2006

Namık Kemal University

Institute of Social Sciences

No: 06 / 2012

Comparison of the Economical Indicators of Turkey and European Union States via Decision Tree Method

Dilek ALTAŞ

Vildan GÜLPINAR

SOSYAL BİLİMLER METİNLERİ

Papers on Social Science

Süreli Hakemli Dergi

ISSN 1308–4453 (Print)
ISSN 1308–4895 (Internet)

Sahibi/ Owner: Prof. Dr. Osman ŞİMŞEK- Rektör
Namık Kemal Üniversitesi Adına

Baş Editör/ Editor in Chief: Doç. Dr. Ahmet KUBAŞ
Sosyal Bilimler Enstitüsü Müdürü

Sosyal Bilimler Metinleri Namık Kemal Üniversitesi Sosyal Bilimler Enstitüsü tarafından online ve basılı olarak sosyal bilimlerin farklı alanlarında yapılan çalışmaların duyurulması ve kamu oyu ile paylaşılarak tartışmaya açılmasına yönelik olarak yayınlanan, farklı üniversitelerdeki öğretim üyelerinden oluşmuş Hakem Kuruluna sahip, ***ASOS***, ***ZDB***, ***PROQUEST*** *ve* ***Index Copernicus*** tarafından indekslenen **uluslararası, akademik hakemli ve süreli** bir yayındır. Çalışmada öne sürülen görüş ve düşünceler yazara ait olup Namık Kemal Üniversitesi Sosyal Bilimler Enstitüsünü bağlamaz.

İndirme Adresi:
http://sosyalbe.nku.edu.tr/

Namık Kemal Üniversitesi
Sosyal Bilimler Enstitüsü
Değirmenaltı Yerleşkesi
TR-59030 Tekirdağ
Tel: +90-282-250 4500
Faks: +90-282-250 9932
E-Posta: sosyalbilimler@nku.edu.tr

Hakem Kurulu

Yusuf ALPER	Prof. Dr.	Uludağ Üniversitesi
Sudi APAK	Prof. Dr.	Beykent Üniversitesi
Neşe ATİK	Prof. Dr.	Namık Kemal Üniversitesi
Hasan BOYNUKARA	Prof. Dr.	Namık Kemal Üniversitesi
Tankut CENTEL	Prof. Dr.	Koç Üniversitesi
Toker DERELİ	Prof. Dr.	Işık Üniversitesi
Nadir DEVLET	Prof. Dr.	İstanbul Ticaret Üniversitesi
Ayten ER	Prof. Dr.	Gazi Üniversitesi
Nalan GÜREL	Prof. Dr.	Marmara Üniversitesi
İsmail Hakkı İNAN	Prof. Dr.	Namık Kemal Üniversitesi
Abdülkadir IŞIK	Prof. Dr.	Namık Kemal Üniversitesi
Cem KILIÇ	Prof. Dr.	Gazi Üniversitesi
Derman KÜÇÜKALTAN	Prof. Dr.	Trakya Üniversitesi
Thomas LOPEZ GUZMAN	Prof. Dr.	Cordoba Universitesi
Ahmet MAKAL	Prof. Dr.	Ankara Üniversitesi
Ahmet SELAMOĞLU	Prof. Dr.	Kocaeli Üniversitesi
Ali Nazım SÖZER	Prof. Dr.	Dokuz Eylül Üniversitesi
Yaşar ŞENLER	Prof. Dr.	Namık Kemal Üniversitesi
Can TUNCAY	Prof. Dr.	Bahçeşehir Üniversitesi
Devrim ULUCAN	Prof. Dr.	Maltepe Üniversitesi
Rasim YILMAZ	Prof. Dr.	Namık Kemal Üniversitesi
Levent AKIN	Doç. Dr.	Ankara Üniversitesi
Şener BAĞ	Doç. Dr.	Namık Kemal Üniversitesi
Süleyman BAŞTERZİ	Doç. Dr.	Ankara Üniversitesi
Petru GOLBAN	Doç. Dr.	Namık Kemal Üniversitesi
Alpay HEKİMLER	Doç. Dr.	Namık Kemal Üniversitesi
Aşkın KESER	Doç. Dr.	Kocaeli Üniversitesi
Ahmet KUBAŞ	Doç. Dr.	Namık Kemal Üniversitesi
Hakan ONGAN	Doç. Dr.	İstanbul Üniversitesi
Todor RADEV	Doç. Dr.	International University College
Abdülkadir ŞENKAL	Doç. Dr.	Kocaeli Üniversitesi
Ali TİLBE	Doç. Dr.	Namık Kemal Üniversitesi
Aykut Hamit TURAN	Doç. Dr.	Namık Kemal Üniversitesi
Banu UÇKAN	Doç. Dr.	Anadolu Üniversitesi
İrfan ATALAY	Yrd. Doç. Dr.	Namık Kemal Üniversitesi
Leyla ATEŞ	Yrd. Doç. Dr.	Namık Kemal Üniversitesi
Sonel BOSNALI	Yrd. Doç. Dr.	Namık Kemal Üniversitesi
Tatiana GOLBAN	Yrd. Doç. Dr.	Namık Kemal Üniversitesi
İmran GÜR	Yrd. Doç. Dr.	Namık Kemal Üniversitesi
Ali GÜREL	Yrd. Doç. Dr.	Namık Kemal Üniversitesi
Ahmet MENTEŞ	Yrd. Doç. Dr.	Namık Kemal Üniversitesi
Lütfü ŞİMŞEK	Yrd. Doç. Dr.	Namık Kemal Üniversitesi
Tevfik SÜTÇÜ	Yrd. Doç. Dr.	Namık Kemal Üniversitesi
Çiğdem VATANSEVER	Yrd. Doç. Dr.	Namık Kemal Üniversitesi
Ahmet Zeki BULUNÇ	Dr.	Başkent Üniversitesi (Emekli
Oscar A. POMBO	Dr.	Colef Üniversitesi

Hakem kurulunda yer alan isimler unvan ve soyadına göre alfabetik sıralanmıştır. Yayınlanmak üzere gönderilen çalışmaların konularına göre hakem ilavesi yapılabilir.

