative Unit of the Distrital Cadaster. For its acronym in Spanish.

**UPZ:** Zonal Planning Units by its acronym in Spanish.

**MDL:** Minimum Description Length

**WEKA:** Waikato Environment for Knowledge Analysis

**PH:** Horizontal Property by its acronym in Spanish

**NPH:** No Horizontal Property by its acronym in Spanish

**LR:** Linear Regression

**RAE:** Relative Absolute Error

**RSSE:** Root Sum of Squares Error

**MAE:** Mean of the absolute error

**RMSE:** Root of the mean square error

**Declarations:**

**Funding:** This research was developed with own resources within the GIGA research group of the Distrital University, within the framework of the doctoral thesis in engineering at the Javeriana University.

**Conflicts of interest/Competing interests:** In the present work there is no conflict of interest by its authors)

**Availability of data and material (data transparency):** The data was obtained legally and are available for consultation

**Code availability:** In the present work, no code was developed because properly licensed commercial software and free software were used

**Authors' contributions:**

Edwin Perez: Methodology, conceptual framework, data analysis, processing, writing and final editing.

Nelson Obregon: Methodology, conceptual framework, Supervision and review.

Maria Albancando: Analysis and data processing, writing.

**Acknowledgements:** Special thanks to the UAECD for providing the information for analysis

## References

Alcaldía Mayor de Bogotá, «Decreto 12 de 1993», Bogotá D.C.

Alcaldía Mayor de Bogotá D.C , «UPZ 65 Acuerdos para construir ciudad,» Bogotá, Oficina asesora de prensa y comunicaciones - Secretaría Distrital de Planeación, 2008.

Cazorla, M, Alfonso, M, Escolano, F, Colomina, O y Lozano, M, Inteligencia Artificial, Modelos, Técnicas y Áreas de aplicación, Alicante: Paraninfo, S.A, 2003.

Corso, C, Aplicación de algorítmos de clasificación supervisada, Buenos Aires: Universidad Tecnológica Nacional, 2009.

Cuevas, A, Teoría de la Información, Codificación y Lenguajes, Madrid: Servicio del Ministerio de Educación y Ciencia, 1975.

García, A, Inteligencia Artificial. Fundamentos, práctica y aplicaciones, Madrid: RC Libros, 2012.

García, F, Aplicación de técnicas de minería de datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA), Granada: Universidad de Granada, 2013.

Hernández, J, Práctica de minería de datos, Introducción a WEKA, Valencia: Universidad Politécnica de Valencia, 2006.

Instituto Geográfico Agustín Codazzi (IGAC), Resolución 620 de 2008, Bogotá, 2008.

McCarthy, J, Mechanisation of Thought Processes., Simposio N.10 Volumen I ed., Londres, 24 - 27 de Noviembre de 1958.

Mitchell, T, Machine Learning, Portland: McGraw Hill, 1997.

Mood, A, Graybill, F y Boes, D, Introduction to the Theory of Statistics, Auckland: McGraw Hill, 1974.

Morate, D, mabual de WEKA, Granada, 2000.

Murphy, K, Machine Learning A Probabilistic Perspective, Cambridge, Massachusetts, 2012.

Pino, R, Gómez, A y de Abajo, N, «Introducción a la Inteligencia Artificial: Sistemas Expertos, Redes Neuronales Artificiales, y computación evolutiva», Asturias, Servicios y publicaciones de la Universidad de Oviedo, 2001.

Sancho, F, «Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Sevilla,» 10 Diciembre 2016. [En línea]. Available: http://www.cs.us.es/~fsancho/?e=104. [Último acceso: 14 Abril 2017].

Secretaría Distrital de Planeación, «Conociendo la localidad de Kennedy. diagnóstico de los aspectos físicos, demográficos y socioeconómicos», Bogotá, 2009.

Secretaría Distrital de Planeación, «21 Monografías de las localidades: Diagnóstico de los aspectos físicos, demográficos y socioeconómicos de las localidades – 2011. # 19 Ciudad Bolívar», 2011, p. Bogotá.

Secretaría Distrital de Planeación , «21 Monografías de las Localidades: Diagnóstico de los aspectos físicos, demográficos y socioeconómicos,» Bogotá, 2011.

Vizcaino, P. A, Aplicación de técnicas de inducción de árboles de decisión a problemas de clasificación mediante el uso de WEKA, Bogotá. KONRAD LORENZ, 2008.

