RESEARCH ARTICLE OPEN ACCESS

Regressiontree Analysis Withcartalgorithm: GII Predictionusing NRI Indicators

Merve Doğruel Anuşlu*, SeniyeÜmitFırat**
*(Department ofIndustrial Engineering, Istanbul Gedik University, Turkey
** (Department of Industrial Engineering, Marmara University, Turkey
Corresponding Author : Merve Doğruel Anuşlu

ABSTRACT
In these days, innovation levels and capacities of the countries are of great importance both in terms of competitiveness and the Industry 4.0 revolution that we have been experiencing. In this context, capacity and level are relative concepts and on a global basis, there is a great need for a common measuring system in terms of comparisons. The Networked Readiness Index (NRI) and The Global Innovation Index (GII) are important global indexes with effective and academic infrastructure in identifying countries' innovation levels. This study has been conducted by using regression tree technique, which is one of data mining techniques, with the GII score and the indicators in the pillar under the sub-indexes of the NRI. Classification and Regression Tree (CART) analysis was applied for GII prediction based on NRI indicators and determining the NRI indicators that provide the best resolution. In the application, the model with the lowest error squares averages has been proposed and thus it is expected that this model will be used to make predictions in future studies.
Keywords- ClassificationandRegressionTree (CART), Data Mining, DecisionTree Learning, Global Innovation Index (GII), Innovation, Networked Readiness Index (NRI), RegressionTree.

Date of Submission: 01-06-2018 Date of acceptance:16-06-2018

I. INTRODUCTION

Innovation, which is the basic tool of creating economic prosperity; contributes to the fight against climate change on a wider scale, triggers sustainable development and promotes social cohesion [1]. Innovation, which seeks solutions to global problems in such a wide range, has become extremely important in terms of sustainable global competition, especially in the last 50 years.

In today's competitive world,both developed and developing countries have to find common innovative solutions to global challenges and simultaneously meet the urgent needs of their own people. Innovation, an element that transcends national borders and strengthens people and politics,is a critical factor in the growth of countries.

The increasing tendency of global connectivity requires the ability to solve problems at individual, societal, regional and global levels as well as a standardized path. It is possible to measure and analyze innovation data through key indicators. Since 2007, Global Innovation Index (GII) has been ranking world economies according to their innovation capabilities and results by using 82 indicators.In the 2016 edition of GII, other important parameters including patent applications, training expenditures, exports of creative products and other international dimensions have been added into those indicators.

On the other hand, we are experiencing the Fourth Industrial Revolution, which represents the transition to a new set of systems, combining digital, biological and physical technologies with new and powerful options. These new systems are built on the infrastructure of the digital revolution. Global Information Technologies Report 2016 has a content that contains countries' state of readiness to benefit from the emerging technologies and evaluates the opportunities offered by the digital revolution and beyond. In this context, the Network Readiness Index (NRI) is very convenient for examining global innovation.

Information and Communication Technologies (ICT) development; innovation in information systems,are effective in the continuous training and improvement of managing competencies and professional skills [2]. The drivers of the ICT revolution can be measured globally by the NRI.

Innovation from the perspective of Human Capital (HC) and Information and Communication Technologies (ICT); new ICT information for businesses, governments or social communities means the capacity to develop new talents skill such as social and managerial competencies. From the perspective of Human Capital (HC) and

Information and Communication Technologies (ICT), innovation is new ICT information for businesses, governments or social communities which means the capacity to develop new talents skill such as social and managerial competencies. These changes are measured with the Global Innovation Index (GII)which has detailed criteria on the innovation performance of countries and economies around the world [2].

GII, considered as one of the most effective indicators of countries' innovation levels and NRI, which deals with innovation as ICT-based, are two important indexes in the field of innovation that have an academic background. These two indexes, both in scope and function,are similar to each otherand seem to be related.

Every day, new data at the level of exabytes are created and carried out through IP (Internet Protocol Network) networks. In 2016, the world has entered the "zettabyte era" and global IP traffic has reached a capacity of 1.1 zettabytes or more than 1 trillion gigabytes. It is estimated that until 2020 global IP traffic will reach to 2.3 zettabytes. This data growth is fueling the economiesand inducing innovation by creating waves of creativity. The Global Information Technologies Report of 2016 highlights the role of technology and broadband in particularto promote global innovation [3]. Without internet networking, it does not seem possible for any innovation to be realized. IP networks; has the capacity to connect each person, each country and every device with the IP feature. Global networks ensure the quick growth of data without interruption and their collaborative innovation in many areas from manufacturing to services and prosesses. Countries that are equipped to promote digital activity continue to contribute to the emergence of new sectors and the rapid development of traditional sectors.