Comparison of the Economical Indicators of Turkey and European Union States via Decision Tree Method

ABSTRACT

The EU membership and accession process are essential in economical and social aspects for Turkey and many other non-member states. In this study the criteria for determining the candidate states and how these criteria affect the accession process have been a question for debate recently. The purpose of the study is to investigate whether the level of economic development criteria had an impact on the EU accession process and if they have an impact, to determine which economic criteria are the most important. The model, developed as a result of this study, allows the states considering applying for full membership to estimate their acceptance time.

In line with the purpose of the study, Inflation Rates, Export, Import, Exchange Rates, Unemployment Rates, Total Labor, Fixed Capital Investments, Gross Domestic Product and Population Density variables of Turkey and 20 EU member states have been analyzed. Macroeconomic data is calculated based on the change in the values between the year of application for full membership and the year they are awarded full membership. Since the founder states were not subject to accession process they are not under scope of the study.

In application, the C4.5 algorithm data was manually derived, and certain rules have been reached. The data used in the manual solution of the C4.5 algorithm were then applied to the J48 and J48-Part algorithms in WEKA (Waikato Environment for Knowledge Analysis) computer program and the obtained results have been discussed.

Key Words: European Union, Data Mining, Decision Trees, C4.5 Algorithm, J48 Algorithm.

Avrupa Birliği Ülkeleri ile Türkiye'nin Ekonomik Göstergelerinin Karar Ağacı Yöntemi ile Karşılaştırılması

ÖZET

Avrupa Birliği üyeliği ve aday ülkelerin değerlendirildiği üyelik süreci Türkiye başta olmak üzere pek çok diğer ülke için ekonomik ve sosyal olarak büyük öneme sahiptir. Bu bağlamda Avrupa Birliği'nin üye ülkeleri hangi kıstaslara göre belirlediği ve ekonomik gelişmişliğin üyelik surecine nasıl bir etkisi olduğu son yılların en çok tartışılan konularından birisi olmuştur. Bu çalışmanın amacı; ekonomik gelişmişlik kıstasının müzakere sürecine etki edip etmediği ve eğer etkili ise bu süreçte hangi ekonomik kıstasların daha belirleyici olduğunu ortaya çıkarmaktır. Ayrıca çalışma sonucunda oluşturulan model, AB'ye tam üyelik başvurusunda bulunacak ülkelerin, müzakere sürecinin kaç yıl süreceğini tahmin etmelerine olanak sağlayacaktır.

Makalenin amacı doğrultusunda AB üyesi 20 ülke ve Türkiye'nin makro ekonomik verilerinden Enflasyon Oranları, Kur Oranları, İşsizlik Oranları, İhracat, İthalat, Toplam İş Gücü, Sabit Sermaye Yatırımları, Gayri Safi Yurtiçi Hâsıla (GSYIH) ve Nüfus Yoğunluğu değişkenleri incelenmiştir. Makro ekonomik veriler, ülkelerin tam üyelik başvurusu yaptıkları yıl ile üye kabul edildikleri yılda görülen değerlerin değişim miktarları alınarak hesaplanmıştır. Kurucu ülkelerin adaylık süreci olmadığından çalışma kapsamı dışında tutulmuştur.

Uygulamada C4.5 algoritmasının elde çözümü yapılmış ve karar kurallarına ulaşılmıştır. C4.5 algoritmasının elde çözümünde kullanılan veriler WEKA bilgisayar programında J48 ve J48-Part algoritmasında da uygulanmış ve elde edilen sonuçlar tartışılmıştır.

Anahtar Kelimeler: Avrupa Birliği, Veri Madenciliği, Karar Ağaçları, C4.5 Algoritması, J48 Algoritması.

İçindekiler

1. INTRODUCTION

Turkey's EU accession process has been the long discussed in Turkey, in EU institutions and among citizens in EU countries. Relationships between Turkey and EU have a long history. Starting with the Ankara Agreement of 1959, the process continued with Turkey's full membership application of 1987. In that period, EU opened its doors to a number of countries in Europe but Turkey was not among those countries. The EU expansion process consisting of 5 phases is as follows: 1st expansion: UK, Ireland and Denmark; 2nd expansion: Greece, Spain and Portugal; 3rd expansion: Sweden, Finland and Austria; 4th expansion: Greek part of Cyprus Poland, Hungary, Czech Republic, Slovenia, Estonia, Latvia, Lithuania, Slovakia and Malta; and 5th expansion: Bulgaria and Romania. Undergoing the longest evaluation process after its full membership application compared to other EU member states, Turkey's full membership negotiations have still not been finalized.

According to research conducted, contrary to what is being covered in the press, the 'religion' factor does not top the list of arguments against Turkey's membership in Europe and its weight is estimated at only 25%. The weight of Turkey's geographic location is 26%, historical and cultural factor – 30% and economic concerns – 40%, topping the list (Kabatepe, 2001).

The purpose of selecting economic study is to determine the efficiency of economic data in the accession process and to make a contribution to scientific research by reaching objective results through numeric data. The economic variables to be selected have been determined in consideration of previous studies and variables included in the Maastricht criteria. Certain value limitations have been set for macroeconomic indicators in scope of the Maastricht criteria. E.g. it is recommended to keep the inflation rate equal to or under 2.93 and keeping interest rates equal to or under 6.84 in EU candidate countries has been pointed out as an important development in the EU accession process.