# Figures

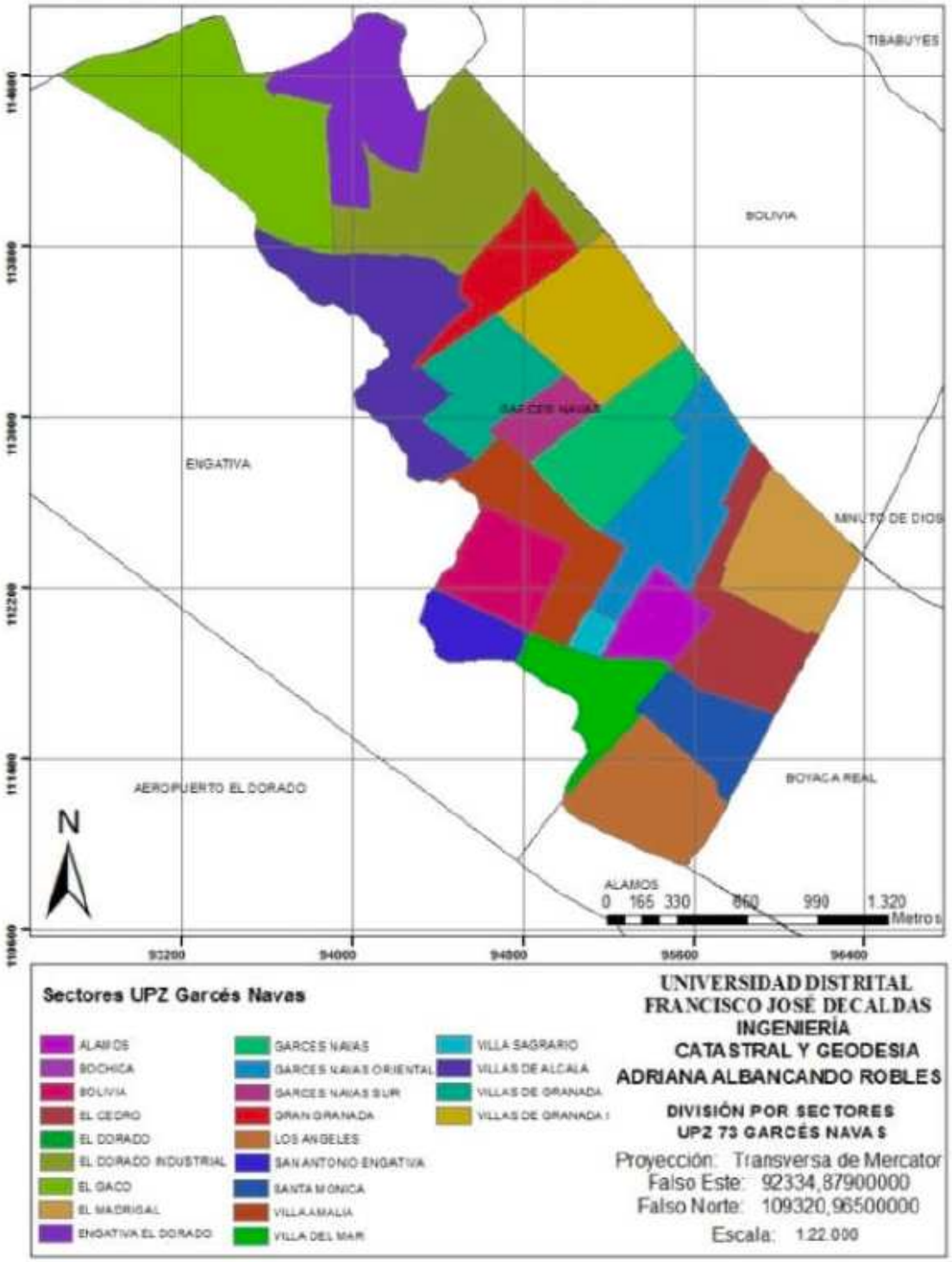



**Figure 1**

Garcés Navas UPZ Zoning

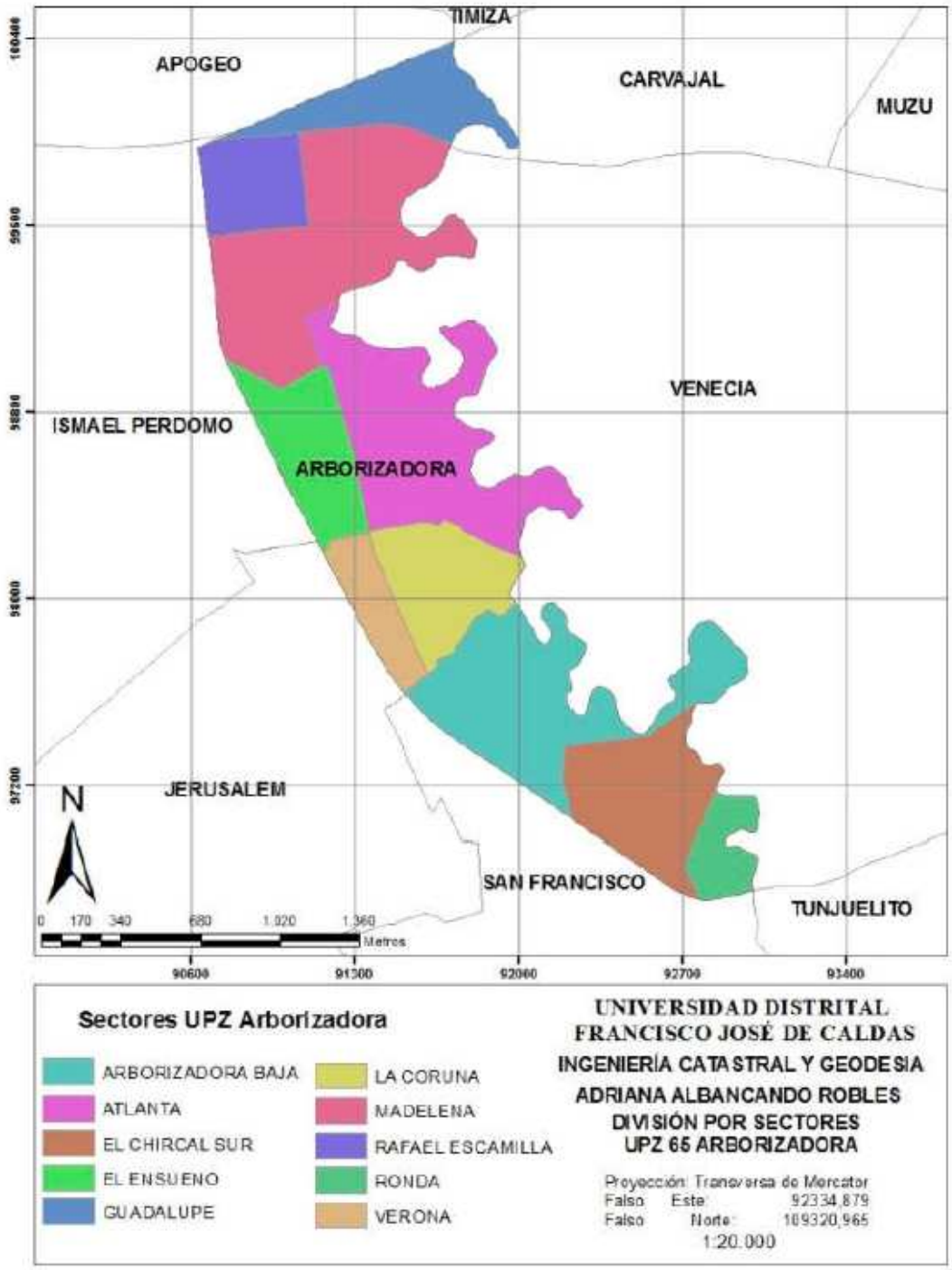