The role of hardware, software,and services have even more critical importance for governments, businesses,and individuals and for this reason, high-speed broadband Internet Protocol (IP) networks have become a part of everyday life. As a matter of fact, it is estimated that there will be more than 26 billion internet connected devicesand more than 4 billion global internet users until 2020. Broadband internet is categorized as one of the most important general-purpose technologies in the world with its social structures and the ability to significantly influence all economies [3]. When these points are taken into consideration, the relation of the innovation with the network systems is more evident. Network systems provide real-time, fast, and very large amounts of data flows and collected - accumulated data increase the need for new methodologies other than traditional techniques at every step of the data processing, from storage to analysis, from summarization to modeling. Data mining has become one of the most popular areas in recent years due to the driving force of these needs.

Data mining is a procedure that is useful for researchers to discover hidden, unknown,interesting relationships in the huge data sets and are widely used both in scientific studies and in industrial applications as a successful approach in terms of prediction [4]. Especially, in order to discover useful and valuable information among the masses, which grow like an avalanche with the big data coming to the agenda and usage of them data mining techniques and algorithms are indispensable tools. The speed of digital convergence offers powerful analysis techniquesin terms of growth in storage volumes and varying raw data.

Data mining is used for functions that can be aggregated into two main categories as descriptive and predictive. To fulfill these functions; data mining has a wide variety of different tasks, such as clustering, association rule mining, and classification [5]. There are numerous algorithms in these three tasks.

In this study; decision trees from data mining classification algorithms have been used. In the literature, regression tree analysis has been applied, which is less studied than the classification tree; thus, it has been aimed to provide a contributionto the lack of implementation of the regression tree analysis. In the research, within the scope of innovation that has been studied globally in the last fifty years, a prediction study has been conducted for GII, which is one of the most important indexes in determining the innovation levels of countries. GII has been predicted by using NRI indicators based on ICT baselines whose theme is "Innovating in the Digital Economy".

While doing the GII prediction in the regression tree model created using the CART algorithm, it has also been aimed to identify NRI indicators that provide the best resolution. In order for the forecasting model to be the most appropriate model,model experiments have been conducted to give the mean of the smallest error squaresandthe optimal tree model predicting the GII with the lowest error squared averages have been obtained. With the optimal CART decision tree model obtained,a model, which can be interpreted visually, whose forecast errors are low and prediction interpretation is easy, has been created. Prediction with this model is also foreseenusing the data in the coming years.

II. GLOBAL VIEW OF INNOVATION

The importance of the innovation was first highlighted by Schumpeter in the early 20th century. In the Oslo Manuel, co-developed by Eurostat and the OECD in 2005, the commonly accepted definition of innovation concept, which has been defined differently, that can be used for all approaches is made as follows: "the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organisational method inbusiness practices, workplace organisation or external relations"[6].

The development and dissemination of new technologies,formal and informal networks and actors composed of institutional sources that regulate these interactions are the basic structural element of innovation systems. Firms, research institutions, government departments, NGOs and other intermediary institutions are actors that contribute to the development and diffusion of innovation [7].

Basically, in innovation, there is an intellectual property including categories of invention, patent, license, other intellectual property, industrial design [8].

Innovation is one of the most important issues of the global agenda. In terms of competition, more importantly, in terms of the digital revolution (Industry 4.0) that is being experienced, it is very important to determine the locations of the countries. In this context, the World Economic Forum and other international organizations are carrying out intensive works on this matter. In this study, two global indexes (NRI, GII), which have the most attention in terms of country comparisons and which have an academic background in terms of indicators, have been taken into consideration.

2.1. The Networked Readiness Index (NRI)

The Global Information Technology Report, which was published in 2016 by World Economic Forum in collaboration with INSEAD Candornell University, measures Information and Communication Technologies (ICT) drivers on a global scale using Network Readiness Index (NRI). The index for 2016 covers 139 countries.

The Global Information Technology Report 2016 was determined as "Innovating in the Digital Economy". The Global Information Technology Report 2016, prepared with the theme "Innovation in the Digital Economy" emphasizes that the digital revolution has changed the nature of innovation and that firms are constantly subjected to increasing pressure to innovate. 4 key messages have been drawn from the report:

1. The digital revolution is changing the nature of innovation.
2. Companies are constantly faced with increasing pressure to make innovation.
3. Businesses and governments cannot keep up with the needs of the rapidly growing digital population.
4. A new economy is emerging that requires urgent innovation in governance and regimes.

These basic results once more draw attention that ICT is the area in which innovation is shaped, triggered and interacted with.

In the 2016 edition of the Global Information Technology Report, which was published in 2001 and developed over time, there are 53 indicators (Appendix 1) for NRI [3].

In Global Information Technology Report 2016; there are four sub-indexes that make up the NRI structure. Measurements that are obtained with the indicators under the ten pillars are used for creating sub-indexes (Fig. 1).