In the search for an answer to the question "What should Turkey do?", the question "What requirements does the EU expect to be satisfied for accession and what requirements affect the accession process and to what extent?" should be asked. The time that passes from candidacy to becoming a "member state" is related to the speed with which the candidate country satisfies the said requirements. This study analyses whether or not the length of the negotiation process[1] of candidate countries is correlated with the change observed in their economic variables.

[1] In compiling the date, in order to get more realistic results, the 'date of application' has been taken into consideration instead of the 'negotiations starting date'. Since the 'negotiations process' concept is used in the literature without considering this distinction, the process between EU candidate countries' application year and the year they are acknowledged as candidates shall be expressed as the "negotiations process" in the scope of the study.

2. METHODOLOGY

2.1. Data Mining

Data Mining (DM) analyses observational data sets, which are often large, for the purpose of detecting unsuspected relationships and summarizing the data in different ways, which are understood by and beneficial for the data owner. Data mining exercises produce relationships and summaries, which are often called models or patterns (Hand, Mannila and Smyth, 2001). Linear equations, rules, clusters, graphs, tree structures and recurrent patterns in the time series can be cited as examples. The data sets studied in data mining are often large, as also mentioned by the definitional. If we only dealt with small data sets, we would then be simply discuss exploratory data analysis, as statisticians do. New problems occur when encountering large bodies of data (Hand, Mannila and Smyth, 2001).

DM helps pull out hidden information from database systems consisting of large data stack. This operation is conducted using the disciplines of statistics and mathematics, modeling techniques, database technology and a variety of computer software. The aim in DM modeling process is to do data research and produce a general model from the results of the research. Each model obtained should be statistically significant and valid. This problem can be solved through pre-processing step of the Knowledge Discovery in Databases (KDD) process, in which case a part of the data is removed (Dunham, 2003). The main difference between DM and statistics is the use of DM not by statisticians but by commercial users. Naturally, DM (especially in terms of a database) includes not only modeling, but also the development of effective and efficient algorithms (and data structures) to conduct modeling on large datasets (Dunham, 2003). The weak part of DM compared to statistics is the fact that DM analyses are deprived of clearly formulated analyses because the data itself is not a pre-defined hypothesis and offers guidance (Dunham, 2003).

Data in this study shall be analyzed and evaluated using Decision Trees (DT), a DM classification method.

2.2. Classification Approach

Classification is a widespread problem involving a number of different applications aiming to place objects into one of a few predetermined categories (Tan et al., 2006). In the literature, classification is under the same title as regression. The main difference between classification and regression models, which use current data to predict the future and have the broadest use among DM techniques is the predicted dependant variable's having a categorical or continuous value (Eker, 2006). While classification predicts categorical values, regression is used in the prediction of continuous values.

Classification techniques are suitable for the prediction and definition of data groups with at most double or nominal categories. The techniques are not as suitable for ordinal categories since the techniques do not consider the hidden order among categories. Moreover, other forms of relations between categories such as lower and upper class are ignored (Tan et al., 2006). The most important function of classification is its revelation of characteristics of persons, objects and institutions in each category after the classification.

Main techniques used in classification and regression models are: Decision Trees, Instance Based Methods- k nearest neighbor, Bayes Classifier, Artificial Neural Networks and Genetic Algorithms.

2.3. Decision Trees

The decision trees are a group of rules detecting statistically significant groups and providing answers in an explicit manner with easily readable tree diagrams, classifying or predicting observations (Doğan et al., 2003). A DT means an orderly division of a data set for the purpose of maximizing differences on the dependant variable (Dormen, 2003). In an algorithm used on a DT whose data is useful for the classification of data in accordance with certain variable values, inputs and outputs are determined variables of the data and the DT algorithm discovers input data variables for output data variables through data structures (Tan et al., 2006). DT is one of the effective methods used to generate classifiers from the data. DT presentation is the most broadly used logical method. Basically, there are a lot of DT induction algorithms defined in machine learning and literature on applied statistics. These algorithms are tested learning methods creating DTs from a serial input-output set. Typically, a DT learning system adopts the top-to-bottom method looking for a solution on a part of the research area. This method guarantees that a simple tree (not necessarily the simplest) can be found. A DT contains nodes in places where variables are tested. Branches proceeding outside from a node match all possible results of the test in that node (Kandartzic, 2003).

The interpretability of decision trees is one of their most attractive sides, especially concerning decision rule construction. Decision rules can simply be established by following any path from a root node to any leaf. A complete set of rules obtained from a decision tree is equivalent to the decision tree itself (for the purpose of limitation). The decision rules appear in the following form: IF (antecedent) THEN (consequent). Concerning the decision rules, the antecedent consists of characteristic values obtained from branches through which a certain path passes on a decision tree while the consequent is comprised of limitation values belonging to a target variable provided by a certain leaf node (Larose, 2005). The IF part consists of all tests on the path and the THEN part is the final classification. Rules of this kind are called decision rules and all the decision rules for all leaf nodes classify the examples just like a tree does. The order of tree rules is unimportant.

DT induction is a non-parametric approach to the creation of classification models. DT algorithms are considerably resistant to noise during the implementation of methods to avoid over-equipment. The existence of more variables than necessary does not negatively impact the accuracy of a DT (Tan at al., 2006). Here are some advantages of decision trees: They are easily understandable, being widely used to explain how decisions are made relying on multiple criteria. Either categorical or continuous data can be used to construct a decision tree. Finally, using a decision tree, a data set can be partitioned into distinct regions on the basis of ranges or specific values (Myatt, 2007).

In addition to the advantages, DTs also have certain disadvantages. Generally, decision making is a serious problem for the DT approach. This is caused by the fact that as the tree gets wider with more sections originating from it, less information will remain in section nodes as a result of the classification made. The DT divides data into a lot of parts. As those parts get more specific, they start getting smaller. As the number of different cases requiring observation increases, each of the training sets gets even smaller. Due to the decrease in the figures, less reliability remains in the accurate depiction of classification. Even through a DT consists of a lot of small branches, finding rules that may pass an accurate statistical observation among such nodes is an optimistic probability since generally, each node leaving those branches shall contain a small proportion of all possible classes. This may lead to problems in application (Seidman, 2001).