**Figure 2**

Arborizadora UPZ Zoning

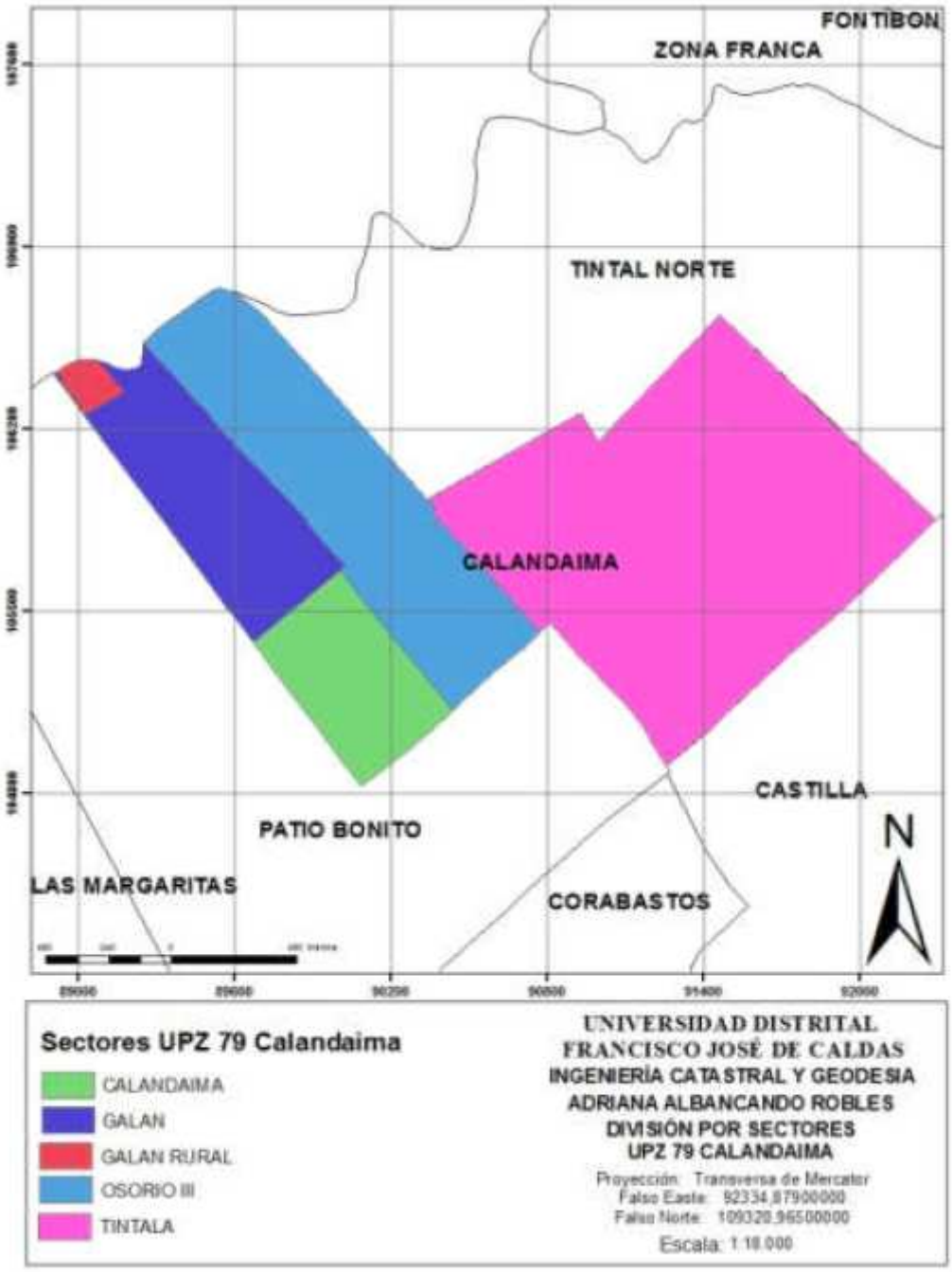



**Figure 3**

Calandaima UPZ Zoning