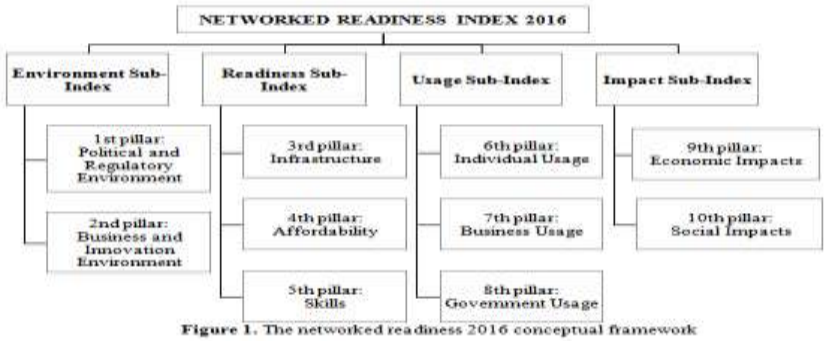

Figure 1. The networked readiness 2016 conceptual framework

2.2. The Global Innovation Index (GII)

First edition of The Global Innovation Index was published in 2007, by Cornell University in collaboration with INSEAD and the World Intellectual Property Organization (WIPO). GII is a common tool for evaluating countries in terms of innovation factors and offers a wide range of global comparative metrics. This research study based on findings and results of the 2016 edition of GII that including 128 countries.

Over the years GII has shown that the innovation capacity of any nation is measured not only by its level of locality but by how it affects the whole world at the same time. Poverty, health, urbanization, access to water and climate change are global issues. However, at the same time, both the challenges and the solutions have local consequences.

Therefore, innovative breakthroughs that offer local solutions in developing countries can have a global impactand they can provide opportunities for mutual benefit among other developing countries. Within this approach, the theme of GII Report 2016 was determined as "Winning with Global Innovation".

GII calculates four measures:

1. Innovation Input Sub-Index
2. Innovation Output Sub-Index
3. The overall GII score
4. The Innovation Efficiency Ratio

GII scores are calculated as simple averages of input and output sub-index values[9].In 2016, input sub-index consists of 5 pillars, output sub-index consists of 2 pillars, each pillar consists of 3 sub-pillars and under these sub-pillars, there are also a total of 82 indicators (Fig. 2).

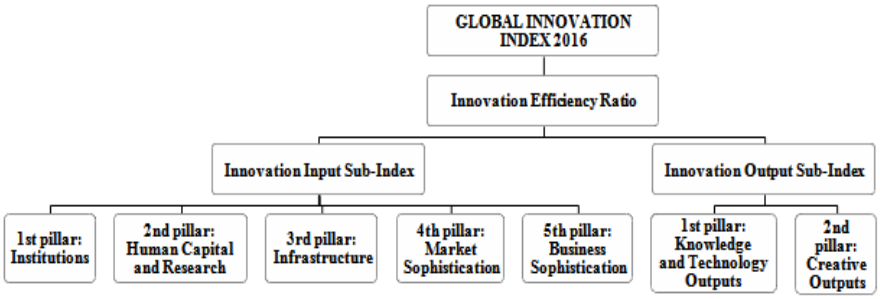

Figure 2. The global innovation index 2016 conceptual framework

When this framework is examined, 3rd pillar "infrastructure" in input sub-indexand the 1st pillar "knowledge and technology output" in the output sub-index seem to belong entirely to the ICT domain. In other words, GII contains variables similar to NRI indicators and common.

2.3. NRI-GII Relationship

Over the last fifty years, technology, innovation,and knowledge have been the three key concepts underlying the evolution of world economy and international business development [10].

ICT is one of the important drivers enabling innovation and growth for developed and developing countries. It has been shown that ICT is one of the most important sources and enables innovationfor economic growth in the developed markets [11].

In 2013, Kononova calculated a correlation coefficient of 0.94 between GII and NRI for 96 countries [12]. Obviously, the variables that can be considered common for both indexes influence these correlations' being significant and high. This finding can be easily explained in practice because the innovation theoretically requires network - internet structures.

Preda et al. established a univariate regression equation between GII and NRI for the 28 countries of the European Union in 2015and the correlation coefficient, R, was found to be 0.918 [13]. In Zoroja's research, it has been stated that ICT has a positive impact on innovation [14]. The results are supporting the widespread opinion "innovation does not happen without ICT".

III. DECISION TREES IN DATA MINING

In a data mining decision tree is a nonparametric prediction model that can be used to represent both classifiers and regression models and are used to indicate hierarchical models and outcomes of decisions in pieces of research[15].

Decision trees are hierarchical, in the form of directional trees composed of nodes and edges (Fig. 3). While a non-leaf node is called an internal or split node, a leaf node is called a terminal node [16]. When two nodes in the tree structure are connected by arrows, the node to which the arrow exists is called the parent node and the node that the arrow is entering is called the child node.

