

The Applicability Of Data Mining Techniques In Eurostat Databases: An Example Of The Decision Tree

1. Introduction

As a result of the rapid development of technology, infrastructure that is used for obtaining and storing data has also been developed continuously. Besides that, importance of knowledge as an indispensable element for individuals and institutions has increased each passing day. However, former data management techniques have become insufficient for mass of data which increases rapidly. Therefore, there was a need for new methods. Data mining is a field that emerges variety of data extraction techniques to meet these requirements.

Eurostat is a statistical office of European Union. Its purpose is to serve objective and accurate data to decision makers. These statistics are open for everyone to use. Although Eurostat databases are very comprehensive and useful, it is particularly difficult to find academic publications related to data mining.

In this study, it is intended to do data mining study using the statistics that provided by Eurostat. In order to accomplish the analysis, “Information Society” field was selected and the data was analyzed with using a decision tree algorithm of data mining. In the end, the analysis results were presented. It is also intended to shed some light to the next studies.

1.1 Data Mining

Data mining is a process of knowledge acquisition from large databases. It is used to predict the future trends and find behaviors behind expectations that experts may miss. Data mining, is a part of large process called knowledge discovery (Figure 1). Specifically, in data mining step, advanced statistical analysis and modeling techniques are applied to the data to find useful patterns and relationships (Hudairy, 2004, 1).

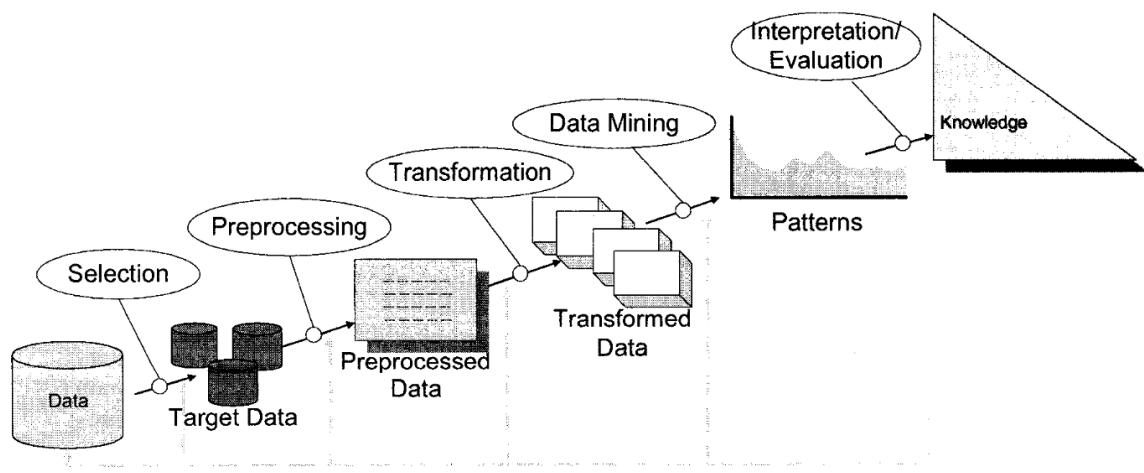


Figure 1. Knowledge Discovery Process (Hudairy, 2004)

Data mining solves user defined problems automatically by using new set of algorithmic modeling. Each of these strong tools are intuitive, easily explainable, understandable and useful (Garango & Raggard, 1999, 81). Data mining is a huge filed contains artificial intelligence, computer science, machine learning, database management, data visualization, mathematic and statistics (Liao, 2003, 157).

1.2 Decision Tree

Decision tree is a data mining approach that mainly used for classification and prediction. Other methods such as neural networks are also used for classification, but decision tree supports decision maker to verify and justify their decisions by the advantages of interpretation and understanding easily. Additionally, decision tree can analyze different type of data without need to predict underlying distribution (Chien & Chen, 2008, 282). Decision tree technique is a very popular technique in data mining tasks. It is used successfully for health, finance, marketing, human resources, sport, telecommunication and etc. (Birant, 2011, 48). A decision tree describes which factors must be taken into account and how each factor is related to different outputs of the decision by historic. Decision tree has value for both predictive and descriptive models (Bounsaythip & Rinta-Runsala, 2001, 20).