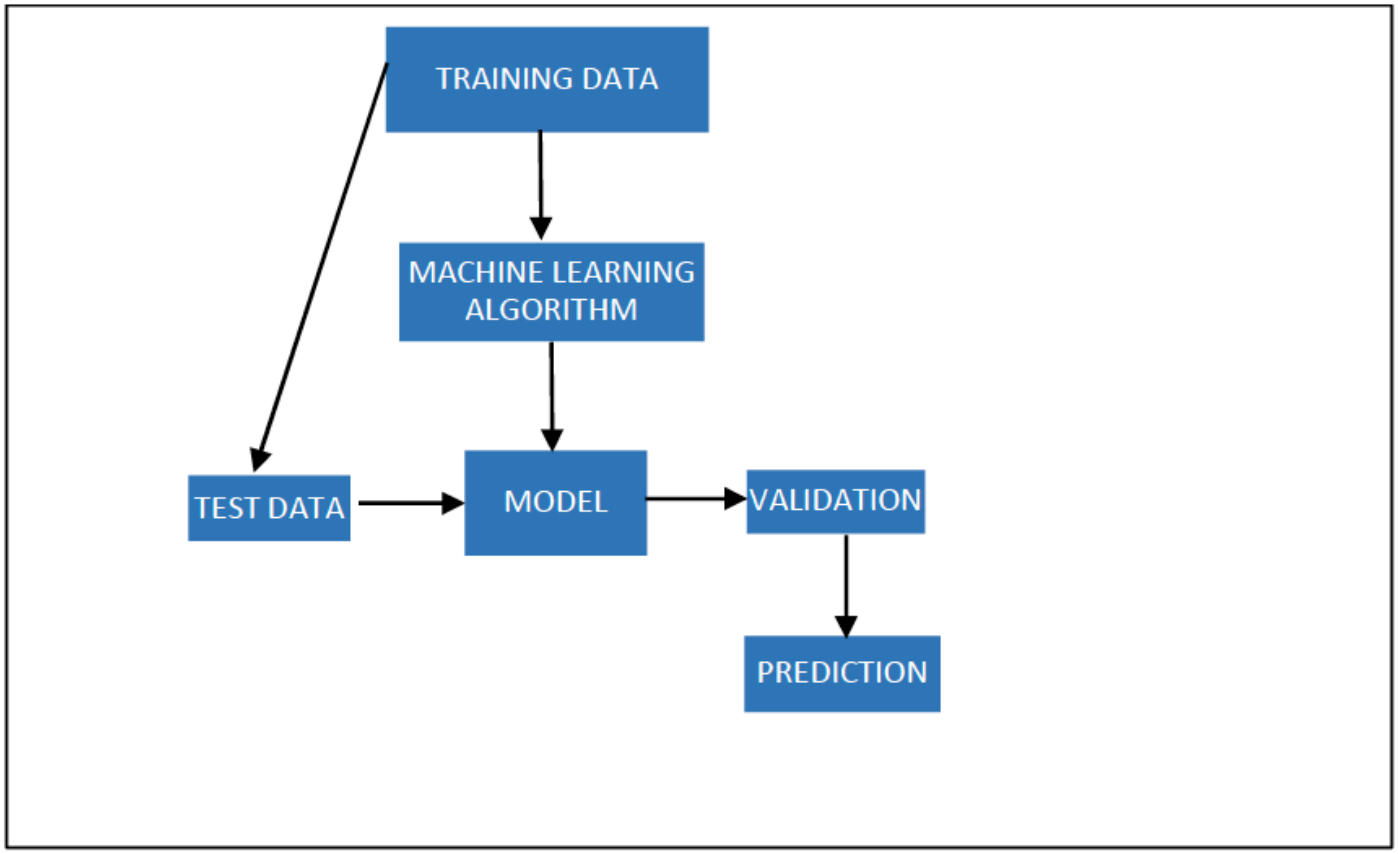



**Figure 4**

Learning Machine Process

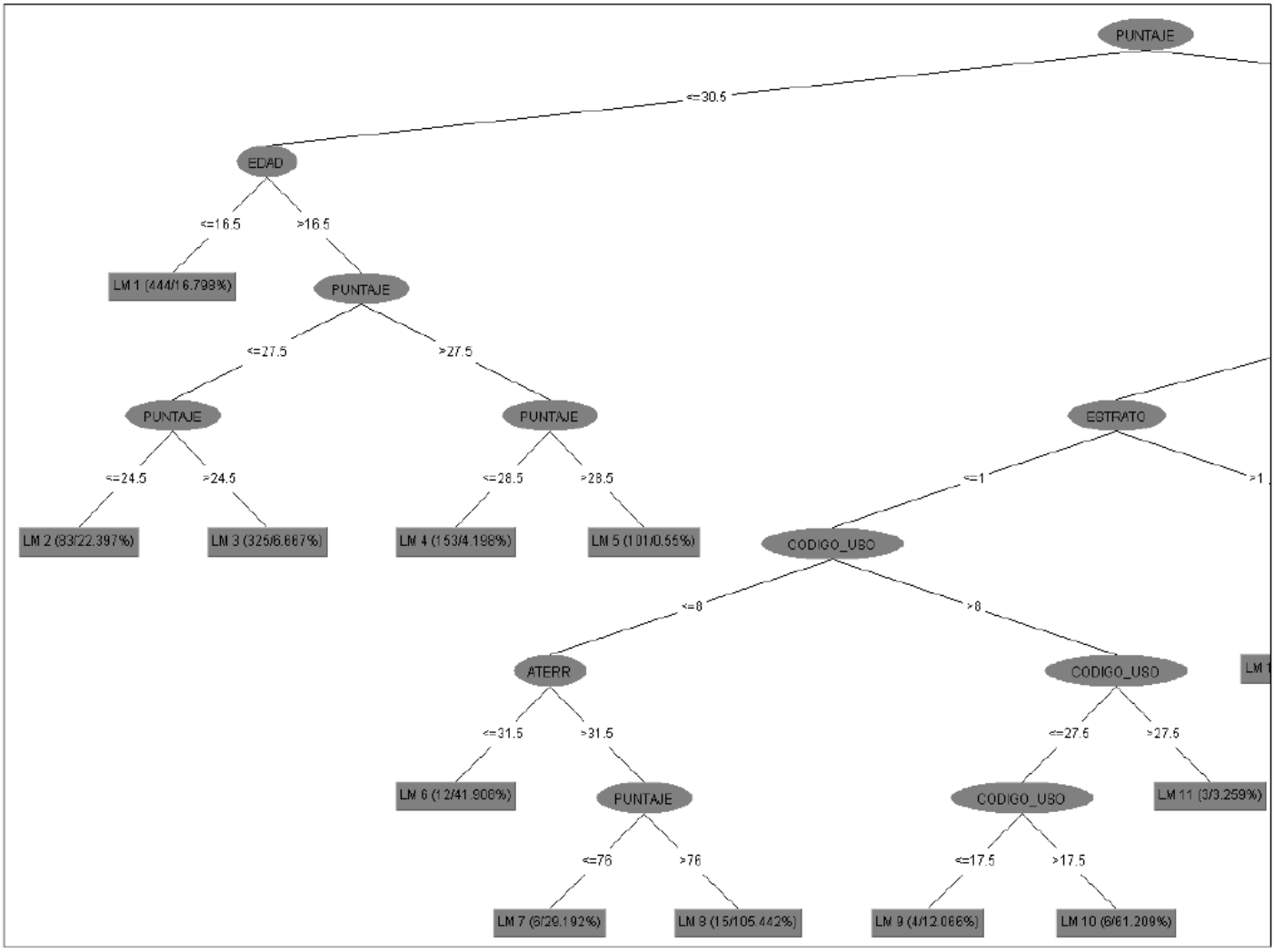