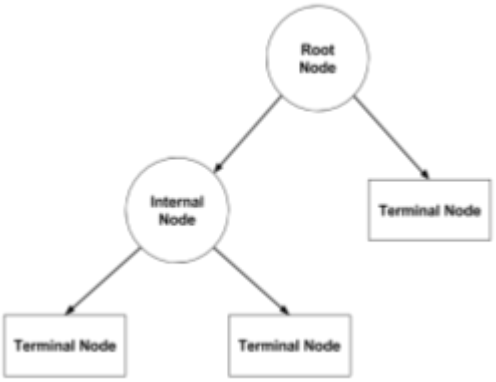

Figure 3*.A tree structure*

In all methods used while constructing decision trees, a set of "if-then" rules is generated which leads to a set of final values for the variable to be predicted. If the final values obtained are the probabilities of a categorical variable, then the created decision tree is described as a classification tree; if they are the quantities of a continuous variable, then the created decision tree is described as a regression tree [17]. A classification tree is an algorithm that represents a summary of decision rules. While the dependent variable is a categorical response variable, independent variables are predictors. Each internal node represents a decision based on an estimate. Each edge guides the potential future decision. Each leaf is labeled with a class. Aim; classification is done by following a path extending from the root to the leaves in accordance with the values of the estimators. The regression tree is also an algorithm represented by a summary tree, but the response variable is a real quantity instead of a class quantity. Decision nodes are similar to classification trees but for every leaf purpose variable, labeling is done with a quantity [18].

Decision trees in data mining, a collection of learning algorithms based on supervised machine learning bases. As in other learning algorithms also in learning decision trees, the chosen decision tree algorithm aims to create the most appropriate model from the learning data. Then the validity of the model created by the test data is testedand if the validity of the created model is confirmed, the corresponding model is used for predicting.

3.1. Regression Tree

The aim of the regression tree is to predict continuous dependent variable (response variable) using continuous and categorical independent variables.

The foundations of regression trees were laid in 1963 with the development of the Automatic Interaction Detection (AID) algorithm by Morgan and Sonquist. In 1984, the most popular version of the CART (Classification and Regression Trees) algorithm was developed by Breinman et al [19]. In the literature, there are also different algorithms such as GUIDE, M5, SUPPORT, SECRET, MART, SMOTI, MAUVE, BART, SERT [20].

Processes of creating and using regression tree models include three basic algorithmic sub-tasks [21]:

1. Regression tree growing
2. Regression tree pruning
3. Regression tree prediction

3.1.1. Regression Tree with CART (Classification and Regression Tree) Algorithm

CART (Classification and Regression Tree)It is a non-parametric and non-linear decision tree algorithm whose algorithm is used to generate both classification and regression trees and which predicts based on repeated duplicate allocations. If the response variable is categorical, the created tree is named a Classification Tree (CT);if the response variable is continuous, the created tree is called a regression tree (RT).

CART is an alternative to regression analysis where assumptions in the regression analysis are not met. CART is also used as an alternative to regression analysisbecause even when the data set has a complex structure, it determines independent variables that affect the dependent variableand presents the model's significance of these variables in an understandable visualization of the relations between them [22].

One of the most important advantages of CART is that it can work with a completely automatic and effective mechanism even when there is incomplete data [23].

In order to create a regression tree, the process of the CART algorithm can be summarized as follows [24; 25]:

1. CART performs all possible splits on each of the arguments starting from the root node and implements a predefined node impurity measure for each split.
2. It determines the reduction in impurity obtained.
3. CART then performs the best splitting by applying goodness-of-split criteriaand separates the data set into right-left child nodes.
4. Since CART is recursive, it repeats steps 1 to 3 for each non-terminal node, producing the largest possible tree.
5. Finally, CART applies pruning algorithm to the tree obtained.

3.1.1.1. Regression Tree Growing by CART

In order to enlarge the regression tree, one of the input variables at each step is selected to

separate the samples. The attribute value test is applied to the splitting point during the selected variableand the best splitting point is determined for the inner node to be split to the next nodes[26]. CART uses the least squared deviation (LSD) measure for splitting operations in the creation of regression trees or uses the least absolute deviation measure (LAD) [23].

The purpose while growing the treeis to split the input field to obtain fewer errors between the predicted output and the actual output. In general, the predicted outputs are determined as follows using the average of the actual outputs of training samples taken from a terminal node [26]:

$$\widehat{y_i} = \frac{\sum_{j \epsilon t_i} y_j}{|t_i|} (1)$$

t_i: the leaf node i

$|t_i|$: the number of samples in the leaf node i

The splitting criterion is based on the least squares deviation (LSD) impurity measure.

$$I(t_i) = \sum_{j \epsilon t_i} \left(y_j - \hat{y}_i\right)^2 (2)$$

$I(t_i)$: impurity measure at node i

Using LSD, the splitting criterionis calculated as follows [26]:

$$\Delta I = I(t_p) - P_l I(t_l) - P_r I(t_r) (3)$$

t_p:the parent node and t_l and t_r are the two child nodes of t_p

P_l and P_r are the proportions of data samples assigned to left and rigt child nodes r

The split point is determined to maximize ΔI

If the splitting rule is generated using a numeric or ordinal variable and the number of child nodes is two,the instances in the parent node are split into two subsets such as$\{x: x_k > s\}$and $\{x: x_k \leq s\}$. Herex$_k$defines the selected variable and s defines the splitting point. The same approach is used for nominal predictors; but for q categorized unordered categorical predictors $2^q - 1$ is found to be the possible splitting [26].