A decision tree is a mapping description technique that consists of tests and attributes which are connected to two or more subtrees and leafs, or decision nodes labeled with a class name (Zorman et al., 2001, 110). Decision tree classifiers starts with training set related to class labels. Root node is the main property. Each inside node symbolizes the test attributes, each branch represents the results of test and each leaf shows class labels. In order to define an unknown sample, tree classifier follows path which starts from rood node to leaf nodes which hold class labels (Sathyadevan & Nair, 2015, 550).

The following decision tree example is for the concept `buy_computer` that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node represents a class (Tutorials Point, 2016).

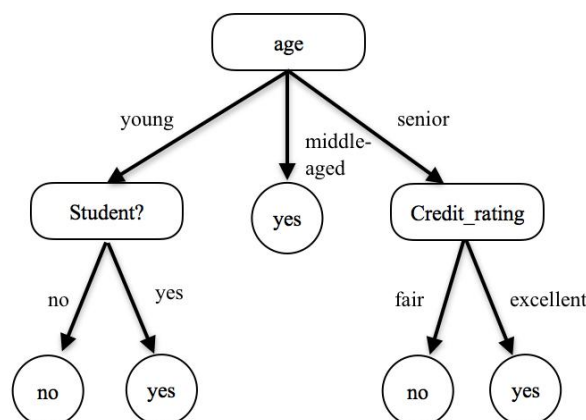


Figure 2. Decision Tree Sample

Decision tree has some advantages; such as it can produce a model that has rules understandable easily and representative, and also can produce highly informative outputs. In artificial intelligence and decision support systems, this technique can be used for both continuous and categoric attributes, but it is available for predict only categorical outputs (Birant, 2011, 48).

2. Method

2.1 Data Selection

In this study, statistics about Information Society is focused. These statistics track the usage of Information and Communications Technologies, which has been one of the main drivers of changes within society and businesses for more than a decade. The regulation contains two modules, covering (Eurostat, 2016):

- Enterprises
- Households and individuals

Data about these two modules can be downloaded as MS Access database format. And also can be accessed via web Interface. From the web interface (Figure 3), “Activities via internet not done because of security concerns” subject was selected to be analyzed.