**Figure 5**

Example of decision tree generated in WEKA

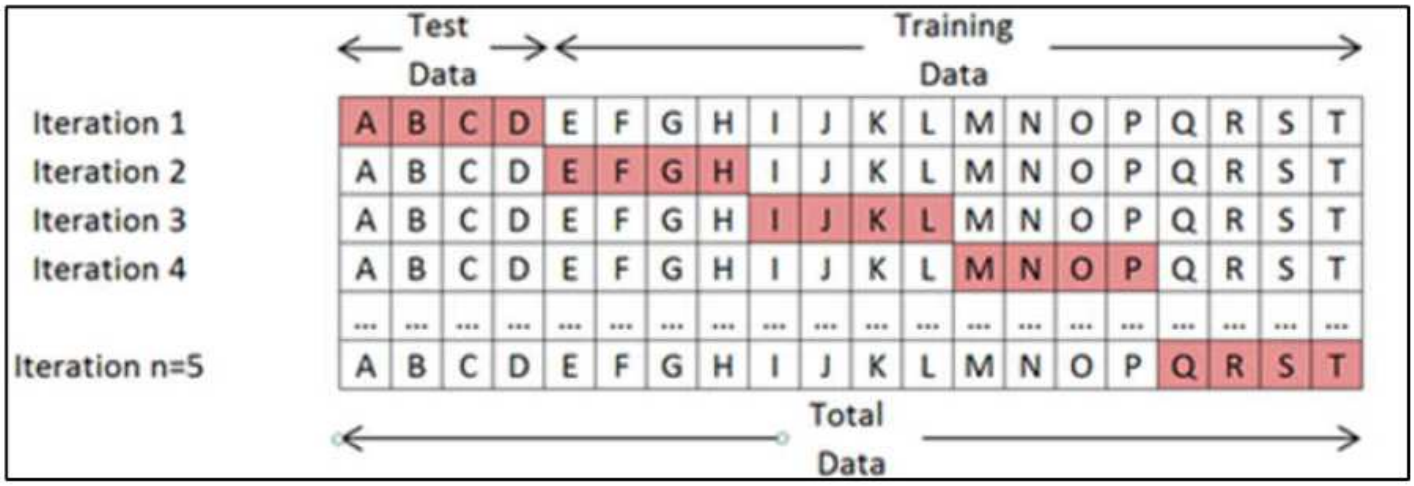



**Figure 6**

Example