

# A Machine Learning Method for the Analysis of Urban Italian Mobility

Gabriella Schoier(✉) and Giuseppe Borruso

DEAMS – Department of Economic, Business, Mathematic and Statistical Sciences "Bruno de Finetti", University of Trieste, Valerio, 4/1 – 34127, Trieste, Italy
{gabriella.schoier,giuseppe.borruso}@deams.units.it

**Abstract.** The aim of this paper is to consider a method of machine learning to analyze the problem of the sustainability of the urban transport in Italian cities. First of all we recall the definition of sustainable mobility then we present some indicators considered in our analysis.

The methodology used in this paper are decisional trees.

We both consider classification and regression trees. We have chosen two different dependent variables one for classification trees (a categorical variable: Macroregion according to NUTS 1: North West, North East, Centre, Islands and South of Italy) and one for regression trees (a quantitative variable: PM10 maximum number of days in excess of the human health protection limit foreseen for PM10). In order to test the performance of this methodology we have applied random forest.

The analysis has been performed using SAS language.

**Keywords:** Spatial Data Mining · Decisional Trees · Urban Mobility

## 1 Introduction

From the beginning of the twenty-first century, a fundamental idea has been how to extract information from big data. Big data analytics could provide opportunities to develop new knowledge to reshape our understanding of different fields and to support decision-making.

A huge amount of data i.e. big data have been collected by private and public organizations (see eg. [12, 14–16]). An example of big data collection regards urban transport. In particular, the problem of sustainability of urban transport has been recognized as a crucial economical social and political objective. To analyze the notion of urban sustainable mobility and to choose appropriate indicators it is necessary to define the notion of urban sustainable mobility and to choose appropriate indicators enabling its measurement.

The aim of this paper is to analyse urban transport mobility at province level in Italy on the base of some indicators.

The aim of this research is to understand how different variables influence urban transport mobility in the territories (provinces).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14107, pp. 590–603, 2023.
https://doi.org/10.1007/978-3-031-37114-1_41

The new idea regards the application of decision trees to some indicators and random forest to test the goodness of fit.

We consider both classification and regression trees; two different dependent variables have been chosen:

– classification trees dependent variable Macroregion according to NUTS 1: North West, North East, Centre, Islands and South of Italy,
– regression trees dependent variable PM10_20 ( maximum number of days in excess of the human health protection limit foreseen for PM10).

Our objective is to see how different indicators move to analyze and then monitor the sustainability of the urban transport in Italian cities in particular regarding the chosen dependent variables.

## 2 The Methodology: Decision Trees and Random Forest

### 2.1 Decision Trees

As it is well known decision trees are a part of hierarchical classification or segmentation techniques [2, 11]. These techniques have the purpose of "sorting" statistical units into the various classes of a dependent variable on the basis of the values of one or more explanatory variables.

In 1984 Breiman and others introduced an innovative segmentation technique. This technique is called Classification and RegressionTrees (CART) (see [3]). It is presented as a recursive and binary partition methodology.

The hierarchical segmentation process used in the construction of decisional trees consists in divide the statistical units in a finite number of disjoint subgroups in order to guarantee an internal homogeneity higher than that of the initial dataset and a high heterogeneity between the subgroups.

A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name says, it uses a tree-like model of decisions. Though it is a commonly used tool in data mining for deriving a strategy to reach a particular goal, it is also widely used in machine learning. They are easy to interpret and make for visualizations, they make possible to reproduce work, they can handle both numerical and categorical data, they perform well on large dataset and are extremely fast.

A decision tree is a type of predictive model that has been developed independently in the statistics and artificial intelligence communities. The HPFOREST procedure creates a tree recursively. An input variable is chosen and used to create a rule to split the data into two segments. The process is then repeated in each segment, and then again in each new segment, and so on until some constraint is met. In the terminology of the tree metaphor, the segments are nodes, the original data set is the root node, and the final segments are leaves or terminal nodes. A node is an internal node if it is not a leaf. The data in a leaf determine the estimates of the value of the target variable. These estimates are subsequently applied to predict the target of a new observation assigned to the leaf.

This type of models are used for both classification and regression (see eg. [6]).

The performance of a tree can be further increased by pruning in particular by using CART (see [3]) example of indices used for pruning are Entropy and Gini for classification trees and RSS (residual sum of squares) for regression trees.