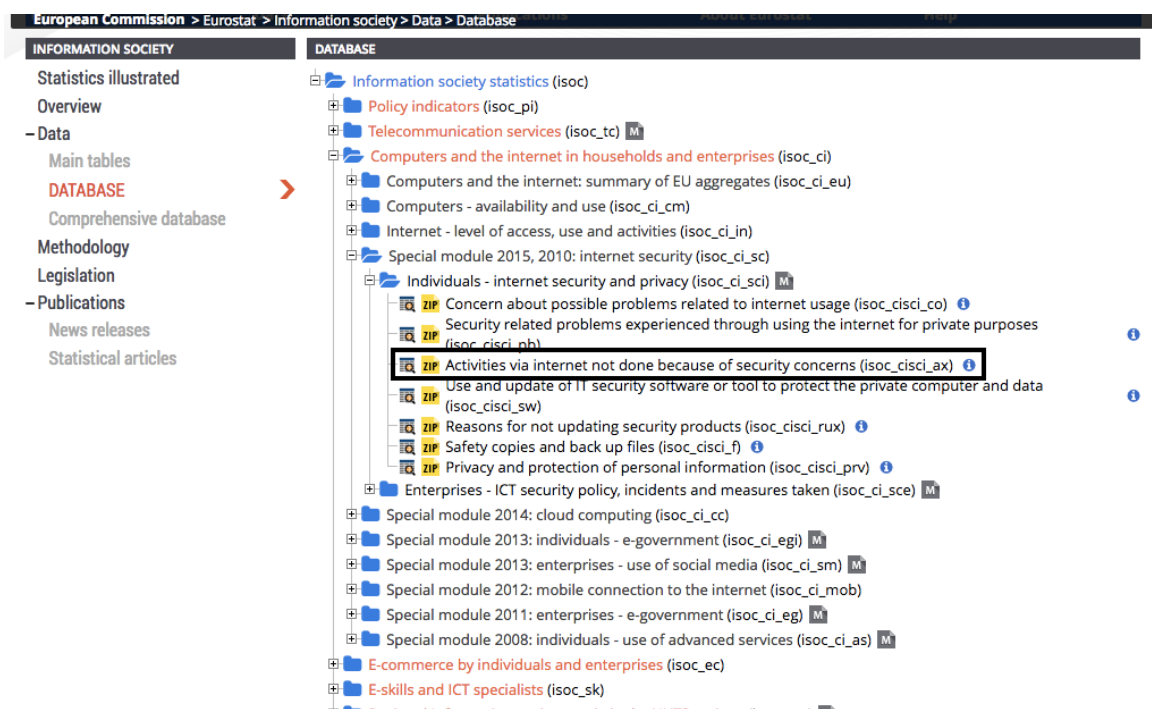


Figure 3. Selection of Statistics

2.2 Data Mining Software: WEKA

As data mining software, Weka was used to do decision tree analysis. Weka has many features. Some features are preprocessing, classification, regression, clustering, association modules and etc. Weka is a free software that has GNU General Public License. Weka is a set of machine learning algorithms for data mining tasks (Machine Learning Group at the University of Waikato, 2016). Weka is also have advanced features, such as ability of producing new modules for programmers, or adding new tools which created by developers.

2.3 Attribute Selection

There are 6 attributes selected for the analysis. These attributes were inferred from statistical data from the topic “Activities via internet not done because of security concerns”. One of them must be used as a class attribute for the classification method. And the other attributes are used as decision variables. These 6 attributes are listed below:

- Year: There are two years (2010, 2015) in the statistics.
- Geographical Region: There are 30 geographical regions listed in the Table 1.
- Types of Security Concern: 5 different security concerns are taken into account. They listed together with the corresponding labels:
 - I_SBGGOOD - Security concerns kept individual from ordering or buying goods or services for private use
 - I_SBBANK - Security concerns kept individual from carrying out banking activities such as account management
 - I_SBPERS - Security concerns kept individual from providing personal information to online communities for social and professional networking
 - I_SBGGOV - Security concerns kept individual from communicating with public services and administrations
 - I_SBSOFT - Security concerns kept individual from downloading software, music, video files, games or other data files
- Gender: Male, Female
- Age Group: There are three age groups (16-24, 25-54 and 55-74)
- Percentage of Individuals: Percentage of individuals related to activities via internet not done because of security concerns

Table 1. List of Geographical Regions

BE - Belgium	HR - Croatia	PL - Poland
BG - Bulgaria	IT - Italy	PT - Portugal
CZ - Czech Republic	CY - Cyprus	RO - Romania
DK - Denmark	LV - Latvia	SI - Slovenia
DE - Germany	LT - Lithuania	SK - Slovakia
EE - Estonia	LU - Luxembourg	FI - Finland
IE - Ireland	HU - Hungary	SE - Sweden
EL - Greece	MT - Malta	UK - United Kingdom

ES - Spain	NL - Netherlands	NO - Norway
FR - France	AT - Austria	TR - Turkey

2.4 Preprocessing

To do the analysis, data need to be preprocessed. Weka has integrated preprocessing module that has many filters to be used. In order to use decision tree algorithm, class value has to be determined. As a class value, percentage of individuals' attribute is used. However, this attribute has many different values, and these have to be limited. This can be done by discretization filter that one of the preprocessing filters of Weka software.

Discretization is the process of putting values into buckets so that there are a limited number of possible states. The buckets themselves are treated as ordered and discrete values (Microsoft Developer Network, 2016). As can be seen in Figure 4, value "2" was determined for bins parameter and value "true" was determined for useEqualFrequency parameter. Regarding the attribute, two classes with equal frequencies were created according to the values.