of cross validation

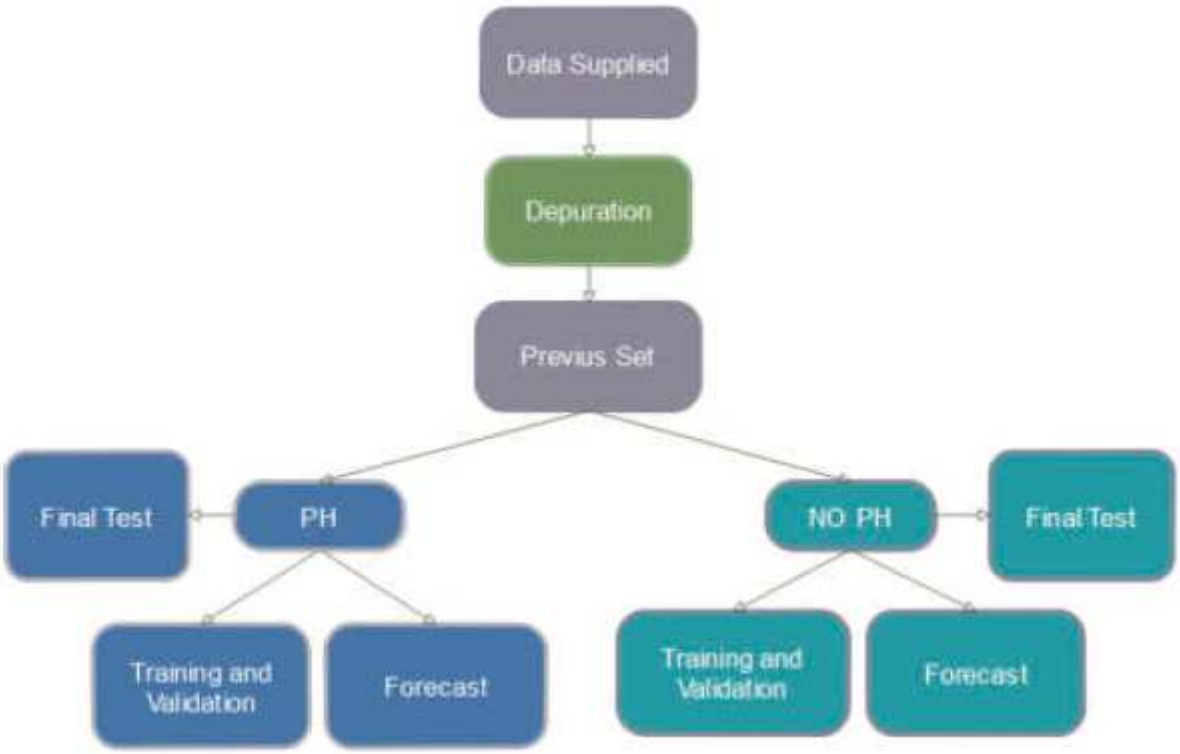



**Figure 7**

Organization of data for experimentation