$G = 1 - \sum_j f_j^2$ Gini index

$H = -\sum_j f_j \log f_j$ Shannon index

where $f_j$ is the relative frequency.

It involves removing the branches that make use of features having low importance. In so doing the complexity of tree is reduced and the power by reducing overfitting is increased.

The simplest method of pruning starts at leaves and removes each node with most popular class in that leaf, this change is kept if it does not deteriorate accuracy. It is also called reduced error pruning.

More sophisticated pruning methods can be used such as cost complexity pruning where a learning parameter (alpha) is used to weigh whether nodes can be removed based on the size of the sub-tree. In order to obtain the sequence of trees of decreasing dimension, one defines, for every tree $T \leq T_{max}$ a cost complexity function $R_\alpha(T)$

$$R_\alpha(T) = \widehat{R}(T) + \alpha \left|\tilde{T}\right|$$

where $\widehat{R}(T)$ estimate of rate of wrong classification,

$\left|\tilde{T}\right|$ the numbers of leaves.

### 2.2 Random Forest

Random forest is a high-performance method that creates a predictive model called a forest that consists of several decision trees (see [4]).

The purpose of a predictive model is to predict a target variable on the basis of independent variables/regressors.

In particular random forest trains the model using training data in which the target values are known then model is applied to observations in which the target is unknown.

If the predictions fit the new data well, the model is said to generalize well. The method has the ability to perform both classification and regression prediction. We can say that random forests are an improved extension on classification and regression trees (CART) with respect to instability and accuracy.

Random forest creates different decision trees. The differences regard the fact that the training data for a tree is a sample with replacement from the original training data and that the regressors considered for splitting a node are randomly selected from all available independent variables. Among these variables only a single one is used to form a splitting rule. The chosen variable is the one that is most associated with the target one.

Training a forest can require training hundreds of decision trees. Specifically, random forests remain relatively stable with changes in data due to the combination of many trees. Random forests are computationally simple to fit, there are no formal distributional assumptions. The Random Forest algorithm consider a forest of $n$ trees for each tree we have to take $B$ random sample with replacement ( boostrap samples) to grow a tree then

we have to select for each node indepeldently and randomly $m$ variables from $M$ possible variables and use them as predictors. Trees are grown until the nodes can no longer be split. These predictors are used as the "out-of-bag" (OOB) data to measure the accuracy of the model [4]. The observation is passed down these trees and it is classified in the class which obtain the majority rule, giving the OOB prediction for the observation. The error rate of all the OOB predictions is the OOB error rate of the forest.

Random forests compute the importance of variables in different ways. For classification problems the Gini criterion can be used. For a given tree, the Gini variable importance measure for a particular variable of interest is the weighted average of the decrease in the Gini impurity criteria of the splits based on this variable. This is averaged over the all the trees in the forest to get the Gini importance for the forest itself. Important variables yield high Gini variable importance measures.

Another variable importance calculation is called permutation importance. Permutation importance is based on predictive accuracy. The out-of-bag error rate is computed from both a data set obtained from permuting the values of a particular variable of interest in the "out-of-bag" OOB data and the original OOB data. The difference between these two OOB error rates gives the permutation variable importance. Important variables yield high permutation variable importance measures.

## 3 The Application

### 3.1 The Data and the Variables

The study of sustainable mobility plays an important role in the socio-economic field and it is of great importance for the politicians and the economists(see e.g. [13]).

To obtain this pourpose it is essential define the notion of sustainable mobility and to measure it. Numerous definitions of sustainable mobility are proposed, the most famous one was introduced in the Brundland report according to which: "Sustainable transport meets the mobility needs of the present without compromising the ability of future generations to meet these needs." (see [17]).

The European Conference of Ministers of Transport (see [5]) further specified the main features that a sustainable mobility system should meet, that is:

- allowing the basic access and development needs of individuals, companies and society to be met safely and in a manner consistent with human and ecosystem health, and promotes equity within and between successive generations;
- being affordable, operating fairly and efficiently, offering a choice of transport mode, and supporting a competitive economy, as well as balanced regional development;
- limiting emissions and waste within the planet's ability to absorb them, using renewable resources at or below their rates of generation, and using non-renewable resources at or below the rates of development of renewable substitutes, while minimizing the impact on the use of land and the generation of noise.