One class is determined as below 9.5, and the other is above 9.5. With reference to this, there are relatively two classes can be named as low and high rank. The distinction depends on researchers' choice. Different classes can be created by using different parameters.

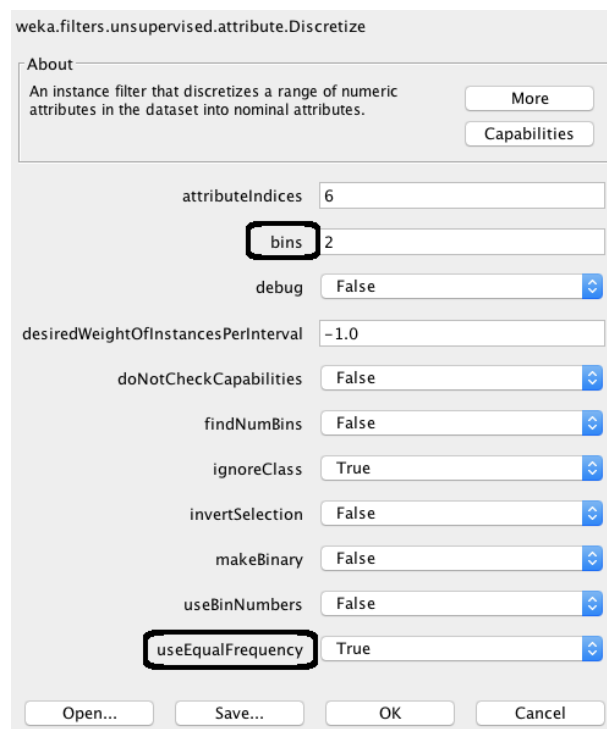
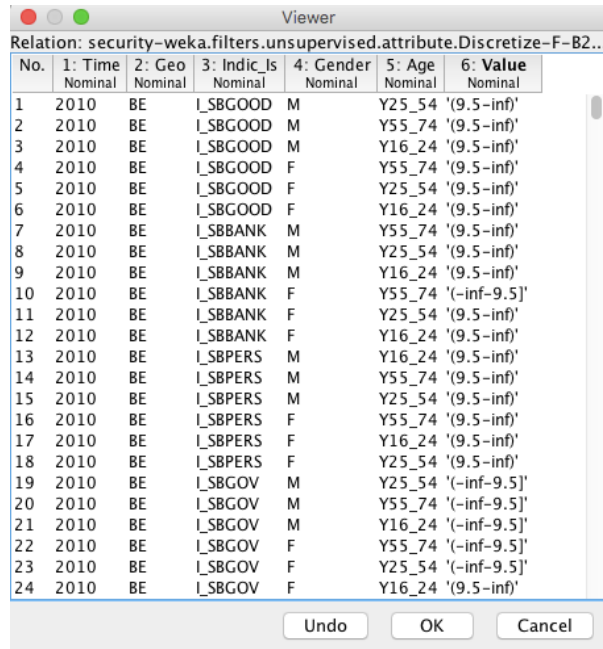


Figure 4. Discretization

After discretization process, variables of “Value” attribute were converted two types of variables as can be seen in Figure 5.