Decision Tree Application Criteria

To apply methods based on the induction learning method, a few important conditions should be observed. These conditions have been given below (Kandartzic, 2003).

Variable-Value Definition points to the regular format of the data to be analyzed, i.e. all information, variables and rates relating to an object or an example should be expressible as a constant sum. Each characteristic may have both differential and numeric values. However, characteristics used to define examples should not vary depending on situation. Where necessary, continuous variables should be made discontinuous, which should be provided by an algorithm. Due to its nature, such limitation leaves definition sets in examples with a variable structure out of the scope.

Pre-Defined Classes point to the necessity to create categories, to which examples are to be assigned, in advance (examined data). For Discrete Classes, regardless of whether the case belongs to a special class, concrete definition of classes should be understandable. A lot more examples are expected than classes.

The criterion of **Sufficient Data** tells us that if a sufficient number of concrete patterns can be separated from coincidences, then the approach is valid. Since this approach is generally based on statistical tests, a sufficient number of examples should be available for these tests to be effective.

A tree consists of decision points, on which the complete set of observations or a subset of the observations is split depending on some criteria. Each point in the tree stands for a set of observations, which is referred to as a node. The relationship between two joined nodes is described as a parent-child relationship. The larger set to be divided into two or more sets of a smaller size is referred to as the parent node. The nodes produced by dividing the parent are called child nodes. A child node that is not divided any further (does not produce any more children) is called a leaf node (Myatt, 2007).

2.4. Decision Tree Algorithms

DT algorithms are used in prediction tasks where a classification model is required. The cases have been designed for partitioning into different groups for the problems to be solved in the best way.

As a principle, there is a number of DTs that could be generated of a series of given variables. While some DTs are more accurate in comparison to others, obtaining the most suitable tree is impossible to calculate due to the increasingly growing dimension of the research field. Nevertheless, effective algorithms have been developed for the induction of the reasonably accurate DT within a reasonable time. Such algorithms, in their majority, use a strategy that generates a DT through the making of most suitable decisions on which characteristic is to use to divide the data into partitions.

One of such algorithms is the Hunt's algorithm, the algorithms that serves as a basis for a number of current DT induction algorithms including ID3, C4.5 and CART (Tan et al., 2006). In the Hunt's algorithm, the DT separates training records into purer sub-groups and grows them in a recursive manner. Let D_t be a set of training records formed with the node t and y = $y_1, y_2, \ldots y_c$ a class label. A recursive definition of the Hunt's algorithm is as follows (Tan et al., 2006). Basically, the Hunt's algorithm consists of two steps. In the first step, if all records in D_t belong to the same y_t class, t is a leaf node labeled as y_t. In the second step, if D_t contains records belonging to more than one class, a variable test condition is selected to partition the records into smaller sub-sets. For the Hunt's algorithm to work, a combination needs to be available for each variable value in the training data and each combination needs to have the same class label. These assumptions are impossible to satisfy in a number of applications.

2.4.1 ID3 Algorithm

Quinlan ID3 and C4.5, the more developed version of the latter, are one of the best known tree development algorithms used to create a DT based on single-variable partitions (Kandartzic, 2003). At the end of the 1970s, J.Ross Quinlan developed Hunt's 'Divide and Conquer' algorithm to create the ID3 DT algorithm. While characteristics in the method used by Hunt were selected randomly, Quinlan

used the following entropy method for the selection of variables, thus resolving the most eliminating the most important weakness of Hunt's method (Berson et al., 1999).

The ID3 algorithm starts with correction examples in the root node of the tree. A variable is selected to divide those examples. A branch is created for each variable value and example sub-sets, which gain a new characteristic from the branch, are placed to the newly created sub-node. The algorithm is repeatedly applied on each sub-node until all examples on a single node belong to a single class. Each path leading to a leaf on a DT is associated with a classification rule. An important aspect in such a top-to-bottom DT deduction algorithm is the selection of the characteristic in the node. The variable selection in ID3 and C4.5 is based on minimizing the entropy criterion information applied for examples in a node (Kandartzic, 2003).

The variable selection part of ID3 is based on the following assumption: To a great extent, the complexity of a DT depends on the amount of information replacing the value of the given characteristics. Information-based tracking selects the characteristic that minimizes the information required in the resulting sub-tree in order to classify the characteristic that provides the highest information gain. In the C4.5 algorithm, the classification field extends from categorical characteristics to numeric characteristics. The criterion supports variables that are received after division into sub-sets with low entropies, provided that the majority of examples inside belong to a single class. Basically, the algorithm chooses characteristics with the highest degree of discreteness locally among classes (Kandartzic, 2003). The basic idea behind the induction algorithm is to ask questions whose answers provide most information.

2.4.2 C4-5 Algorithm

C4.5 is the developed version of ID3. C4.5 is capable of dealing with deficient and continuous variable values and of conducting operations such as DT pruning and rule induction (Dunham, 2003).

C5.0 has been developed with rule deriving speed and quality at a better level than C4.5, its previous version. Additionally, C5.0 has also implemented the technique called boosting, combining multiple DTs in a single classifier. Boosting is an approach of using different classifiers together. While normally, boosting requires more time to operate a certain classifier, it increases the accuracy rate. It has been observed that error rate on some datasets is less than a half of what is found for C4.5. Boosting is not always effective in cases where the training set includes a lot of noise. The operation principle of boosting is the creation of multiple training sets from one training set. Each item in the training set is assigned a weight. The weight represents the importance of that item for classification. A classifier is created for each combination of weights used. Thus, a lot of classifiers are actually crated. When classifying using C5.0, each classifier is assigned a vote, voting is conducted and the target variables group is allocated to the class getting most votes (Dunham, 2003).