To measure the environmental sustainability of the urban mobility requires taking into account both the environmental, the economic and the social aspect [17]. Moreover, the selection process should be made explicit and should adhere, according to the COST

action 356, to ten criteria: validity, reliability, sensitivity, measurability, data availability, ethical concerns, transparency, interpretability, target relevance, and actionability [11].

On the basis of these guidelines 13 indicators (reported in Table 1) have been selected and have been measured with respect to Italian provincial towns.

As regards the economic aspect three indicators have been selected: Income, Salary, Rate of Employment.

The environmental indicators chosen are six: Gasoline Fuel, Diesel Fuel, Low Emission, E-Charging Stations Density, $PM_{10}$, Urban Green Density.

Finally the indicators used to measure the social dimension of the Italian urban transport systems are four: Rate of Fatal Accidents, Dead Pedestrians, Vehicles Density per KM2, Population_Density.

The variables considered in this study are reported in the following Table 1 (see [7–10]).

**Table 1.** Indicators for Sustainable mobility of Italian cities

| Name | Description |
|---|---|
| City | Province |
| Macroregion | Macroregion according to NUTS 1: North West, North East, Centre, Islands and South of Italy |
| RateFatalAccident_19 | Fatality accident rate |
| Dead_Pedestrians_19 | Pedestrians died in accidents |
| Population_Density_21 | Popolation density |
| Vehicles_DensityKM2_19 | Vehicle density per $km^2$ |
| RateEmployment_21 | Employment rate 20-64 years |
| Income_17 | Income |
| Salary_20 | Salary |
| GasolineFuel_Car_21 | Petrol/gasoline vehicles % |
| DieselFuel_Car_21 | Diesel-fuelled vehicles % |
| LowEmission_Car_21 | Low emission vehicles % |
| Echarging_Stations_Density_19 | Density of electric car charging columns |
| PM10_20 | maximum number of days in excess of the human health protection limit foreseen for PM10 |
| Urban_Green_Density_20 | Urban green density |

The data font is the Italian National Institute of Statistics (Istat). Some variables are not available for all the provincial towns.[1]

[1] These cities have not be considered in this analysis: Andria e Trani.

At provincial level, the most up-to-date available data refer to the year 2021 and 2019, unless the economic variables - income and salary - which refer respectively to the years 2017 and 2020, but they have been include in the analysis as they have been considered fundamental for our purpose. The criteria for the choice of each indicators inside the dimension is done as [13], i.e. the correlations within each domain, the variability indexes (i.e. coefficient of variation and the quartile difference for standardized data) and the adequacy of the indicator to the analysis to be carried out.

### 3.2 Classification Trees

The aim of the first part of our analysis is to consider a categorical dependent variable in order to forcast it. The dependent variable is Macroregion according to NUTS 1: North West, North East, Centre, Islands and South of Italy.

We have applied different models with and without pruning. At the end on the base of a compromise between complexity and rate of error we have chosen a classification tree with split Entropy and prune Gini index.

As we can see on the base of the model the rate of error is very low as regards the Island it is equal to zero for Centre, North West and South is less than 0.2 while for the North East is less than 0.25 see Table 2.

**Table 2.** Confusion matrix based on the model.

| Macroregion | Centre | Islands | North East | NorthWest | South | Error rate |
|---|---|---|---|---|---|---|
| Centre | 19 | 0 | 0 | 1 | 3 | 0.1739 |
| Islands | 0 | 13 | 0 | 0 | 0 | 0.0000 |
| North East | 1 | 0 | 17 | 4 | 0 | 0.2273 |
| North West | 2 | 0 | 2 | 21 | 0 | 0.1600 |
| South | 1 | 2 | 0 | 0 | 21 | 0.1250 |

In Fig. 1 the selected model using the Gini's index for the pruning is reported.

As one can see the number of chosen leaf according to the Gini index are 10; the value of the Gini index is 0.237.