No.	1: Time Nominal	2: Geo Nominal	3: Indic_Is Nominal	4: Gender Nominal	5: Age Nominal	6: Value Nominal
1	2010	BE	I_SBGGOOD	M	Y25_54	'(9.5-inf)'
2	2010	BE	I_SBGGOOD	M	Y55_74	'(9.5-inf)'
3	2010	BE	I_SBGGOOD	M	Y16_24	'(9.5-inf)'
4	2010	BE	I_SBGGOOD	F	Y55_74	'(9.5-inf)'
5	2010	BE	I_SBGGOOD	F	Y25_54	'(9.5-inf)'
6	2010	BE	I_SBGGOOD	F	Y16_24	'(9.5-inf)'
7	2010	BE	I_SBBANK	M	Y55_74	'(9.5-inf)'
8	2010	BE	I_SBBANK	M	Y25_54	'(9.5-inf)'
9	2010	BE	I_SBBANK	M	Y16_24	'(9.5-inf)'
10	2010	BE	I_SBBANK	F	Y55_74	'(-inf-9.5]'
11	2010	BE	I_SBBANK	F	Y25_54	'(9.5-inf)'
12	2010	BE	I_SBBANK	F	Y16_24	'(9.5-inf)'
13	2010	BE	I_SBPERS	M	Y16_24	'(9.5-inf)'
14	2010	BE	I_SBPERS	M	Y55_74	'(9.5-inf)'
15	2010	BE	I_SBPERS	M	Y25_54	'(9.5-inf)'
16	2010	BE	I_SBPERS	F	Y55_74	'(9.5-inf)'
17	2010	BE	I_SBPERS	F	Y16_24	'(9.5-inf)'
18	2010	BE	I_SBPERS	F	Y25_54	'(9.5-inf)'
19	2010	BE	I_SBGGOV	M	Y25_54	'(-inf-9.5]'
20	2010	BE	I_SBGGOV	M	Y55_74	'(-inf-9.5]'
21	2010	BE	I_SBGGOV	M	Y16_24	'(-inf-9.5]'
22	2010	BE	I_SBGGOV	F	Y55_74	'(-inf-9.5]'
23	2010	BE	I_SBGGOV	F	Y25_54	'(-inf-9.5]'
24	2010	BE	I_SBGGOV	F	Y16_24	'(9.5-inf)'

Figure 5. Data List After Discretization Filter

2.5 Application

After the necessary arrangements, decision tree analysis can be done. There are a lot of decision tree algorithms to be utilized. These algorithms vary depending on types of variables and creation steps of decision tree. The algorithm C4.5, which is one of the most widely used algorithms was used in this study. Weka software equivalent of this algorithm is J48. This model was run using the J48 algorithm with 10 folds cross-validation parameters.

3. Findings

As a result of the conducted analysis, there is an analysis summary obtained as Figure 6. One of the most important points is the value of “Correctly Classified Instances”. The value 86.45% means that approximately 86 percent of individuals were classified correctly by the model. Another important point is “Confusion Matrix”. This matrix gives the number of correctly classified and misclassified instances by class labels. The results also provide Kappa statistics. Kappa statistics gives agreement between predicted and true class. In this study, the value of Kappa statistics is 0.729. This means that there is a strong agreement between the predicted and true class.