The C4.5 algorithm is superior compared to ID3 in many aspects. The algorithm offers solutions in cases of missing observation and/or presence of continuous variables in the data while pruning and discretion strategies and rules ensure the obtainment of a more distinct DT.

3. APPLICATION

This study uses the C4.5 DT method and Classification Technique from among DM techniques to analyze and evaluate the criteria, according to which EU candidate countries are assessed, whether or not economic criteria affect the EU accession process and, if they do, what macroeconomic variables are effective and finally, what developments an EU candidate or prospective candidate needs to demonstrate after the EU application in order to accelerate its accession process.

In line with the above-described purpose, Inflation Rates, Export, Import, Exchange Rates, Unemployment Rates, Total Labor, Fixed Capital Investments, Gross Domestic Product (GDP) and Population Density variables from among macroeconomic data parameters of 20 EU member states[2] and Turkey have been analyzed. The reason why Population Density, Unemployment and Labor data were included in the dataset is their being included in macroeconomic data in documentation published by the European Union and their great importance for a study analyzing the relationships between the EU and Turkey.

In the selection of these parameters, Progress Reports (2002-2007) annually published for Turkey since 1998 and evaluating Turkey's compliance with acquis communautaire and the macroeconomic indicator tables of Eurostat have been taken reference.[3]

Dates when relations with the EU started shall be considered the full membership application dates for EU member states and Turkey. This will ensure the obtainment of more realistic results in the comparison of economic data for all countries.

3.1. Findings Obtained as a Result of Manual Solution

The DT diagram obtained as a result of the solution of the dataset produced manually (using formulas) by using C4.5, a DT algorithm and decision rules obtained from the diagram have been provided below. (First class: ≤ 9 years, second class: > 9 years)

[2] Since France, Germany, Italy, Belgium, Netherlands and Luxemburg are founding states and their joining did not occur in the scope of an expansion, they have been left out of the study scope.

[3] See http://ec.europa.eu/economy_finance/indicators/annual_macro_economic_database/ameco_applet.htm

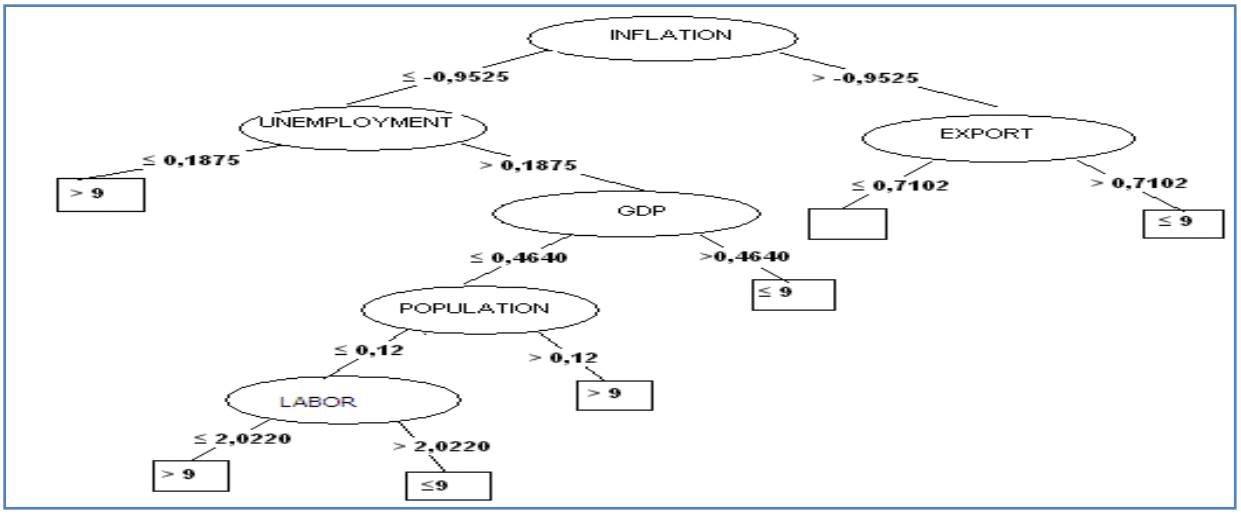

Figure 1. Appearance of the decision tree obtained through manual solution

The rules generated from the nodes of the DT created through manual solution are listed below.

RULES 1. If INFLATION R. ≤ -0,9525 and UNEMPLOYMENT R. ≤ 0,1875
Then Classification: Class 2 (greater than 9)

RULES 2. If INFLATION R. ≤ -0,9525 and UNEMPLOYMENT R. > 0,1875 and GDP >0,4640
Then Classification: Class 1 (less than or equal to 9)

RULES 3. If INFLATION R. ≤ -0,9525 and UNEMPLOYMENT R. > 0,1875 and GDP ≤ 0,4640 and POPULATION > 0,12
Then Classification: Class 2 (greater than 9)

RULES 4. If INFLATION R. ≤ -0,9525 and UNEMPLOYMENT R.>0,1875 and GDP ≤ 0,4640 and POPULATION ≤ 0,12 and LABOR ≤ 2,0220
Then Classification: Class 2 (greater than 9)

RULES 5. If INFLATION R. ≤ -0,9525 and UNEMPLOYMENT R.> > 0,1875 and GDP≤ 0,4640 and POPULATION ≤ 0,12 and LABOR > 2,0220
Then Classification: Class 1 (less than or equal to 9)

RULES 6. If INFLATION RATE >-0,9525 and EXPORT > 0,7102
Then Classification: Class 1 (less than or equal to 9)

3.2. Findings obtained using the WEKA software

In the solution of the DT method using computer software, the J48 method, the WEKA version of the C4.5 algorithm, which is, in turn, a DT algorithm developed by Quinlan has been used in the WEKA program, which is frequently used in machine learning studies.