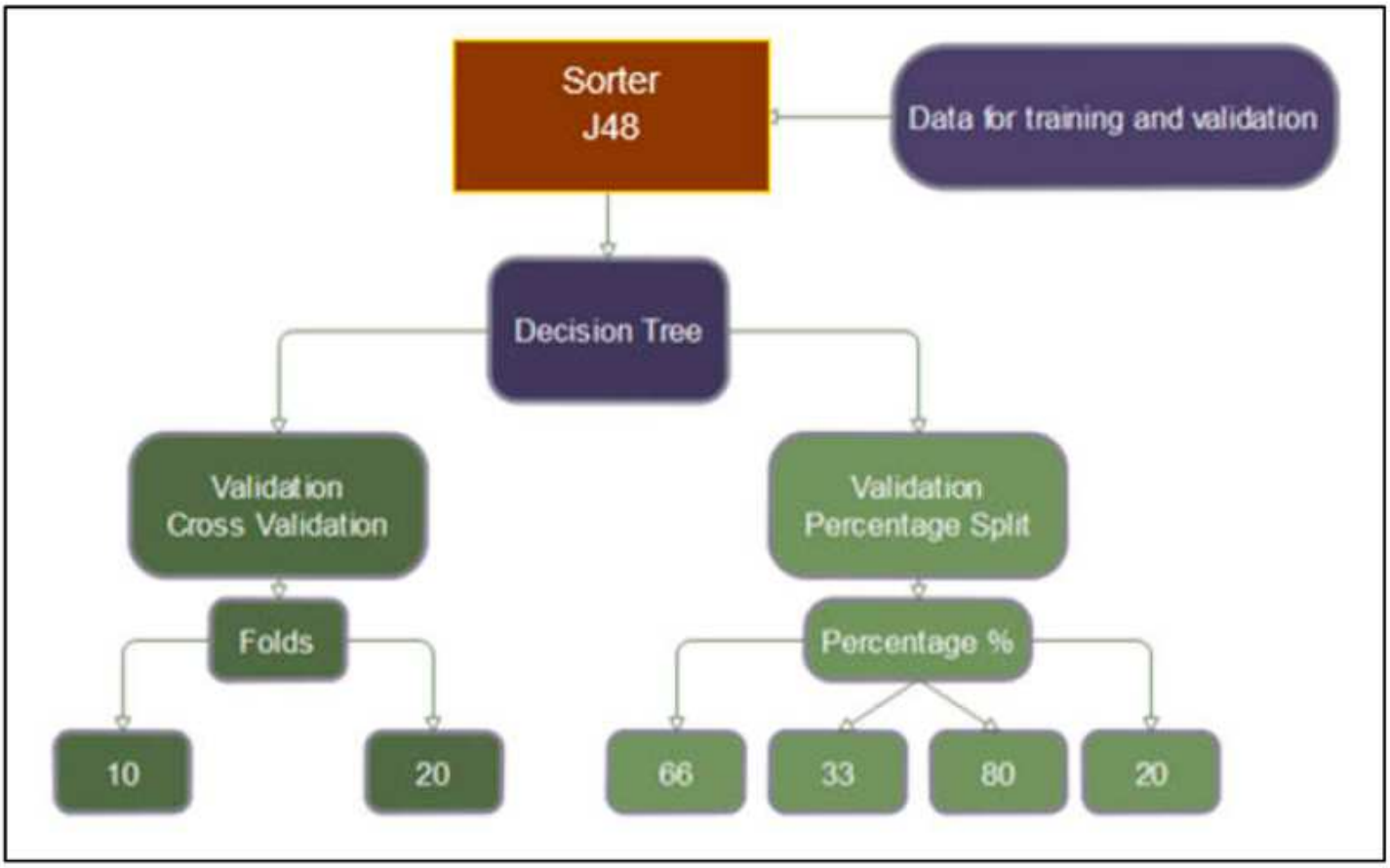



**Figure 8**

Training and Validation process

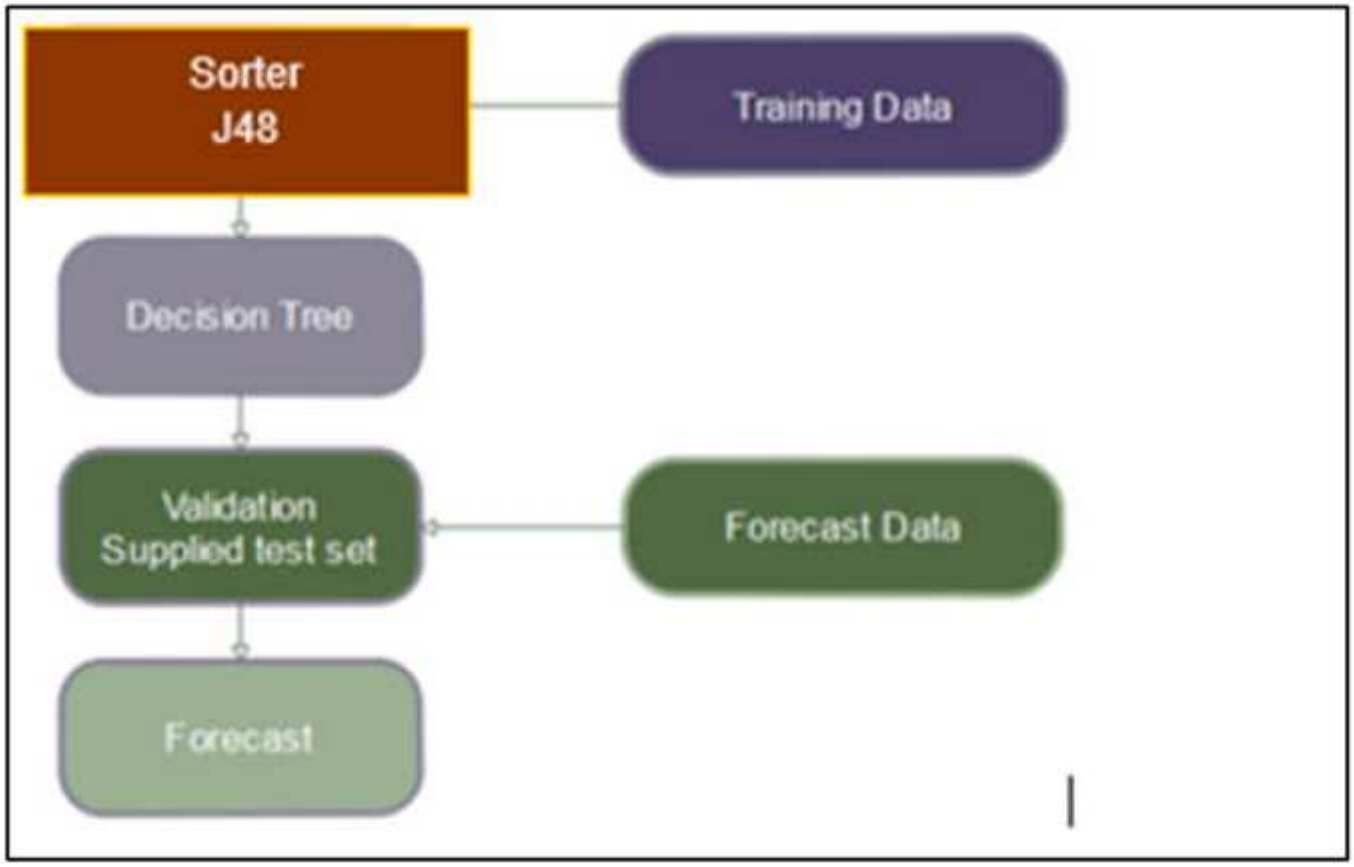