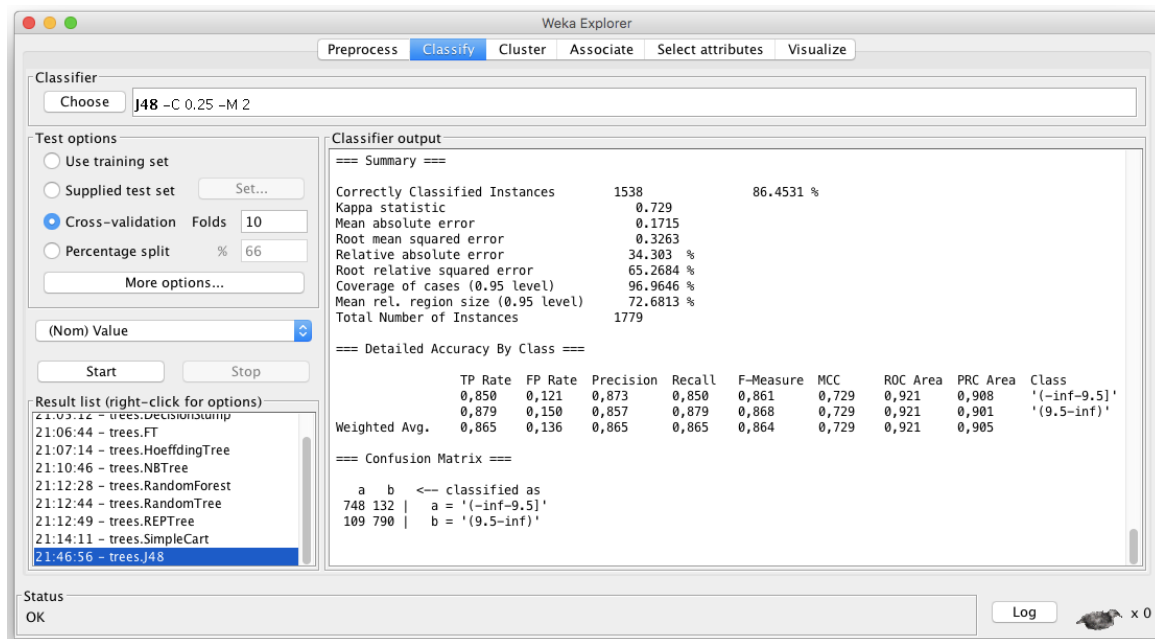


Figure 6. Analyze Results Window

A decision tree was constructed by the end of the analysis. Beginning part of the decision tree can be seen in text format as follow:

Geo = BE: '(9.5-inf)' (60.0/9.0)

Geo = BG

| Age = Y16_24

| | Indic_Is = I_SBGGOOD: '(9.5-inf)' (4.0)

| | Indic_Is = I_SBBANK: '(9.5-inf)' (4.0)

| | Indic_Is = I_SBPERS: '(9.5-inf)' (4.0)

| | Indic_Is = I_SBGOV: '(-inf-9.5]' (4.0)

| | Indic_Is = I_SBSOFT

| | | Time = 2010: '(9.5-inf)' (2.0)

| | | Time = 2015: '(-inf-9.5]' (2.0)

| Age = Y25_54

| | Indic_Is = I_SBGGOOD: '(9.5-inf)' (4.0)

| | Indic_Is = I_SBBANK: '(9.5-inf)' (4.0)

| | Indic_Is = I_SBPERS: '(9.5-inf)' (4.0)

| | Indic_Is = I_SBGOV: '(-inf-9.5]' (4.0)

```

| | Indic_Is = I_SBSOFT: '(-inf-9.5]' (4.0)
| Age = Y55_74: '(-inf-9.5]' (20.0)
Geo = CZ: '(-inf-9.5]' (58.0/1.0)
Geo = DK
| Indic_Is = I_SBGGOOD: '(9.5-inf)' (12.0)
| Indic_Is = I_SBBANK: '(9.5-inf)' (12.0/1.0)
| Indic_Is = I_SBPERS: '(9.5-inf)' (12.0)
| Indic_Is = I_SBGOV: '(-inf-9.5]' (12.0)
| Indic_Is = I_SBSOFT: '(9.5-inf)' (12.0)
Geo = DE: '(9.5-inf)' (60.0/9.0)
Geo = EE
| Indic_Is = I_SBGGOOD
| | Age = Y16_24
| | | Time = 2010: '(-inf-9.5]' (2.0)
| | | Time = 2015: '(9.5-inf)' (2.0)
| | Age = Y25_54: '(9.5-inf)' (4.0)
| | Age = Y55_74: '(-inf-9.5]' (4.0/1.0)
| Indic_Is = I_SBBANK: '(-inf-9.5]' (12.0)
| Indic_Is = I_SBPERS: '(9.5-inf)' (12.0/3.0)
| Indic_Is = I_SBGOV: '(-inf-9.5]' (12.0)
| Indic_Is = I_SBSOFT
| | Time = 2010
| | | Age = Y16_24: '(-inf-9.5]' (2.0)
| | | Age = Y25_54: '(9.5-inf)' (2.0)
| | | Age = Y55_74: '(-inf-9.5]' (2.0)
| | Time = 2015: '(9.5-inf)' (6.0)

```

When the decision tree is analyzed, it can be seen that the initial distinction is geographical region. This means that, region has more prior distinction to the other attributes. Each region has own unique characteristics. For example, first region is BE (Belgium). This region appears to remain in the second class, and doesn't have another clear distinction. However, it is primarily subjected to distinction based on age groups with regard to the region coded BG (Bulgaria). It is observed that individuals of the same age group generally have similar

characteristics. After the age distinction, classes are divided according to the type of security concerns.