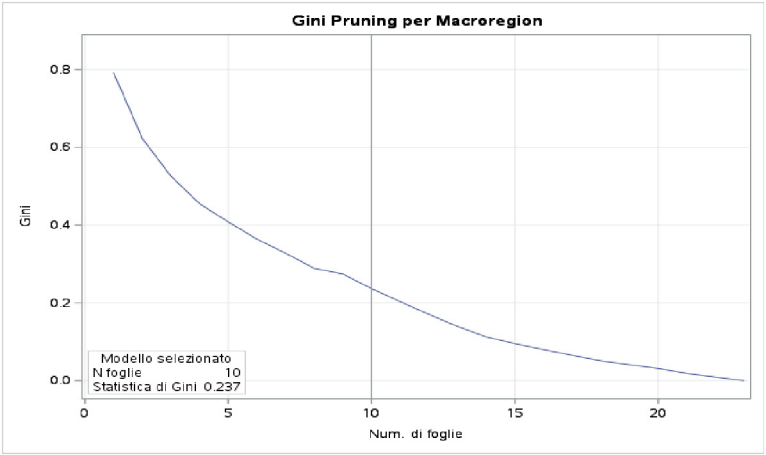


**Fig. 1.** The selected model using the Gini's index for pruning

The importance of the variables for the construction of the classification trees is reported in Table 3.

**Table 3.** Confusion matrix based on the model

| Variable | Realative importance | Importance | Count |
|---|---|---|---|
| RateUnemployment_21 | 1.0000 | 4.2501 | 1 |
| GasolineFuel_Car_21 | 0.9972 | 4.2381 | 2 |
| Salary_20 | 0.7066 | 3.0030 | 2 |
| Dead_Pedestrians_19 | 0.5125 | 2.1780 | 1 |
| DieselFuel_Car_21 | 0.4802 | 2.0407 | 1 |
| Vehicles_DensityKM2_19 | 0.4624 | 1.9654 | 1 |
| Low_Emission_Car_21 | 0.2954 | 1.2553 | 1 |

In the Fig. 2 the obtained classification tree is produced. The leaves (terminal nodes) are: node3, node 7, node 8, node A, node B, node C while node 9 is divided in node D (terminal node) and node E which is divided in two terminal nodes: node H and node 1.

