MAKING THE DECISION ON BUYING SECOND-HAND CAR MARKET USING DATA MINING TECHNIQUES

Assistant PhD. Student Cristina OPREA
Petroleum-Gas University, Ploiesti, Romania
oprea_cris2005@yahoo.com

Abstract:

According to a recent study by GfK Romania, the institute number one of market research in Romania, 55% of Romanians that plan to buy a car and would buy a second hand car.

Given the high demand for such cars, this study is to facilitate the acquisition of second hand cars. This will be achieved through analysis of existing data on the auto market using data mining techniques.

The paper has determined the price of a car using linear regression and the of score each type of car, which, on a scale from 1 to 5, will show whether or not a car worths to be bought. Also there have been identified the most important criteria considered when choosing a car using the algorithm "InfoGainAttributeEval" Ranker search method. The results obtained indicate a strong correlation between price and features set cars: class, fabrication_year, no_km, combustible, engine capacity, registration_statementand and emissions_class.

Keywords: data mining, linear regresion, decision tree, J 48 algoritm, ID3 algoritm.

JEL Classification: C38, C13

INTRODUCTION

Data mining consists of an evolving set of techniques that can be used to extract valuable information and knowledge from massive volumes of data. [Ioniţă, 2005].

Data mining process includes the following activities [Tudor, Carbureanu, 2007]:

- *data selection* aimed retrieval of massive data only for relevant data for analysis;
- *data cleaning* dealing with data cleaning and preparation of the activities that are necessary to ensure accurate results,
- *data transformation*, converts the data into a bidimensional table and eliminates unwanted fields so that the results to be valid;
- extracting patterns from data (*data mining*) is to analyze the data by a suitable set of algorithms to discover patterns and significant rules and to produce predictive models
- *data validation* which requires proper interpretation of the results of data mining and aims to select those models that are valid and useful in future decisions in different areas.

Data mining techniques to discover segments, clusters, subgroups to classify and better understanding of the phenomenon analyzed and implement details of its forecasts evolution.

1. DATA PRESENTATION

The data is supplied by ads in newspapers Raid and Automar for selling second hand cars in Ploieşti and in the country published in May 2010. It was created an Excel database with the criteria considered when choosing a second hand car. Thus, we identified nine attributes, namely:

Table1. The Attributes of the model

Nr.	Attribute	Description	Possible values
1.	combustible	Type of combustible used	0, 1
2.	cylinder_capacity	Engine capacity	1, 2, 3
3.	emissions_class	Class of emissions pollutants	1, 2, 3
4.	endowments	Endowments car: the trunk, air conditioning, airbags, etc.	0, 1
5.	registration_statement	Car registration statement	0, 1

6.	class	Car class	1, 2, 3, 4, 5
7.	fabrication_year	Fabrication year of the car	1, 2, 3
8.	no_km	Number of kilometers	1, 2, 3
9.	price	Price of the car	1, 2, 3

The data quality in terms of attributes used for completing the degree is 99%. *Combustible attribute* values is the type of combustible used by car and can take two possible values, namely 0 for gasoline and 1 for diesel.

Cylinder_capacity attribute is the engine capacity and may take the following values:

Table 2. The value of cylinder_capacity attribute

The values of cylinder_capacity attribute	Cylinder_capacity
1	<=1400 cmc
2	>1400 cmc and <=1800 cmc
3	>1800 cmc

Emissions_class attribute is the class of vehicle classification by pollutant emissions and has a value of 1 if it is Euro 2, 2 if it is Euro 3 and the value 3 if it is Euro 4.

Features attribute refers to the facilities available on the vehicle, e.g. trunk, right passenger airbag, air conditioning etc. If your vehicle has such features, attribute receives a value and 0 if no such facilities.

Registration_statement attribute value is 0 if the vehicle is not registered in Romania and one if it is registered.

Cars are divided into classes. Cars are divided into classes by certain criteria: destination, car body type, weight, length etc. Each class is denoted by a letter in Latin alphabet. Attribute class has the following values:

Table 3. The value of class attribute

The values of class attribute	Class
1	B
2	C
3	D
4	E
5	L

Fabrication_year attribute refers to the year the car was manufactured and may take the following values:

Table 4. The value of fabrication_year attribute

The values of fabrication_year attribute	Fabrication_year
1	<=2000
2	>2000 and <=2005
3	>2005

No_km attribute indicates the number of kilometers the car has and can take the following values:

Table 5. The value of No_Km attribute

The values of no_Km attribute	No_Km
1	<=50000
2	>50000 and <=150000
3	>150000

Price attribute refers to the price offered by the seller and has the following values:

Table 6. The value of price attribute

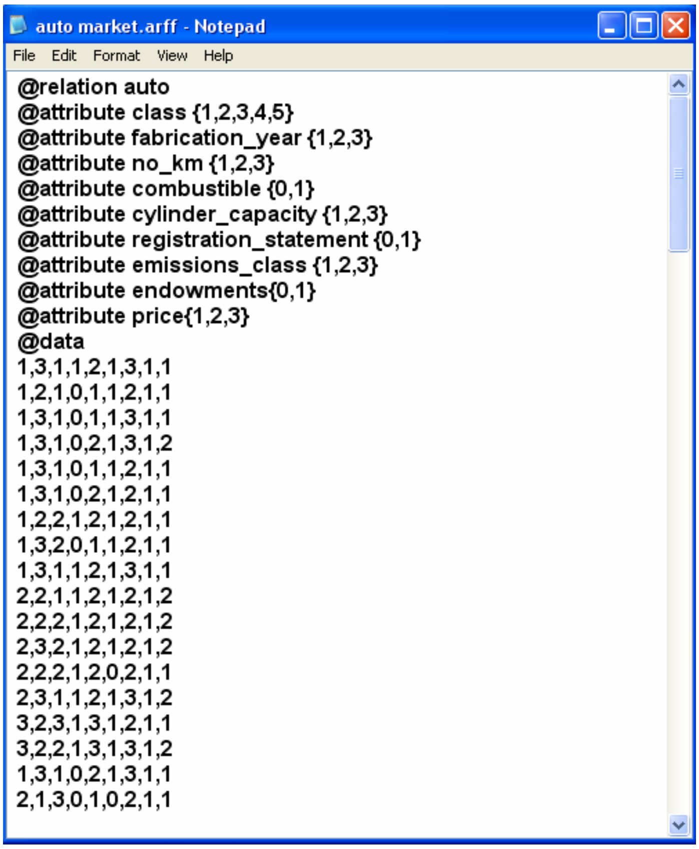
The values of price attribute	Price (Euro)
1	<=5000
2	>5001 and <=10000
3	>10000

Before starting the tests, given that the algorithms selected for analysis established a relationship between input attributes and output attribute (the cause-effect) it was required a modification. Therefore, all instances of the database must contain at least two attributes not null, one for input and one for output, since it is impossible to establish a relationship with a single attribute. As a result, in the database there were deleted all instances that had a single attribute.

2. THE PROPOSED MODEL

As environmental data mining it was used WEKA platform. Weka stands for Waikato Environment for Knowledge analysis (Waikato Environment Knowledge Analysis) software and is a University of Waikato, New Zealand.

Weka is a collection of automatic learning algorithms for data mining in Java. Algorithms can be applied directly on a data set or can be called from code written by the programmer [Oprea, Tudor, Carbureanu, 2007]. Weka contains tools for data preprocessing, classification, regression, clustering, association rules and visualization. Contains a collection of visualization tools and algorithms for data analysis and predictive modeling associated with graphical user interfaces to offer easy access to instruments.



```
@relation auto
@attribute class {1,2,3,4,5}
@attribute fabrication_year {1,2,3}
@attribute no_km {1,2,3}
@attribute combustible {0,1}
@attribute cylinder_capacity {1,2,3}
@attribute registration_statement {0,1}
@attribute emissions_class {1,2,3}
@attribute endowments{0,1}
@attribute price{1,2,3}
@data
1,3,1,1,2,1,3,1,1
1,2,1,0,1,1,2,1,1
1,3,1,0,1,1,3,1,1
1,3,1,0,2,1,3,1,2
1,3,1,0,1,1,2,1,1
1,3,1,0,2,1,2,1,1
1,2,2,1,2,1,2,1,1
1,3,2,0,1,1,2,1,1
1,3,1,1,2,1,3,1,1
2,2,1,1,2,1,2,1,2
2,2,2,1,2,1,2,1,2
2,3,2,1,2,1,2,1,2
2,2,2,1,2,0,2,1,1
2,3,1,1,2,1,3,1,2
3,2,3,1,3,1,2,1,1
3,2,2,1,3,1,3,1,2
1,3,1,0,2,1,3,1,1
2,1,3,0,1,0,2,1,1
```

Figure 1. Source file containing the input dataset

Weka strengths of this package are:

- is available under GNU (General Public License)
- is very portable as it is implemented in Java programming language, language that runs on any platform;
- contains a collection of techniques for preprocessing and data modeling;
- is easy to use even by a beginner because it uses graphical user interfaces.