If no stopping rule has been applied, the process that follows a consecutive sequence continues until the homogeneity criteria are met and the maximum tree is reached, or until some stopping rules are applied [22].

3.1.1.2. Regression Tree Pruning by CART

If the tree structure is made too large while learning with training data, a tree model with zero defects that have each leaf in a single training is created. Especially, when working with small samples, the model can hardly generalize against situations that were not previously encounteredand therefore the predictions are not correct. This is known as overfitting the training data. To minimize this problem, pruning rules known as pre-pruning used in stopping the growth of the treeor pruning rules made after the tree grows, known as post-pruning, are applied [27].

For the pruning of the CART algorithm, a popular solution called cost complexity, which involves taking complexity into account with an explicit punishment for complexity has been identified [28]. The error-complexity measure, which is tried to be minimized, $R_\alpha(T)$consists of two parts as the total cost of classification error for T treeand punishment for complexity [29]:

$R_\alpha(T) = R(T) + \alpha|T| (4)$

R(T):total cost of classification error for T tree

|T|: number of terminal nodes

α: the penalty value applied to each terminal node

αvalue is equal to zero or greater than zero. If $\alpha = 0$, there is no punishment value, the cost complexity is the maximum level and it is a saturated tree. If α value is increased, R(T) the cost complexity will be reduced because the splits below the tree that reduce the value will be cut off [23].

In recent studies on pruning, αchanges place with cp[28].

$R_{cp}(T) \equiv R(T) + cp|T|R(T_1) (5)$

T_1:the tree with no split

|T|:the number of splits for a tree

R: risk of the tree

For a T tree, the overall risk in the K terminal node is as follows [28]:

$$R(T) = \sum_{j=1}^{K} P(A_j)\, R(A_j) (6)$$

This value is the sum of the risk associated with each node through terminal nodes.

The value of cp ranges from 0 to 1. When cp=0, one has a saturated tree. When cp=1, there are no splits.

A tree of the optimum size is selected among the different candidate trees by using independent test data or cross-validation [30]. If the data set is not large enough,use of the cross-validation method is recommended despite the computational complexity [31].

3.1.1.3. Regression Tree Prediction by CART

After selecting the optimal tree,CART calculates summary statistics for each terminal node. If the splitting rule is set as LSD,CART calculates the mean and standard deviation of the dependent variable. The mean of the terminal node is the predicted value of the dependent variablethese terminal node states. If LAD is selected, CART produces the median of the

dependent variable and the average of absolute mean deviations. For the terminal node, it is the predicted value of the median dependent variable [24].

IV. APPLICATION AND RESULT

For creating and using regression trees using the CART algorithm in the application phase, the flow in Fig. 4 was improved so that the application was carried out in the direction of this flow plan. The first aimof the practice is to predict the target variable GII using NRI indicators as predictors. On the other hand, when this prediction is made, it is aimed to determine the NRI indicators which provide the best splitting. It is aimed to create the most appropriate model to provide these two aims together with CART analysis.

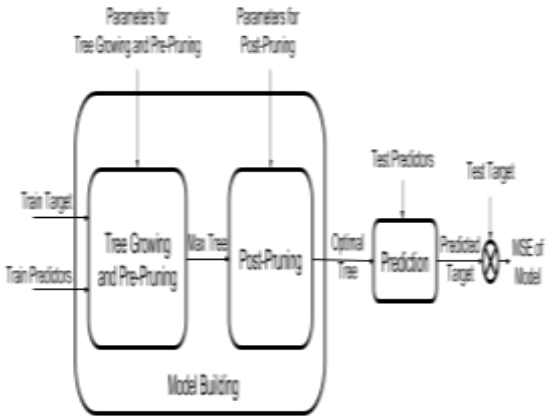

Figure 4*.The block diagram of Regression Tree with CART*

The R programming language was used to implement the analysis. The dataset will be generated from thepredictors of NRI indicators belonging to 2016 and the target variable the GII variable. Various operations have been performed in the data pre-processing such as: in R language the selection of relevant data for 2016 only; Implementation of the transpose process to ensure that the NRI data matches the GII; matching country names in order to be able to use data from the same country but written differently in the NRI and GII; the determination of the 26 countries that are in the NRI but not in the GII data and the 3 countries that are in the GII but not in the NRI data and excluding them from the scope; integration of NRI and GII data according to country names; and the exclusion of countries whose indicators consist entirely of empty data. Consequently, a data set, consisting of 54 variables as 53 predictors and 1 target variable (GII) belonging to the NRI dataset of 123 countries, was obtained.