**Figure 9**

Forecast process

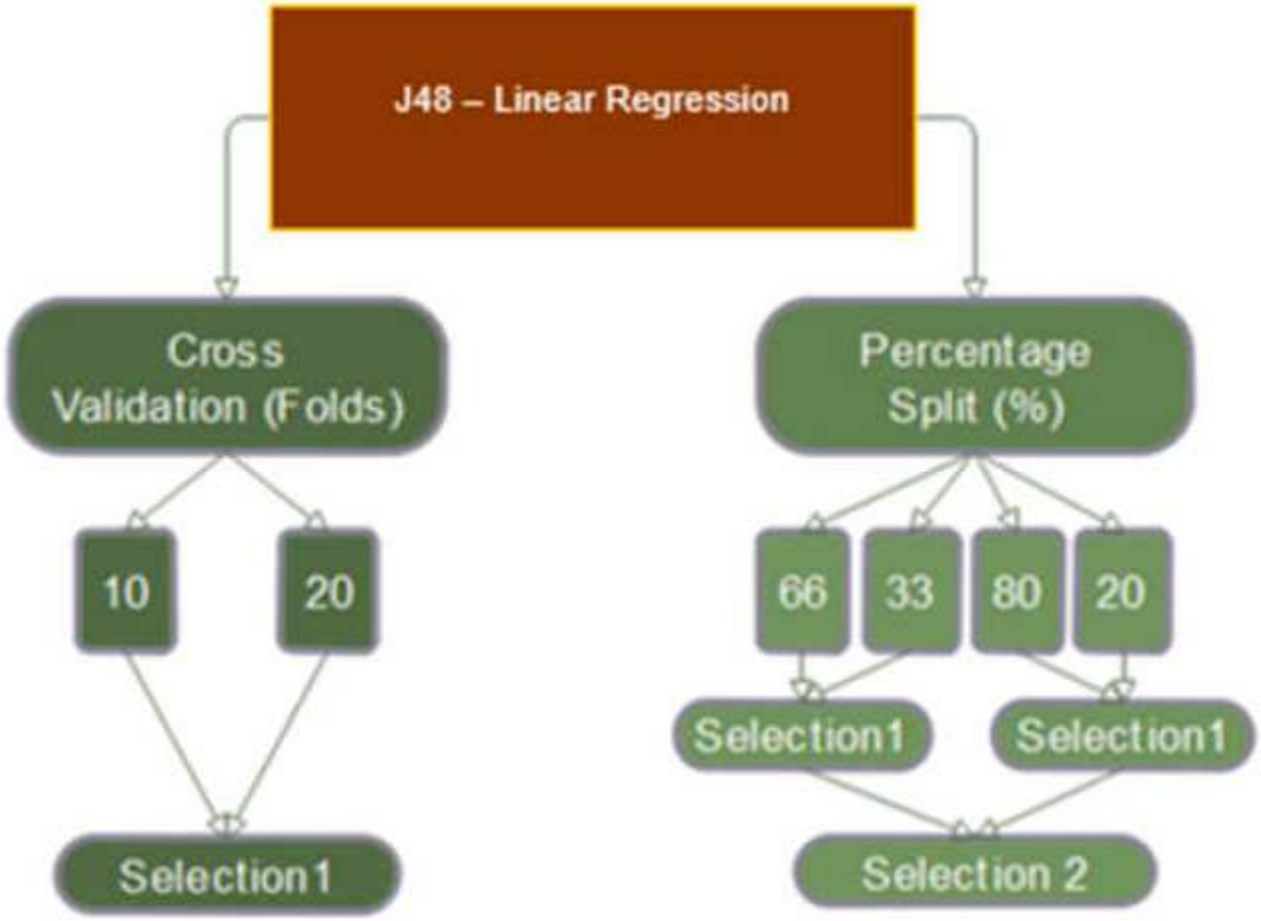



**Figure 10**

Model selection for J48 and for LR