Dataset used in WEKA programming environment must be in CSV or ARFF to be processed. The data come mostly from an Excel table or a database and must be converted to CSV or ARFF format, the most widely distributed database in text files. Using this format in parallel with direct support for databases is another advantage of WEKA. In addition to these favorable factors characterizing WEKA system, there are some disadvantages, namely that use interface requires learning, understanding algorithms and the interpretation of numerical and graphical results.

In addition, WEKA uses statistical terms instead of using appropriate terms of input (e.g., in economic applications) like other specialized software business and more intuitive for a manager or economist

Dataset used was converted into ARFF format to be processed and it was added a header containing the description of attributes, types and their values. In figure 1 are the attributes that were used for modeling application.

The final form of the database contains 9 attributes which describe the existing information about cars and 329 instances. Analyzing the distribution of car prices according to the 9 attributes we observe the following (Figure 2):

- Only 49 of the 329 machines have higher selling price of 10,000 euros;
- Nearly half of all machines are considered Class B
- Only 75 of all cars are unregistered.
- Very few cars have the production year since lower then 2005.

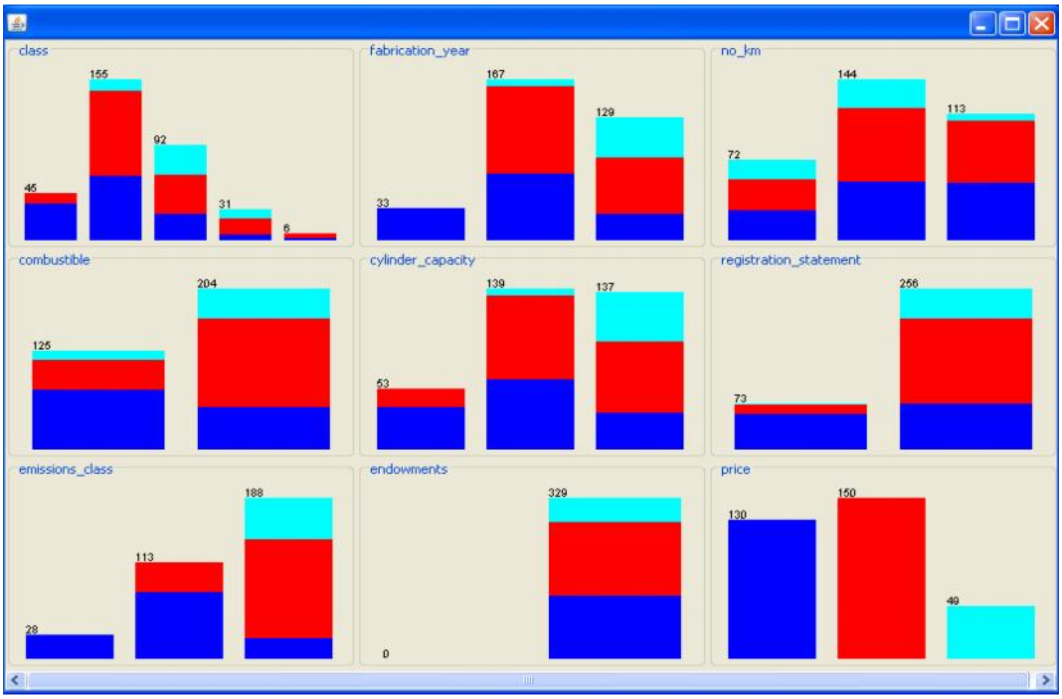
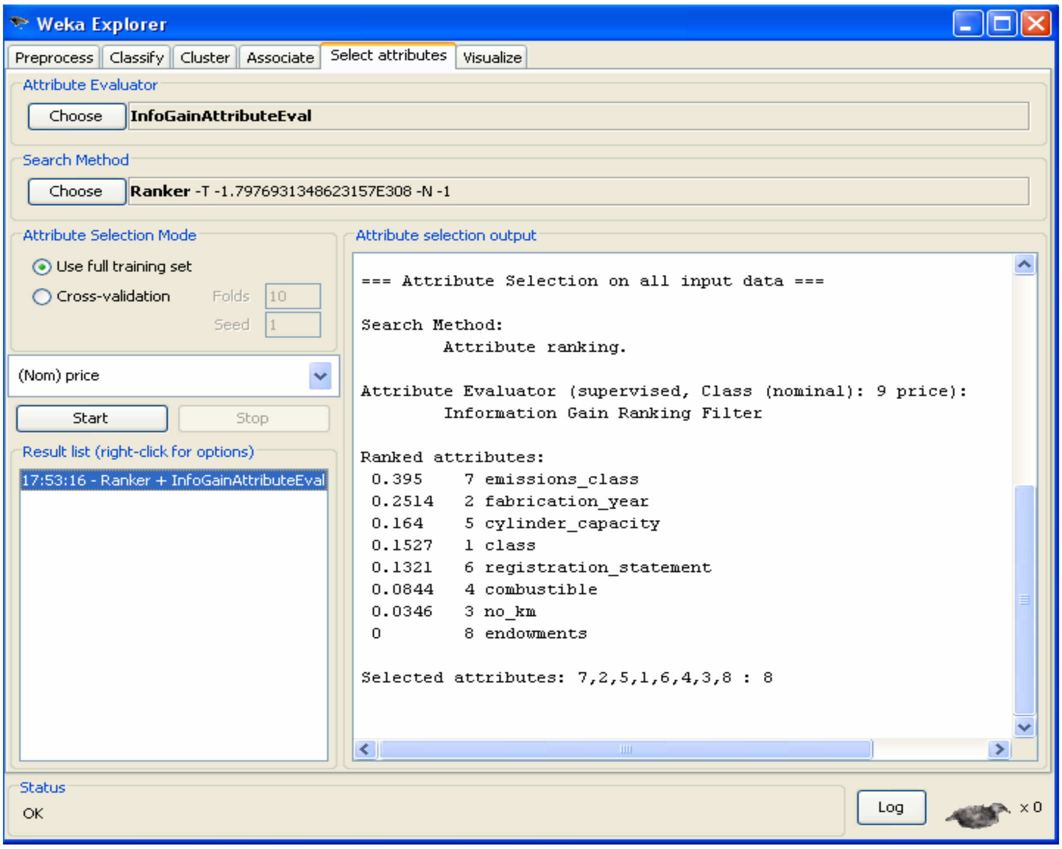