For the model to be able to learn; 0.60 of data set has been used as training set and the remaining 0.40 has been separated as a test set.

In order to maximize the maximum regression tree with the CART algorithm, the "rpart" library is used in the R program. Once the relevant rpart library has been downloaded,the values of the arguments needed to grow the tree are defined for the rpart function. For this purpose, the cross-validation number "xval" has been determined as 10; the "minsplit" value, which is the minimum number of observation that should be found in the node, as 5; "minbucket" value, which is the minimum number of observations that should be present at any terminal node, as 5 and the value of the complexity parameter "cp" for pre-pruning has been determined as 0.001. While setting the cpvalue, care has been taken to ensure that the tree is not very complex, but that it is a value allowing the split to be determined at an optimum level.

In the tree enlarged with the specified arguments,we have obtained a structure consisting of 12 internal nodes and 13 terminal nodes that provide splitting for the largest tree used in the prediction of the response variable GII (Fig. 5).

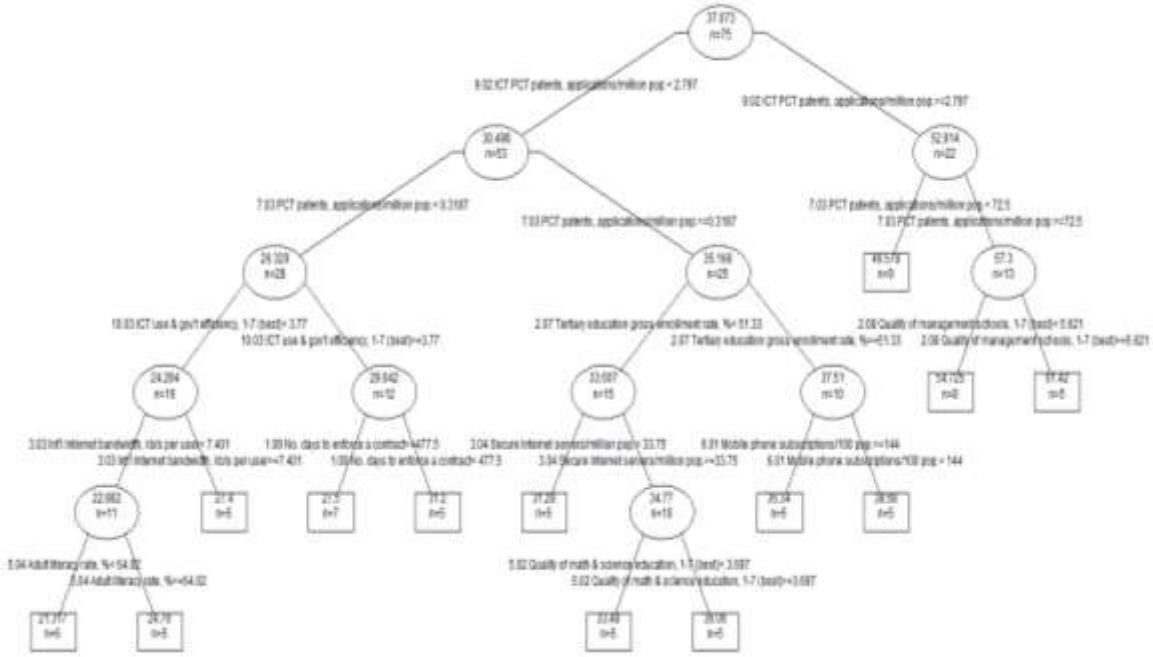

Figure 5. Maximum tree (train proportion: 0.60, xval: 10, minsplit: 5, minbucket: 5, cp: 0.001)

Table 1.Cross-validation errors

	CP	nsplit	rel.error	xerror	xstd
1	0.75311	0	1	1.01565	0.14136
2	0.09949	1	0.24689	0.30702	0.04912
3	0.05895	2	0.1474	0.2374	0.04296
4	0.0149	3	0.08845	0.21487	0.04949
5	0.0133	4	0.07355	0.20895	0.048
6	0.00881	5	0.06025	0.2117	0.04896
7	0.00677	6	0.05144	0.20485	0.05855
8	0.00391	7	0.04467	0.19666	0.05793
9	0.00385	8	0.04076	0.19687	0.05791
10	0.00312	9	0.03691	0.19478	0.05781
11	0.0016	10	0.03379	0.19623	0.05812
12	0.00132	11	0.03219	0.1953	0.05815
13	0.001	12	0.03087	0.19723	0.05826

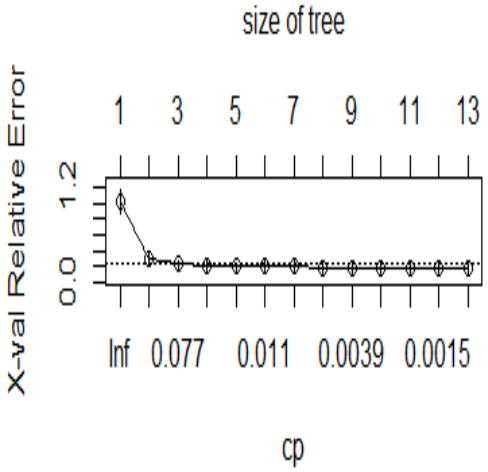

***Figure 6**.Cross-validation errors*

The post-pruning application will be done to find the optimal size of the tree from the largest tree created according to the values of the specified arguments. To carry out the post-pruning, the cp value with the least cross-validation error value of the maximum tree obtained should be selected.