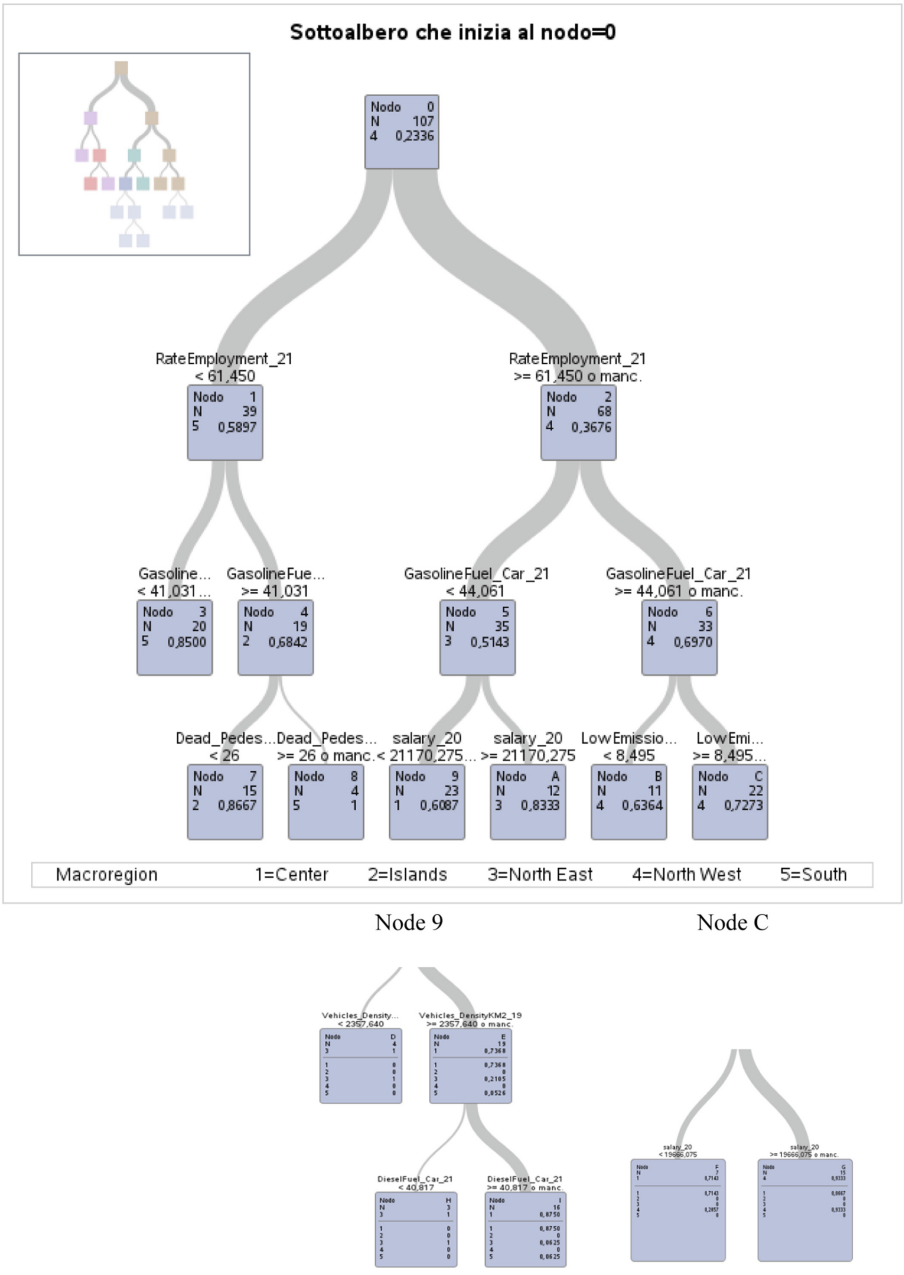


**Fig. 2.** Classification tree

There are 107 units in the root node (node 0).

These units are divided into 39 units with RateEmployment_21 < 61.450 (node 1) and 68 units for node 2 with RateEmployment_21 ≥ 61.450 (node 2).

Node 1 is divided in two nodes:

- node 3 (terminal node) if GasolineFuel_Car_21<41.031, twenty units assigned to class 5 (South) (prob. 0.85),
- node 4 if GasolineFuel_Car_21≥41.031, divided in two nodes:
  node 7 (terminal node) if Dead_Pedestrians_19<26 fifteen units assigned to class 2 (Islands) (prob. 0.87),
  node 8 (terminal node) if Dead_Pedestrians_19≥26 four units assigned to class 5 (South) (prob. 1).

Node 2 is divided in two nodes:

-node 5 if GasolineFuel_Car_21<44.061, divided in two nodes:
  -node 9 if Salary_20<21170,275 divided in:
  -node D (terminal node) if Vehicles_DensityKM2_19<2357,640 four units assigned to class 3 (North East) ,
  -node E if Vehicles_DensityKM2_19≥2357,640 divided in:
  -node H (terminal node) if DieselFuel_Car_21<40.817 three units assign to class 3 (North East),
  -node I (terminal node) if DieselFuel_Car_21≥40.817 sixteen units assigned to class 1 (Centre)
  -node A (terminal node) if Salary_20≥21170,275 twelve units assigned to class 3 (North East) (prob. 0.8333).
- node 6 if GasolineFuel_Car_21≥44.061, divided in two nodes:
  node B (terminal node) if Low_Emission_Car_21<8.495 eleven units assigned to class 4 (North West) (prob. 0.6364),
  node C if Low_Emission_Car_21≥8.

In order to test the classification model a random forest model has been performed. He results regarding the OOBS are reported in the next figure (Fig. 3).