Figure 2. Attribute Distribution

After applying the algorithm "InfoGainAttributeEval" with Ranker search method we notice that the most important attribute to consider for the price of a car is the class attributes, followed by year of manufacture and the number of kilometers traveled.

Features attribute that obtained coefficient of rank 0 will be removed from the model. Eliminating this attribute will lead to growth speed in the model creation and a better accuracy.



Attribute Evaluator

InfoGainAttributeEval

Search Method

Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) price

Result list (right-click for options)

17:53:16 - Ranker + InfoGainAttributeEval

Attribute selection output

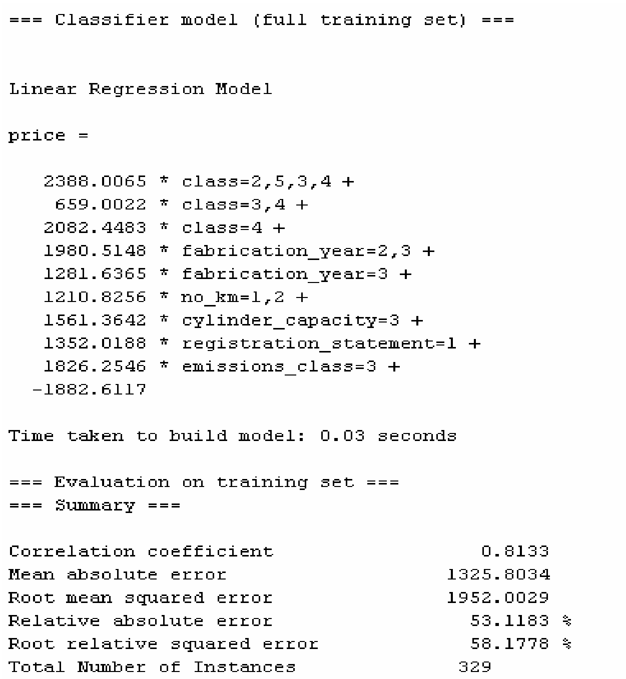
```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 9 price):
        Information Gain Ranking Filter

Ranked attributes:
 0.395     7 emissions_class
 0.2514    2 fabrication_year
 0.164     5 cylinder_capacity
 0.1527    1 class
 0.1321    6 registration_statement
 0.0844    4 combustible
 0.0346    3 no_km
 0         8 endowments

Selected attributes: 7,2,5,1,6,4,3,8 : 8
```