As shown in Table 1 and Fig. 6, the least cross-validation error value was obtained with a value of 0.00312 cp. With this cp value, an optimal tree has been obtained and there are 9 internal nodes and 10 terminal nodes that provide optimal tree splitting (Fig. 7).

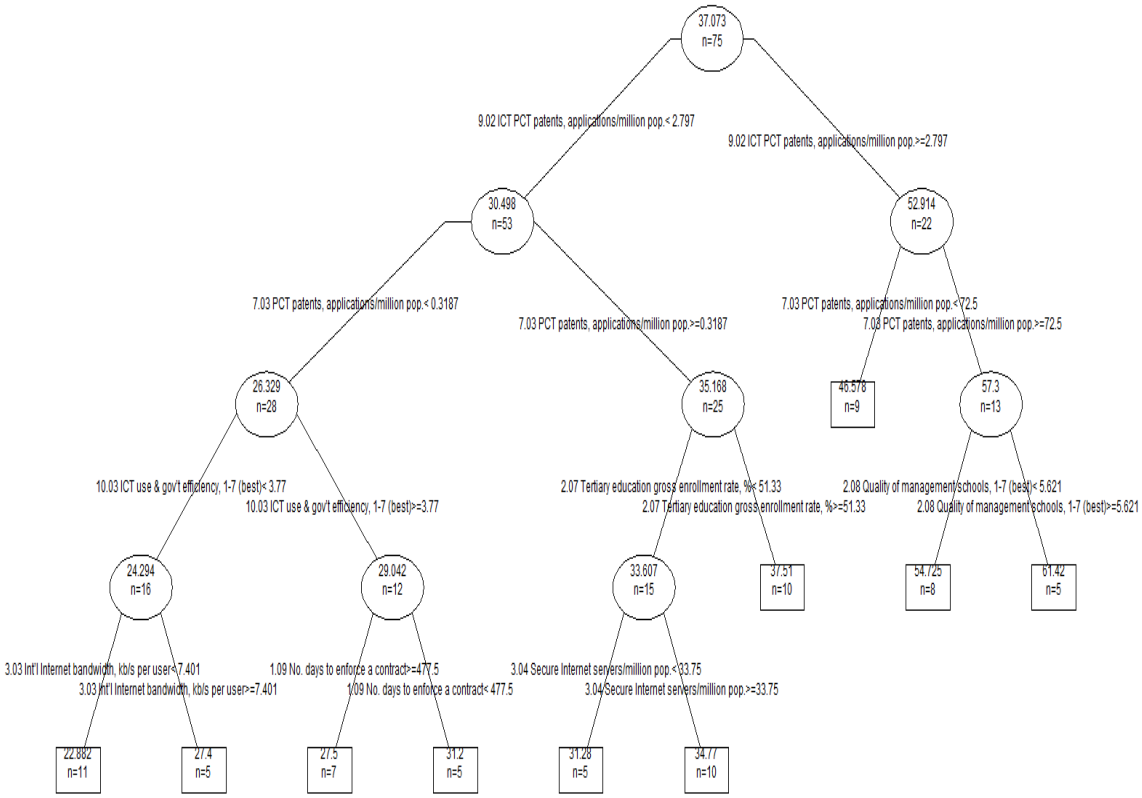

Figure 7. *Optimal tree (cp: 0.00312)*

GII values were predicted by using the obtained optimal tree and test data set. In order to compare predicted GII values with actual GII values, the minimum square error (MSE) was calculated and this value was found to be 15.927.

Table 2. Error squares average of different train scales and models set for XVAL values

Train/ XVAL	5	10	15	20	25
0.6	26.345	15.927	15.550*	19.111	15.883
0.7	23.496	23.496	25.604	22.787	22.780
0.8	24.985	24.985	25.215	26.884	26.426

In order to create the most appropriate model,model tests for different training set scales and xval values were made to investigate whether there is a lower average of error squares (keeping other arguments constant) (Table 2).

It has been decided to grow the maximum tree size with the percentage of the train with 0.60 which ves lowest error squares averageand 15 cross-validation counts. The optimal tree obtained by the cp value (0.0149) having the smallest cross-validation error value of this maximum tree is shown in Fig. 8 and this model will be used for predictions.