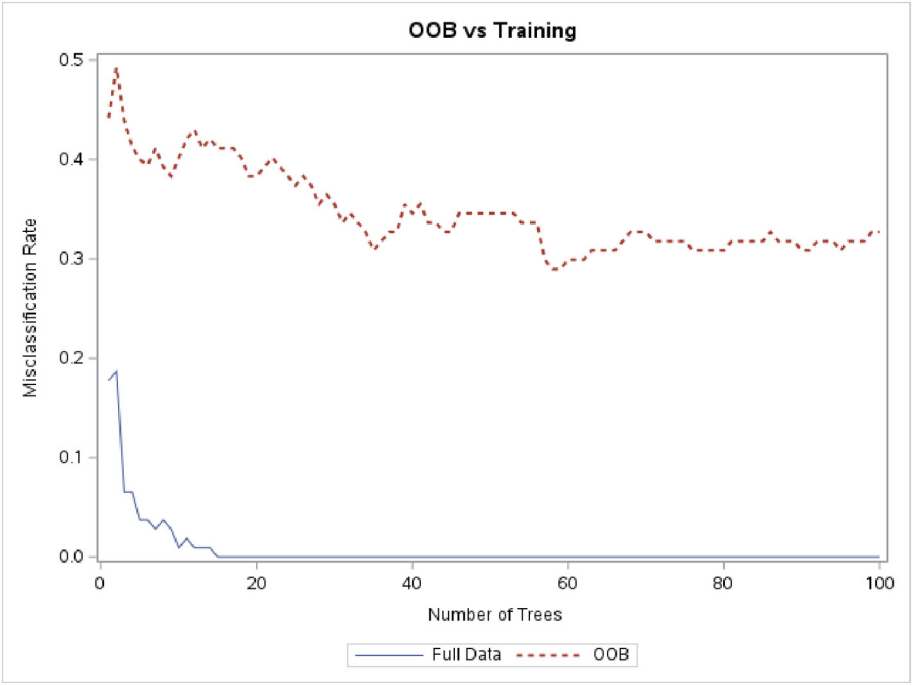


**Fig. 3.** OOB vs Training

### 3.3 Regression Trees

In order to apply regression trees we have chosen as dependent variable PM10_20 We try to predict the dependent variable on the base of the other variables that are.

We have applied different models with and without pruning. At the end on the base of a compromise between complexity and rate of error we have chosen a regresssion tree with split RSS and prune cost complexity.

We have carried out the pruning of the tree to avoid overfitting of the model on the data and to find a compromise between simplicity and discriminatory power.

As regards the pruning we have preferred cost complexity which is an algorithm based on a trade off between the complexity (size) of the tree and the error rate to prevent overfitting.

The final tree has a depth equal to 6 leaves. The ASE (Average Square Error for Regression) is given by the ratio between RSS (residual sum of squares) and the number of units of the node. In our case the ASE medium minimum is equal to 30.63 while the parameter of cost complexity is 145.4; this refers to a numbers of leaves equal to 6 (see Fig. 4).

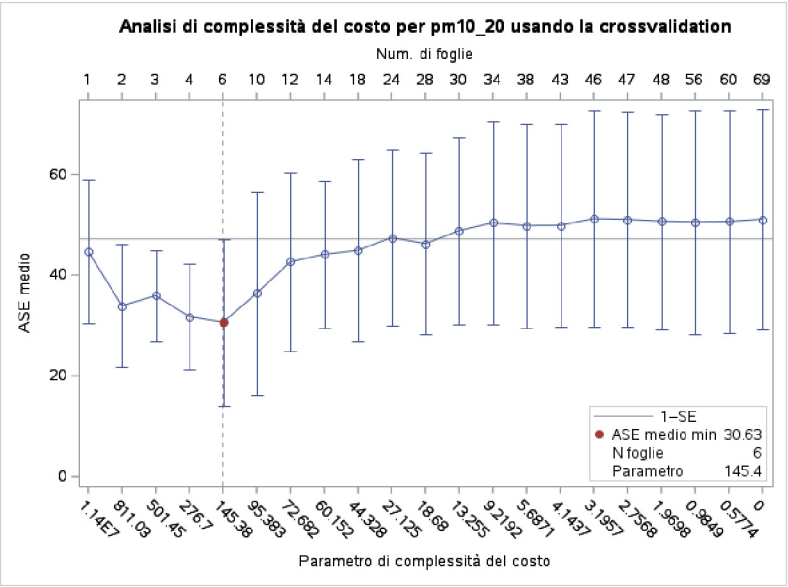


**Fig. 4.** The selected model using pruning costcomplexity

The final consideration regards variable importance; from Table 4 we can see which variables are the most important: Low_Emission_Car_21, Macroregion, Population_Density_21, Income_17, Dead_Pedestrians_19. The presence of the Macroregion variable shows us once again that there is actually a basic difference between the various Italian macro areas (Fig. 5).