Status

OK

Figure 3. Attribute evaluation results

Between the analyzed characteristics and the cars price is a direct relationship, this being represented by a linear connexion. The corresponding algorithm for a linear connexion between the input attributes and the output attribute is the model of linear regression, model that is found in the set of functional algorithms in Weka package.



```
=== Classifier model (full training set) ===


Linear Regression Model

price =

    2388.0065 * class=2,5,3,4 +
     659.0022 * class=3,4 +
    2082.4483 * class=4 +
    1980.5148 * fabrication_year=2,3 +
    1281.6365 * fabrication_year=3 +
    1210.8256 * no_km=1,2 +
    1561.3642 * cylinder_capacity=3 +
    1352.0188 * registration_statement=1 +
    1826.2546 * emissions_class=3 +
  -1882.6117

Time taken to build model: 0.03 seconds

=== Evaluation on training set ===
=== Summary ===

Correlation coefficient                  0.8133
Mean absolute error                   1325.8034
Root mean squared error               1952.0029
Relative absolute error                 53.1183 %
Root relative squared error             58.1778 %
Total Number of Instances              329
```

Linear regression algorithm implemented in Weka is not limited to numeric attributes, being an excellent method to analyze the training database were 7 attributes determine directly the 8th attribute (price). The most important parameter to the algorithm of linear regression is the attribute selection method with 3 possible options.

At the extreme there are “without selection” method that leads to rapid obtain of some results, but it is less selective, respectively, the “Greedy” method that is considerably slower but with precise results. Between them, there is “M5” method that makes a compromise between speed and accuracy. Though, we must consider a restriction: The “Greedy” method determines the most precise formula, but it needs that the relationship between the input and the output attributes is as close to a linear order to determine the correct local maximum values. If the relationship is not linear, it will lead to an incorrect formula.

By applying linear regression classification algorithm based on complete data (329 records), so method "Greedy" and method "M5" found exactly the same car pricing formula based on its characteristics.

Correlation coefficient was 0,8133, which means that the relation between car characteristics and price can be well approximated with a linear relation. A correlation coefficient of value 1 indicates a perfect linear relation and 0 means that there is no relation between the input and output. The rate of linear regression considered relevant only 6 input attributes of 7.

659,0022+
1281,6365+
1210,8256+
1561+
1352,0188+

1826,2546-
1882,6117=6008,34 Euro

Using the found formula we can easily calculate the price of a car by its characteristics. For example, a car class D, made in the last 5 years, with less then 50000 km, engine capacity greater than 1800 cc, registered, has an estimated price by simple adding the corresponding values for each attribute.