WEKA is DM solution software developed at the Waikato University, New Zealand. The software is user friendly with 'the WEKA Explorer'. This program enables the users to actively create a DT. The program offers two variables, which you can select along with a data field. Once a pair of attributes that well discriminate the classes, a two-way split can be created by drawing a polygon around the appropriate data points on the scatter plot (Witten and Frank, 2005)

Given below are the results obtained from the J48 algorithm in WEKA.

```
J48 pruned tree
------------------
GDP <= 0.2605: <= 9 (4.0)
GDP > 0.2605
|   INFLATION <= -1.491: > 9 (8.0)
|   INFLATION > -1.491
|   |   EXPORT <= 0.7102
|   |   |   CAPITAL INVEST. <= 0.117: > 9 (4.0)
|   |   |   CAPITAL INVEST. > 0.117: <= 9 (2.0)
|   |   EXPORT > 0.7102: <= 9 (3.0)
Number of Leaves  :     5
Size of the tree :      9
Time taken to build model: 0.03 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances          3               60      %
Incorrectly Classified Instances        2               40      %
Kappa statistic                          0
Mean absolute error                      0.42
Root mean squared error                  0.5745
Relative absolute error                 85.9091 %
Root relative squared error            116.7808 %
Total Number of Instances               5
```

Figure 2. Results of the J48 algorithm

To test the model accuracy, other methods are created such as cross-validation and percentage split alternatives. The cross-validation splits the dataset into layers of equivalent size (as a default, the program uses 10 layers) and uses n-1 piece in the training of the model and 1 piece in model testing at each iteration. In Figure 2, to determine the classification accuracy of the model, 80% of the data is used for the training set and 20% of it for testing.

Considering the DT, the first split has been determined as the GDP variable with subsequent second-level splits being the Inflation rate and Exports respectively. In the tree structure, a colon introduced the class label that has been assigned to a particular leaf, followed by the number of instances that get to that leaf, expressed as a decimal number because of the manner the algorithm employs fractional instances to handle missing values (Witten and Frank,2005). In other words, the first number in the parenthesis indicates how many cases in the dataset have been correctly classified for the node and the second numbers shows the number of cases that have been incorrectly classified by the node. For instance, the value (4.0) in the expression (GDP <= 0.2605: <= 9 (4. 0)) indicates that there are no cases that have been incorrectly classified and sent to the related leaf. However, if this value were, e.g. (4.2), that would mean that there are 4 cases, of which 2 have been classified incorrectly and sent to that leaf.

We see that the number of leaves is 5 and the size of the tree is 9. The confusion matrix indicates that the two cases of the 'lower than or equal to' class have been assigned to the 'large' class and there are no cases assigned from the 'large' class to the 'lower than or equal to' class. Along with the classification error, the evaluation module also produces Kappa statistics.[4] The mean absolute error has been found as 0.42, the root mean squared error of class probability estimates assigned by the tree as 0.5745 and the relative error depending on previous probabilities as 85.90%. The root mean squared error is the square root of the mean quadratic loss. The mean absolute error, similarly, is calculated using an absolute value instead of a squared difference (Witten and Frank, 2005).

Again, as can be seen on the table above, when 80% of the data is selected for training and 20% for testing, the number of correctly classified cases is observed to be 60% (the rate of performance of the derived model in the training set). If 66% of the data is selected for training and 20% for testing, the number of correctly classified cases becomes 62.50% and the mean absolute error falls down to 0.37. The accuracy values obtained from both classifications are lower than those obtained from previous DT applications.

The DT has been partly developed by Frank and Witten. The J48 Part algorithm is obtained through the pruning of the DT obtained from the J48 algorithm. Rules trained by the J48 Part algorithm, whose rules are created using the J48, are independent of one another (Witten and Frank, 2005).

[4] Kappa statistics measures the degree of compatibility among diagnostic methods in the Mc Nemar test, a chi-square test adaptation for the comparison of rates in dependent groups (the before-and-after comparisons).

The results obtained by applying the J48 Part technique on the data used in the J48 technique have been indicated on Figure 3 below.

```
PART decision list
------------------
GDP > 0.2605 AND
INFLATION <= -1.491: > 9 (8.0)
EXPORT > 0.0851: <= 9 (11.0/2.0)
: greater than (2.0)
Number of Rules  : 3
Time taken to build model: 0.06 seconds
=== Evaluation on test split ===
=== Summary ===
Correctly Classified Instances          3               60      %
Incorrectly Classified Instances        2               40      %
Kappa statistic                          0
Mean absolute error                      0.42
Root mean squared error                  0.5745
Relative absolute error                 85.9091 %
Root relative squared error            116.7808 %
Total Number of Instances                5
```

Figure 3. Results of the J48 Part Algorithm

Comparing the results of the J48 and J48-Part, it is observed that all values are equal except for the number of nodes. The most important point requiring attention here is the pruning of the capital investment variable, which is observed in the J48 output and has the lowest information gain and its absence in the output of the J48Part. Thus, the most important 3 variables among the results of the J48 and J48-Part algorithms have been determined as the same variables through the manual solution.

The algorithm automatically excludes meaningless variables doing the variable selection in the new training process on its own. The reasons for the exclusion of the said 5 variables from the analysis have been examined using statistical analyses. The first reason for the exclusion of the variables from the analysis that can be thought of is the low correlation between the dependent variable (class variable) and independent variables and the exclusion from the analysis of independent variables whose correlation coefficient is lower than a certain limit. The correlation coefficient shows the direction and degree of the relationship between variables. In order to find the appropriate technique for the correlation analysis, the variables have been first tested for being parametric. According to the results of the Kolmogrov-Smirnov test conducted using SPSS 16.0 software, it has been observed that

some variables have not been normally distributed and the Spearman Correlation Coefficient has been calculated using non-parametric methods. In order to test the correctness of the hypothesis, Spearman Correlation Coefficient was calculated using the SPSS Software Package and the results have been provided in Table 1.