Weka has a feature of visualization of decision trees. The decision tree scheme created by the feature can be seen in Figure 7. Decision makers can benefit from a decision tree diagram by easily detect the current situation and give more accurate decisions. For example, administrators of multinational companies are able to understand easily the characteristics of countries where they intend to activate web services.

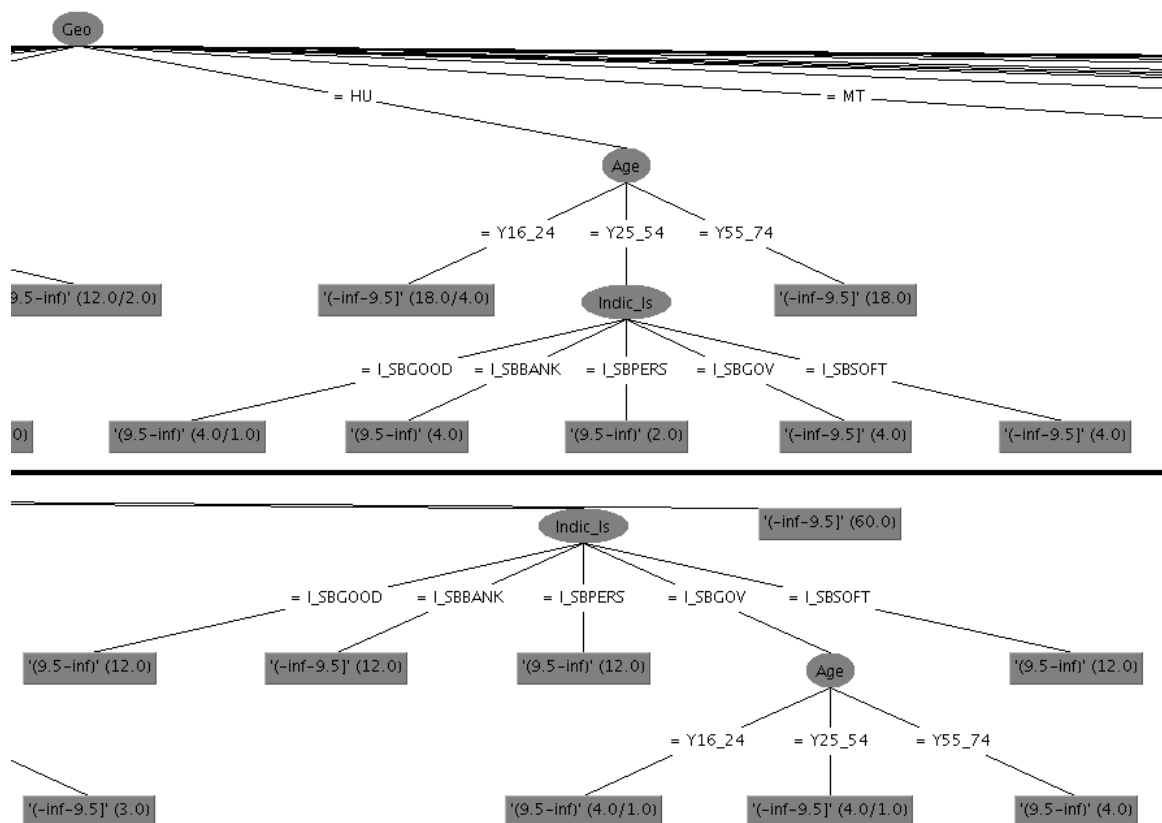


Figure 7. Two Parts of the Visualized Decision Tree

4. Conclusion

Data Mining is a new area include a lot of techniques that obtain meaningful information from large databases, today. Most important feature of data mining is facilitation to find unknown, unpredictable patterns and interrelations. Decision tree is one of the classification techniques of data mining. Decision tree represents an informative map related to the decisions which belongs to the decision makers' responsibility. There are some advantages of decision tree. It can be applied easily, and produce easily understandable models. Besides, it can generate a model without requiring to know underlying distributions of variables.