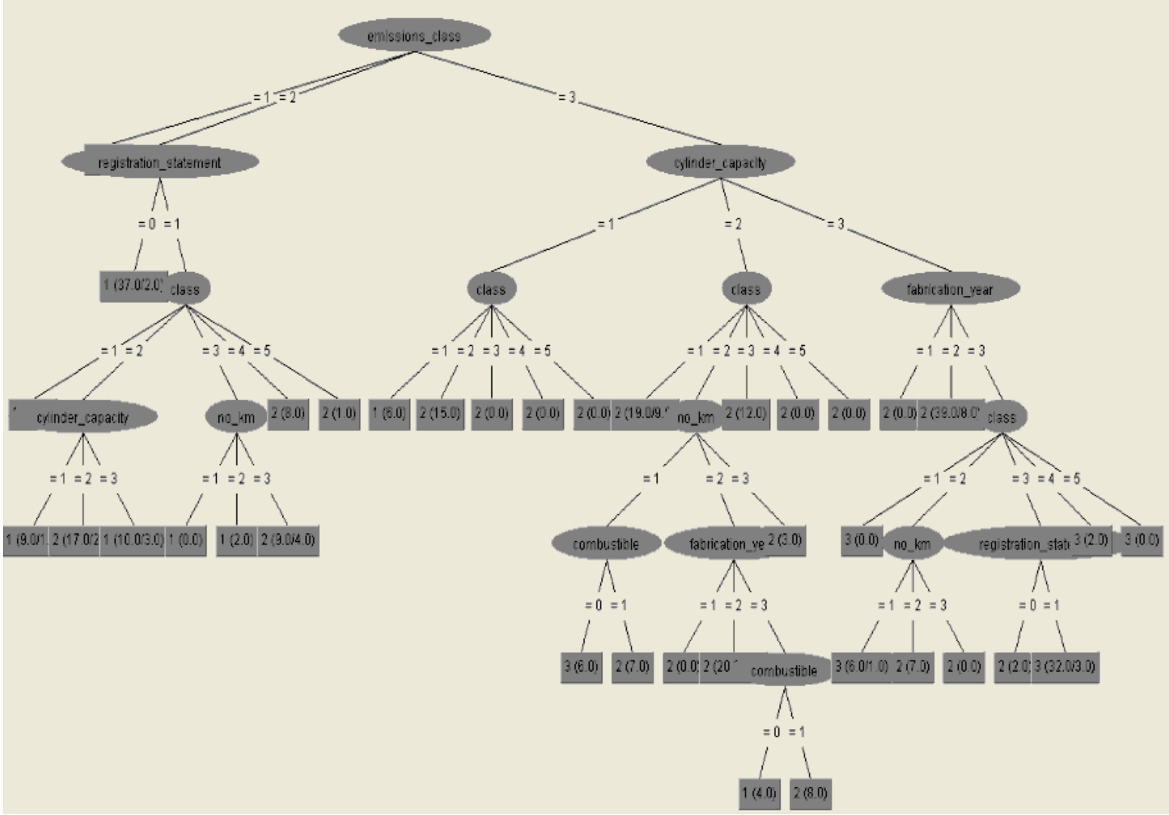


Figure 4. Decision tree algorithm resulting from applying J 48

J.48 algorithm is C4.5 algorithm implementation in Java. and uses top-down inductive method of building decision trees. They are built on testing each node of the tree from the root node for each record. Each node represents an attribute name. it tries placing the instance in an existing class based on common characteristics, evaluating the corresponding attribute for the reached node. Depending on the value, the instance will follow a branch. When there are no more nodes to evaluate, the instances is classified. If a particular class no more obviously differs from a different class by the introduction of more and more records, the two unite, using a process called "pruning."

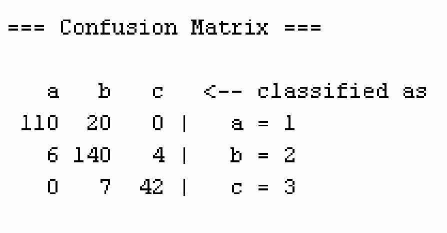
After applying the J48 classification algorithm we achieved an accuracy of 88.75% which means that 292 of 329 instances were classified correctly in the model created.

The value of kappa statistic is 0,8152 and indicates a strong correlation between the attributes analyzed.

Weka generated the following confusion matrix in which the columns indicate the previewed classes (classified), and the rows are the actual classes (real).

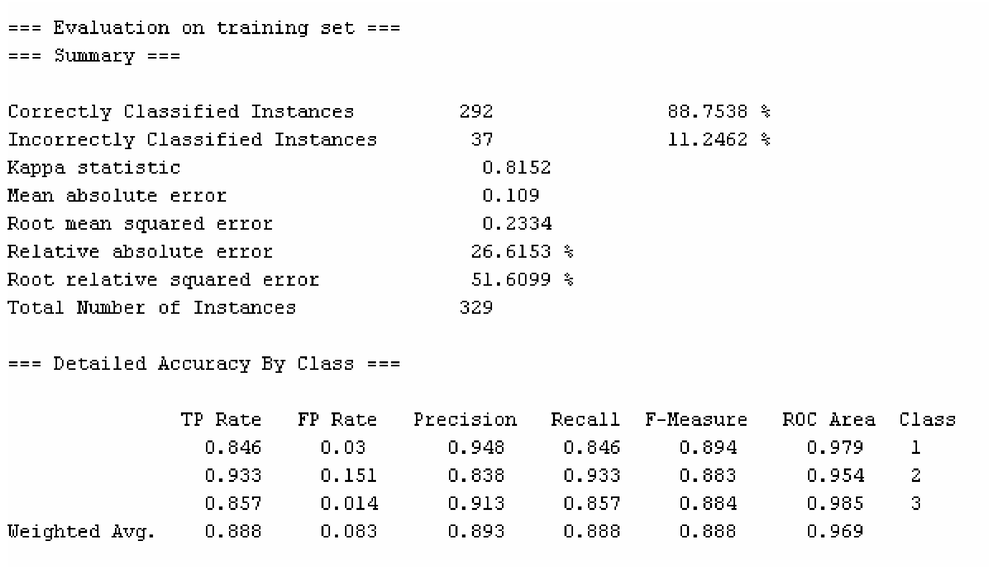
Diagonal matrix indicates correct predictions (110 +140 +42 = 292), and the other elements of the matrix shows incorrect predictions (20 +6 +7 +4 = 37).