Table 1. Spearman Correlation Test Results

Spearman Correlation Coefficient	Class Variable (annual difference)
Inflation rate	-,239
Exports	-,125
Imports	,071
Exchange rate	,097
Fixed capital investments	,050
Unemployment rate	-,228
Labor	,027
GDP	,241
Population density	-,062
Class variable (annual difference)	1,000

As seen in the Table 1, the most important 3 variables analyzed in the WEKA software are the ones with the highest coefficients in the Spearman correlation analysis. GDP of 0.241, Inflation rate of 0.239 (negative) and Exports of 0.125 (negative) are the variables providing the highest correlation coefficient. As seen from the DT obtained using the WEKA software in Figure 2, the coefficient sizes have been determined to be same as the correlation coefficient rating.

The DT based on the J48 results below has been visualized using the WEKA classifier tree visualizer. For the DT rules to be more understandable, they are going to be explained on the figure. The decision rules are as indicated in Figure 5.

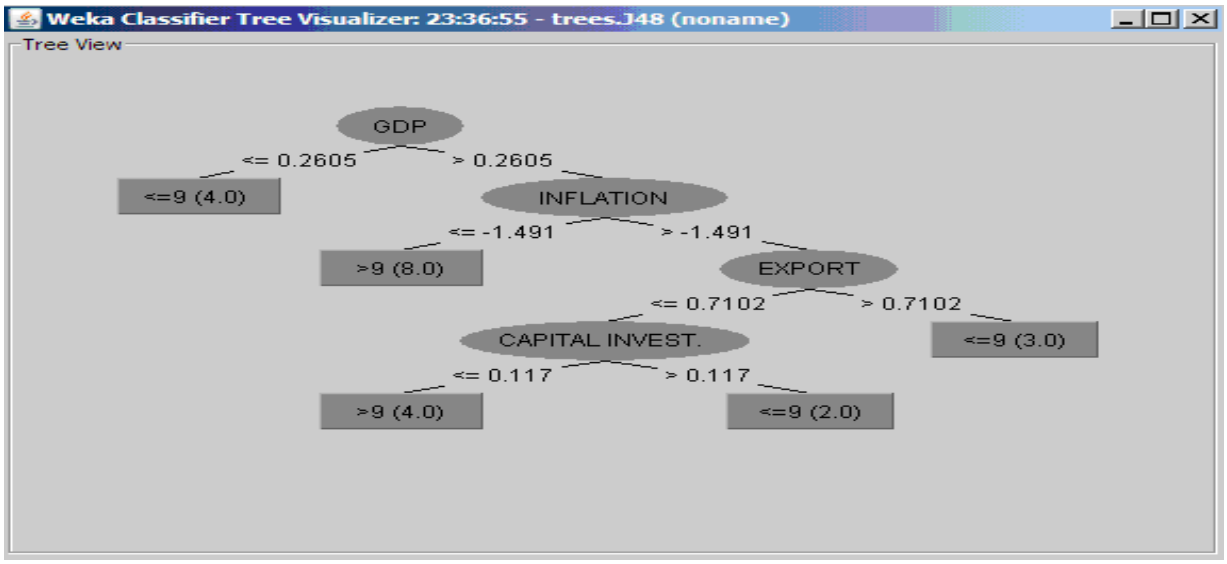

Figure 5. Visualization of the J48 results in the WEKA software

Rules obtained from the decision tree nodes,

RULES 1. If GDP ≤ 0.2605
Then Classification: Class 1 (less than or equal to 9)

RULES 2. If GDP >0.2605 and INFLATION RATE ≤ -1.491
Then Classification: Class 2 (greater than 9)

RULES 3. If GDP >0.2605 and INFLATION RATE > -1.491 and EXPORT > 0.7102
Then Classification: Class 1 (less than or equal to 9)

RULES 4. If GDP >0.2605 and INFLATION RATE > -1.491 and EXPORT ≤ 0.7102 and FIXED CAPITAL INVESTMENTS ≤ 0.117
Then Classification: Class 2 (greater than 9)

RULES 5. If GDP >0.2605 and INFLATION RATE > -1.491 and EXPORT ≤ 0.7102 and FIXED CAPITAL INVESTMENTS > 0.117
Then Classification: Class 2 (greater than 9)

Interesting results are obtained using the above rules.
According to Rule 1, if the GDP of an EU candidate country is less than or equal to 0.2605, the negotiations process will take less than 9 years. According to rule 2, if the GDP of an EU candidate country is more than 0.2605, no information can be provided on the duration of the negotiations process. Therefore, it is necessary to consider the Inflation rate variable, whose information gain is lower than that of GDP and higher than that of other variables. Here, it is indicated that the negotiation

process of candidates with a GDP greater than 0.2605 and Inflation rate smaller than -1.491 will take more than 9 years.

4. CONCLUSION

As a result of the findings obtained in the study, the manual solution has determined that the most important variable affecting the economic performance demonstrated by countries applying for the EU membership until their accession is Inflation rate. Considering the general appearance of the tree, variables with the highest gain rate are the Inflation rate, Exports, Unemployment rate and GDP. Since the Imports characteristic in the DT have demonstrated a lot of similarity to Exports values, they have been omitted. Likewise, since the Exchange rate characteristic does not include cases in the last node that would be affected by class values in the leaf, it has been excluded from the decision tree. The smaller than or equal to 0.7102 branch of the Exports characteristic could not be classified because the cases in Imports and Exchange rates were left out of the scope of observation and because they could not be labeled by a single leaf as a result of the calculations. All variables other than those reached the leaf, labeled by one of the class labels on the DT. Apart from that, all remaining variables have been labeled by one of the class labels on the DT and have reached a leaf.