Data Mining can be applied in many fields, such as banking, marketing, human resources, etc. The data needed to perform data mining applications can be obtained from

many different places. One of the resources supports decision makers to make decision is Eurostat. Eurostat builds databases containing statistical data about many views of European Union. These databases are open for public and free.

In this study, decision tree analysis, one of the data mining techniques, was performed with Eurostat data. In the title of the Information Society, data of “Activities via internet not done because of security concerns” topic in “Households and Individuals” category was chosen for the analysis. In the result of the analysis, a decision tree was established by C4.5 decision tree algorithm (J48 in Weka software) with the accuracy rate of 86.45%.

In conclusion, the applicability of data mining techniques in Eurostat databases with an example of the decision tree was analyzed. The most two important issues of the feasibility of the analysis are:

1. Multidimensional data: More attributes make available more comprehensive analysis. Eurostat has the possibility to multidimensional data, that both can be reached over the web page or can be obtained by queries from MS Access database.
2. Categorical class values: Limiting the number of different values is used as the feature of the class variable. In this study, discretization filter was used to determine the number of classes. The number of classes and frequencies can be change regarding to the purpose of the study, or researchers' choices.

For the future researches, micro data analysis can be run thanks to the opportunity offered by Eurostat. Eurostat allows access to micro data for academic research. This study is also intended to give an idea for further studies.

5. References

- Birant, D. (2011). Comparison of Decision Tree Algorithms for Predicting Potential Air Pollutant Emissions with Data Mining Models. *Journal of Environmental Informatics*, 17(1).
- Bounsaythip, C., & Rinta-Runsala, E. (2001). Overview of data mining for customer behavior modeling. *VTT information Technology*, 18, 1-53.
- Chien, C. F., & Chen, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with applications*, 34(1), 280-290.
- Eurostat. (2016). *Overview - EuroStat*. Retrieved from <http://ec.europa.eu/eurostat/web/information-society/overview>
- Gargano, M. L., & Raggad, B. G. (1999). Data mining-a powerful information creating tool. *OCCLC Systems & Services: International digital library perspectives*, 15(2), 81-90.
- Hudairy, H. (2004). Data mining and decision making support in the governmental sector. *Kentucky: Master Thesis, Louisville University*, 1-5.
- Liao, S. H. (2003). Knowledge management technologies and applications—literature review from 1995 to 2002. *Expert systems with applications*, 25(2), 155-164.
- Machine Learning Group at the University of Waikato. (2016). *Weka 3: Data Mining Software in Java*. Retrieved from <http://www.cs.waikato.ac.nz/ml/weka/>

- Microsoft Developer Network. (2016). *Discretization Methods (Data Mining)*. Retrieved from <https://msdn.microsoft.com/tr-tr/library/ms174512%28v=sql.110%29.aspx?f=255&MSPPErr=-2147217396>
- Sathyadevan, S., & Nair, R. R. (2015). Comparative Analysis of Decision Tree Algorithms: ID3, C4. 5 and Random Forest. In *Computational Intelligence in Data Mining-Volume 1* (pp. 549-562). Springer India.
- Tutorials Point. (2016). *Data Mining - Decision Tree Induction*. Retrieved from http://www.tutorialspoint.com/data_mining/dm_dti.htm
- Zorman, M., Podgorelec, V., Kokol, P., Peterson, M., Šprogar, M., & Ojsteršek, M. (2001). Finding the right decision tree's induction strategy for a hard real world problem. *International journal of medical informatics*, 63(1), 109-121.