The number of correctly predicted values in the total number of predicted values is indicated by the Precision parameter that takes values between 0 and 1. Accuracy equal to 0 indicates that the model doesn't have predictive power, is not conclusive.



```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 110  20   0 |   a = 1
   6 140   4 |   b = 2
   0   7  42 |   c = 3
```

TP rate (true positive rates) is the fraction of positive instances predicted as positive and equals to recall parameter.



```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         292               88.7538 %
Incorrectly Classified Instances        37               11.2462 %
Kappa statistic                          0.8152
Mean absolute error                      0.109
Root mean squared error                  0.2334
Relative absolute error                 26.6153 %
Root relative squared error             51.6099 %
Total Number of Instances              329

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.846     0.03       0.948     0.846     0.894      0.979    1
                 0.933     0.151      0.838     0.933     0.883      0.954    2
                 0.857     0.014      0.913     0.857     0.884      0.985    3
Weighted Avg.    0.888     0.083      0.893     0.888     0.888      0.969
```

Rate parameter FP (False Positive Rate) is the fraction of negative instances predicted as positive.
The confusion matrix, PF Rate is calculated using the formula:

$$FP\_Rate = (suma\_elem\_pe\_coloana - elem\_de\_pe\_diag)/(suma\_elem\_pe\_celelalte\_lin$$

Parameter F-Measure is calculated using the formula:

$$F\_Measure = (2 * TP\_Rate * \Pr ecision)/(TP\_Rate + \Pr ecision)$$

ROC curves (Receiver Operating Characteristic) are widely used in evaluating the results of predictions (forecasts) The Area Under the ROC Curve (AUC) is a measure of performance that encapsulates many of the advantages of ROC curve.

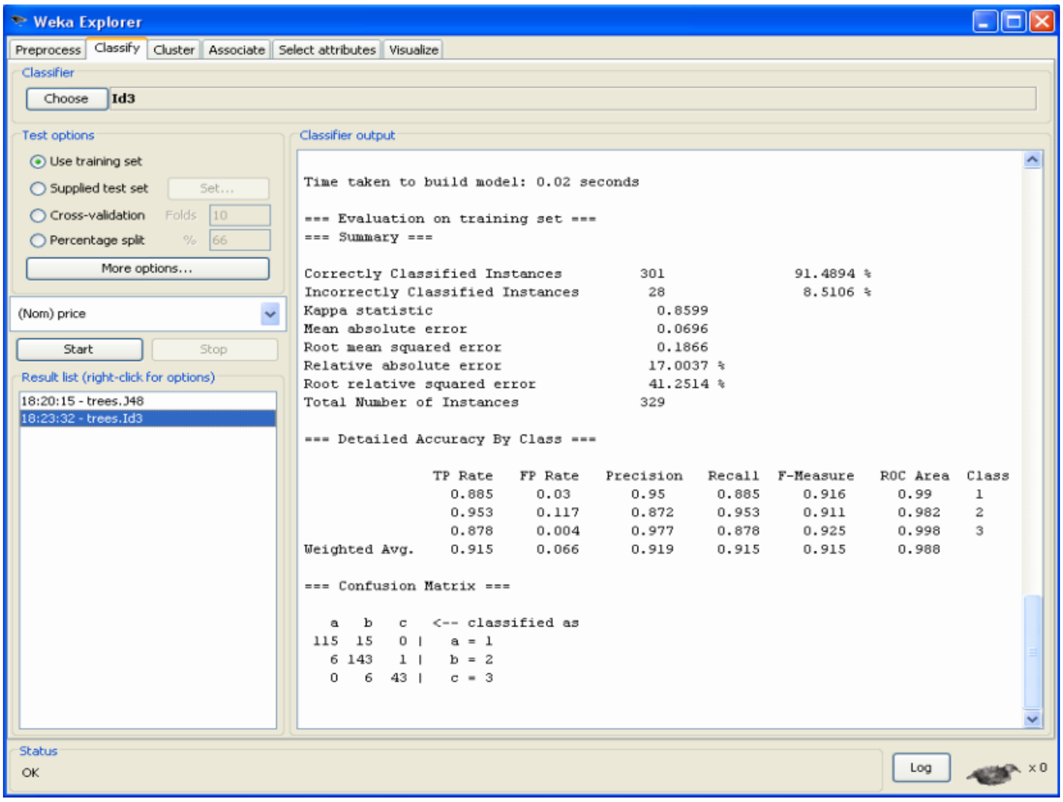
The generated decision tree has as root node the class of pollutant emissions, this attribute being the main mean to differentiate the cars. In the model there were also found as relevant for car differentiation the attributes: class (present in every main brach of the tree), the registration state and the manufacturing year.

For E class cars, the model cannot be successfully applied in the dataset because there are only six cars registered for the class E. The lack of a greater number of records made it impossible to create a viable model for this class.