**Table 4.** Variable importance.

| Variable | Realative importance | Importance | Count |
|---|---|---|---|
| Low_Emission_Car_21 | 1.0000 | 36.4136 | 2 |
| Macroregion | 0.6583 | 23.9698 | 1 |
| Population_Density_21 | 0.5745 | 20.9200 | 1 |
| Income_17 | 0.3812 | 13.8826 | 1 |
| Dead_Pedestrians_19 | 0.3374 | 12.2875 | 1 |

As regards the goodness of fit of the model with 7 leaves ASE is equal to 16.21 while the RSS is 1589.1. Using cross validation with 6 leaves ASE is 34.7613.

There are 107 provinces that have been collected. The procedure HSPLIT[2] use 98 provinces in the root node (node 0). These units are divided into 37 units with

[2] This is the procedure used by SAS language for decision trees.

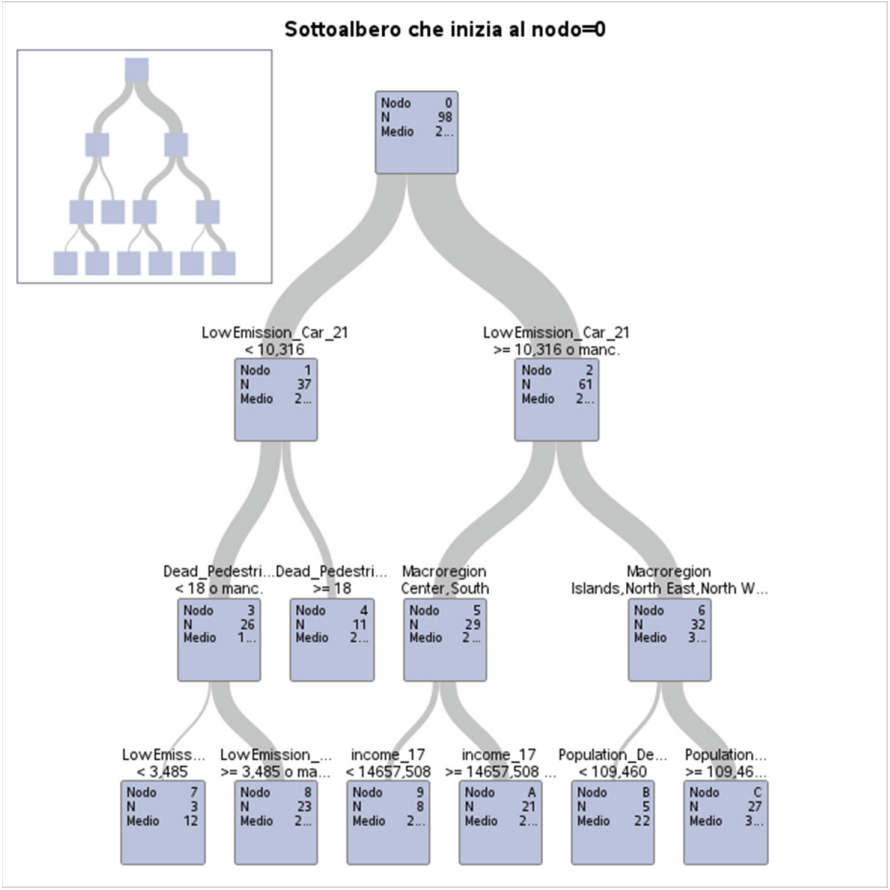


**Fig. 5.** Regression trees with pruning cost complexity

Low_Emission_Car_21 $<$ 10.316 (node 1) and 61 units with Low_Emission_Car_21 $\geq$ 10.316 (node 2).

Node 1 is divided in two nodes: node 3 if Dead_Pedestrians_19 are less than 18 or there is a missing value[3] and node 4 if Dead_Pedestrians_19 $\geq$ 18. Node 4 is a terminal one while node 3 is divided by considering Low_Emission_Car_21. If Low_Emission_Car_21 $<$ 3.485 the units are assigned to node 7 otherwise to node 8.

Node 2 is divided in node 5 if people live in Centre or South and node 6 if the live in Islands, North East, North West. Node 5 generates two leaves: node 9 if Income_17 $<$ 14657.508 and node A if Income_17 $>$ 14657.508. Node 6 generates two leaves: node B if Population_Density_21 $<$ 109.60 and node C if Population_Density_21 $\geq$ 109.60.

The regression tree provides simple rules for predicting Pm10_20. This can be generalized to different data sets in fact the applications of random forest show a not elevated average squared error both in training and in OOBS.