In scope of the analysis conducted using C4.5, a DT algorithm, the J48 and J48-Part algorithms have been applied in the WEKA software and Capital Investments, the least important variable for the DT, has been pruned in the J48-Part algorithm and left out of the analysis scope.

As a result of the correlation analysis conducted to see the degrees of relationships between the variables, low correlations have been detected between the dependent variable and independent variables. The WEKA program has specifically not included into the decision tree the variables whose correlations between the dependent variable and independent variables for this case are under 0.10. Accordingly, the variables that can be included in the DT are required to have a certain correlation value.

Comparing the findings of the WEKA program and findings of the manual solution, it is concluded that despite the rating of the selection of most important variables is the same, the root node selection is different. The difference found in the selection of variable, which affects the dependent variable to the greatest extent can be explained by the digital operation of the independent variable data in the WEKA program and performance of binary branching by the program itself, despite the classes relating to the independent variables having been created in advance in the manual solution.

Interesting results have been encountered in the evaluation of the decision rules obtained from the DT. According to the Rule 1, if the GDP of an EU candidate country is less than or equal to 0.2605, the negotiation process will take less than 9 months. However, both progress reports and Maastricht

criteria recommend and require GDP values to be high. As understood from this result, EU has omitted this general criterion when granting membership to candidate countries even resulting in countries with lower GDP becoming Union members in a shorter time. Similarly, if the GDP of a candidate country is greater than 0.2605, this situation fails to provide sufficient information on the negotiations process. Therefore, the Inflation rate variable, whose information gain is lower than that of GDP and higher than that of other variables, needs to be considered. Here, it is shown that the negotiations process for countries with a DGP higher than 0.2605 and Inflation rate lower than -1.491 will exceed 9 years. Likewise, a conclusion is reached telling us in rule 2 that low inflation rates will also increase the length of the period between the application for EU membership and accession date, which also contradicts to the limits set for economic values in the Maastricht treaty.

Finally, the EU does not consider high level of economic development an indispensable prerequisite for membership as it describes in the progress reports via which it tracks the level of development of countries applying for EU membership, which is proven by the economic data of countries previously admitted to the Union.

This study aimed to provide answers to the questions 'Does the economic performance demonstrated by countries applying for EU membership observed from the date of application to the date of accession contribute, in any manner, to the length of the period from the application to accession (9 years in our example)?' and 'What extent of development in which economic variable will shorten that period?'. The conclusion we aimed to reach would be to 'Estimate the period of EU accession for a country intending to apply for EU membership in consideration of the rules we were going to obtain and taking into account that country's economic data and level of development.'. In accordance with the results obtained, it has been concluded that the EU does not consider economic and demographic performance of applicants in evaluating them and that strong economic performance demonstrated by a candidate country would not shorten its accession period.

Different studies can be conducted on the selection of variables affecting the EU negotiation progress in order to develop the study and obtain results that are more detailed in the light of this study. A DT can be created by selecting variables that affect the negotiations progress, which is a dependent variable, to the greatest extent by conducting a regression analysis using a number of variables to be covered by such a study. Furthermore, different results may be obtained from studies to be conducted by covering variables that include the countries' political and social conditions in addition to the economic data.

REFERENCES

Berson, A., Smith S. & Thearling, K. (1999). **Building Data Mining Applications for CRM.** McGraw-Hill Companies.

Doğan, N. & Özdamar, K. (2003). Chaid Analizi ve Aile Planlaması ile Bir Uygulama, **T.Klin Journal of Medical Science**. vol.23, No. 5, pp. 392-397.

Dormen, D. (2003). **Bankacılık Sektöründe Müşteri İlişki Yönetimi: CRM Acısından VM Yöntemi.** Doctorate Thesis.

Dunham, M.H. (2003). **Data Mining Introductory and Advanced Topics.** Southern Methodist University: Pearson Education Inc.

Eker, H. (2006). **Veri Madenciliği veya Bilgi Keşfi 1-2.** http://www.ikademi.com/insan-kaynaklari-bilgi-sistemleri/621

Han, J. & Kamber, M. (2006). **Data Mining: Concepts and Techniques.** San Francisco: Morgan Kaufman Publishers.

Hand, D., Mannila, H. & Smyth P. (2001). **Principles of Data Mining.** MIT Press

Kabatepe, E., 2005. **Müzakere Sürecinde AB ve Türkiye.** Ankara: TURKAB.

Kandartzic, M. (2003). **Data Mining Concepts Models and Algorithms.** John Wiley& Sons.

Myatt, J.G. (2007). **Making Sense of Data A Practical Guide to Exploratory Data Analysis and Data Mining.** A John Wiley& Sons, Inc., Publication.

Larose, T.D. (2005). **Discovering Knowledge in Data An Introduction to Data Mining.** A John Wiley& Sons, Inc., Publication

Quinlan, J.R. (1993). **C4.5 Programs For Machine Learning.** California: Morgan Kaufman Publisher

Seidman , C. (2001). **Data Mining with Microsoft SQL Server 2000.** Washington: Microsoft Press.

Tan, P.N., Steinbach M. & Kumar, V. (2006). **Introduction to Data Mining.** Boston: Pearson Addison Wesley.

Witten, I.H. & Frank, E. (2005). **Data Mining Practical Machine Learning Tools and Techniques.** Second Edition, Department of Computer Science University of Waikato: Morgan Kaufman Publisher

Namık Kemal Üniversitesi Sosyal Bilimler Metinleri

Namik Kemal University Papers on Social Science

Namık Kemal Üniversitesi Sosyal Bilimler Enstitüsü
Namık Kemal Üniversitesi Değirmenaltı Yerleşkesi
Rektörlük Binası 4. Kat Tekirdağ 59030
e-posta: sosyalbilimler@nku.edu.tr
http://sosyalbe.nku.edu.tr

Sosyal Bilimler Metinleri
ISSN 1308-4453 (Print)
ISSN 1308-4895 (Internet)