Example of interpretation of a decision tree branch: "*If class of vehicle emissions is Euro 4, engine capacity greater than 1800 cc, it is Class D, with over 50000 miles on board and year of manufacture between 2000 and 2005, then the price greater then 5000 euros.*"

emissions_class = 3(Euro 4)
| cylinder_capacity = 3(>1800 cmc)
| | class=2(C)
| | | no_km=2 (50000-150000 Km)
| | | | fabrication_year = 1(2000-2005): 2 (5000-10000 Euro)

To make a comparative study it was applied ID3 algorithm. This verifies the number of instances classified correctly, respectively incorrectly.



Id3

● Use training set
○ Supplied test set Set...
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) price

18:20:15 - trees.J48
18:23:32 - trees.Id3

OK

```
Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         301               91.4894 %
Incorrectly Classified Instances        28                8.5106 %
Kappa statistic                          0.8599
Mean absolute error                      0.0696
Root mean squared error                  0.1866
Relative absolute error                 17.0037 %
Root relative squared error             41.2514 %
Total Number of Instances              329

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.885     0.03       0.95      0.885     0.916      0.99     1
                 0.953     0.117      0.872     0.953     0.911      0.982    2
                 0.878     0.004      0.977     0.878     0.925      0.998    3
Weighted Avg.    0.915     0.066      0.919     0.915     0.915      0.988

=== Confusion Matrix ===

   a   b   c   <-- classified as
 115  15   0 |   a = 1
   6 143   1 |   b = 2
   0   6  43 |   c = 3
```

Figure 5. Results obtained from applying the ID3 algorithm

The results obtained can be viewed in Figure 5. In the example shown, the case of ID3, achieved an accuracy of 91.48% which means that 301 of 329 instances were classified correctly and 28 incorrectly in the created model.

The value of kappa statistic is 0.8599 and indicates a strong correlation between the attributes analyzed.

Weka generated following confusion matrix.

Confusion matrix columns indicate the predicted class (classified), and the raws, the existing classes (real). Class 1 contains 130 instances, class 2, 149 instances, class 3, 49 instances.

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 115  15   0 |   a = 1
   6 143   1 |   b = 2
   0   6  43 |   c = 3
```

Diagonal matrix indicates correct predictions (115 +143 +43 = 301), and other elements of the matrix shows incorrect predictions (15 +6 +6 +1 = 28). The number of correctly predicted values in the total number of predicted values is stated in the Precision parameter that indicates that the proposed model has predictive power and is conclusive.

CONCLUSIONS

This study shows that data mining is an instrument of analyze very strong that allows extracting new information regarding second hand cars prices by their characteristics and using these information in buying decision.

Based on a set of learning it was built a model that can be applied to estimate second car price as to identify defining characteristics client car buying decision.

The obtained results indicate a string relation between the car price and the characteristics presented: Class, year of fabrication, number of km, engine capacity, registration, and emissions. The most important attributes for cr price analyze are class, manufacturing year and number of km.

The study can be extended be including others factors, such as interest rate.

REFERENCES

1. Gorunesu, F., *„Data Mining. Concepte, modele şi tehnici”*, Editura Albastră, Cluj-Napoca, 2006.
2. Ioniţă, A., „*Asupra termenului de data mining*”, Revista Română de Informatică şi Automatică, vol. 15, nr. 2, 2005.
3. Leontin, T. L., Moldovan, D., Rusu, M., Secară, D., Trifu, C., „*Data mining on the real estate market*”, Revista Informatica Economică, nr. 4 (36)/2005
4. Oprea Cristina, Zaharia, M., Gogonea, M., “ *Analysis of student performance to license exam using data mining techniques*”, The 14th IBIMA Conference on Global Business Transformation through Innovation and Knowledge Management, Istanbul, 2010
5. Oprea Cristina, Zaharia, M., Enăchescu, D., „K*nowledge discovery using data mining techniques: A case study*”, „Annual Session Of Scientific Papers IMT”, Oradea, 2010”
6. Oprea, M., *„Sisteme bazate pe cunoştinţe Ghid teoretic şi practic”*, Matrix ROM, Bucureşti, 2002
7. Oprea, M., Tudor, I., Cărbureanu, M., *„Prediction of Student Professional Evolution with Data Mining Techniques”*, The eight international conference on informatics in economy, Bucharest, Romania, May 17-18, 2007
8. Tudor, I., Cărbureanu, M., *Tehnici de data mining în managementul cunoaşterii într-o universitate*, în Managementul cunoaşterii în universitatea modernă, coord.: Bodea C.N. şi Andone I., Editura ASE Bucureşti, 2007, p.293.
9. Zaharia M., Gogonea R. M., *Econometrie. Elemente fundamentale*, Editura Universitară, Bucureşti, 2009.
10. http://www.cs.waikato.ac.nz/~ml/weka/