[3] Missing values are treated using the option of HPSLIT assignmissing = similar.

## 4 Conclusions

In this paper we have analyzed big data, we have also illustrated processes of extraction of information from the data. In order to promote more sustainable urban transport systems it is important to be able to measure and assess the sustainability of present and future transport trends and policies. Yet, there is no general agreement both on the concept of sustainable transport and on which indexes should be used to measure it appropriately, let alone the policies that should be implemented. This comes as little surprise considering that the range of stakeholders involved are numerous and that the impacts that transport has on the economy, the society, and the environment are very complex.

Different indicators of urban transport sustainability referring to Italian provincial towns have been collected.

The statistical methodology used to analyze the Italian urban sustainability is decision trees and random forest to obtain a testing measure.

In more detail: classification trees with dependent variable Macroregion and regression trees with dependent variable PM10 have been used.

The purpose is to see if, on the basis of the previous indicators, there is the possibility of predicting the trend of variables of interest.

The applications of decision trees are in conjunction with random forest. Random forest can to improve the forecasting of regression trees as it is an elaboration of classification and regression trees that average the predictions of a large number of randomly subsampled regression trees so to obtain a more stable solution.

## References

1. ANFIA (associazione nazionale filiera industria automobilistica): Cars In Use By Province And Fuel in 2021 (2021)
2. Bolasco S.: Analisi Multidimensionale dei Dati. Metodi, strategie e criteri d'interpretazione, Carrocci editore, Roma (2005)
3. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classsification and Regression Trees. Taylor & Francis, New York (1984)
4. Breiman, L.: Random Forest. Machine Learning, vol. 45, pp. 5–32, Kluvier Accademic Publisher, Manifactured in The Nederlands (2001)
5. ECMT: Strategy For Integrating Environment and Sustainable Development Into The Transport Policy. Adopted by the Ministers responsible for Transport and Communications at the 2340th meeting of the European Union' Council of Ministers, Luxembourg, April 4–5, 2001 (2001)
6. Gareth, J., Witten, D., Hastie, T., Tibshirani, R.: Introduzione all'apprendimento statistico. In: Salini, S., Gaito, S., Boracchi, P., Ambrogi, F., Manzi, G., Biganzoli, E. (eds) Piccin Editore (2021)
7. ISTAT(a): Nota sulle misure del Benessere Equo e Sostenibile dei territori. Anno 2021. Retrieved from October 2022 (2022) https://www.istat.it/it/files//2019/05/Nota-metodologica.pdf
8. ISTAT(b): Il BES dei territori. https://www.istat.it/en/well-being-and-sustainability/the-measurement-of-well-being/bes-at-local-level Accessed Oct 2022

9. ISTAT(c): BES-2021, Il Benessere Equo e Sostenibile in Italia. https://www.istat.it/it/files//2022/04/BES_2021.pdf. Accessed June 2022
10. Joumard, R., Gudmundsson, H.: Indicators of environmental sustainability in transport. An interdisciplinary approach to methods. INRETS, Lyon, France (2010)
11. Jan, A.K.: Data Clustering. 50 years beyond K-means. Pattern Recogn. Lett. 651–666 (2010)
12. Mayer-Schonberger, V., Cukier, K.: Big Data: A Revolution That Will Transform How We Live, Work, and Think. Mariner Books, Boston (2013)
13. Monte, A., Schoier, G.: A multivariate statistical analysis of equitable and sustainable well-being over time. Soc. Indic. Res. , 1–16 (2020). https://doi.org/10.1007/s11205-020-02392-x
14. Pubblico registro automobilistico Morti e feriti in Incidenti, Accessed 25/11/22 I.Stat
15. Schoier, G., Borruso, G.: A methodology for dealing with spatial big data. Int. J. Bus. Intell. Data Mining **12**(1), 1–13 (2017)
16. Software Testing Help: Data Mining Vs Machine Learning Vs Artificial Intelligence Vs Deep Learning (2019). http://www.intelligenzaartificiale.it/data-mining/. Accessed March 20
17. Zietsman, J., Rilett, L.R. . Sustainable transportation: Conceptualization and performance measures. (No. SWUTC/02/167403-1). Southwest University Transportation Centre, Texas Transportation Institute, College Station, Texas: Texas A & M University (2002)