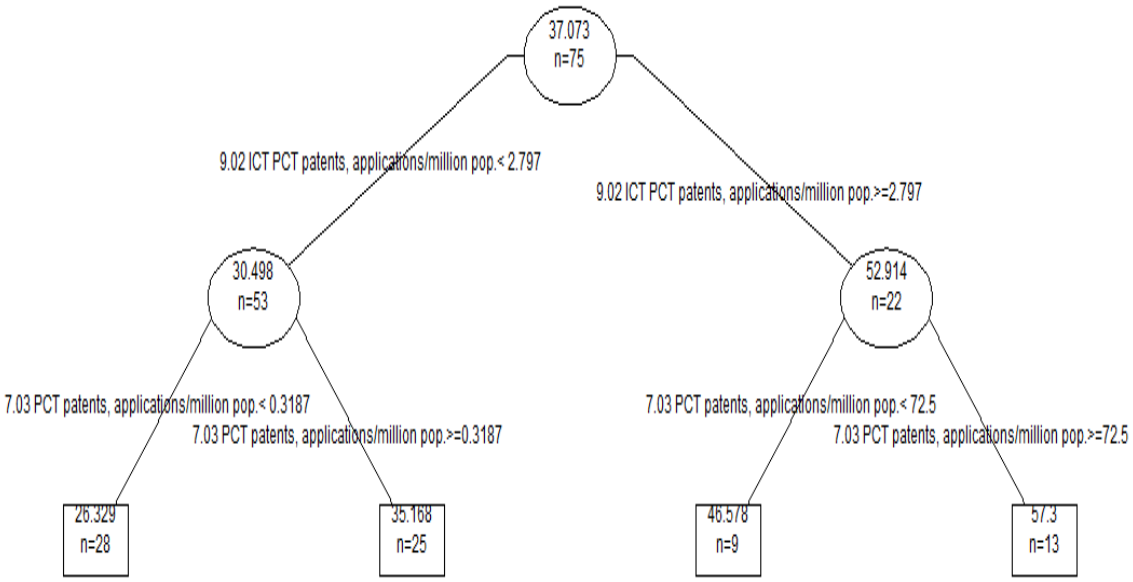

Figure 8. Optimal tree (cp: 0.0149)

V. CONCLUSION

In the practical section, which aims to predict GII scores through NRI indicators, variables that can be suggested as predictors have been determined on the basis of the ultimate optimum tree.

It is seen that the two most effective predictors for predicting GII are “ICT PCT patent applications per million population” and “PCT patent applications per million population”. The "ICT PCT patent applications per million population" predictor, which forms the root node, is one of the 9th pillar, or "economic impact", indicators under the “Impact Sub-Index” of the NRI. Another significant predictor, the “PCT patent applications per million population”, is one of the 7th pillar or "business usage" indicators in the "Usage Sub-Index" of the NRI.

When predicting GII, the variable that enables the split best is the “ICT PCT patent applications per million population”variables. According to the optimal tree model, four basic rules can be established:

1. If ICT PCT patent applications per million population < 2.797 and PCT patent applications per million population < 0.319, GII = 26.329.
2. If ICT PCT patent applications per million population <2.797 and PCT patent applications per million population ≥ 0.319, GII = 35.168.
3. If ICT PCT patent applications per million population ≥ 2.797 and PCT patent applications per million population < 72.469, GII = 45.578.
4. If ICT PCT patent applications per million population ≥ 2.797 and PCT patent applications per million population ≥ 72.469, GII = 57.300.

The predicted GII scores for 123 countries created by using the optimal tree model generatedand real GII scores (for 2016) are included in Appendix2. Although differences in eye examination may seem small, it was deemed appropriate to conduct a correlation analysis to confirm a statistic with a metric. The Pearson Correlation Coefficientbetween the predicted GII and the actual scores, has been found 0.95(alpha = 0.01) and this is extremely high and significant. The GII scores to be predicted by using the model in which these two NRI indicators are active have been very close to real values. Findings in this study strengthenauthors’ recommendation that "these two indicators are very appropriate to use in predicting GII scores".

On the other hand, the fact that the algorithm used, and the optimal tree metrics obtained are extremely good and valid enhances the reliability of the conceptual meaning reached in practice.

Consequently, the applied algorithm and the metric values of the optimized tree developed and the developed tree model, in which the determined and suggested NRI indicators "ICT PCT patent applications per million population" and "PCT patent applications per million population" are used dominantly in the research topredict the GII scores in practice, have given

results with conceptual meaning. This indicates that the study will contribute to the field.

In the future studies GII scores can be predicted by using the NRI indicators and data set by applying this proposed model that if NRI indicators remain the same in the next years. The possibility that, some indicators of NRI can be change in the future is the most important limitation of the prediction model developed in this study. But it is not a big difficulty, because a new version of the model can be rapidly developed by small modifications as changing predictor variables or adding new indicators.

REFERENCES

[1]. F.Gault,Definingandmeasuringinnovation in allsectors of theeconomy,ResearchPolicy, 47, 2018,617-622.

[2]. J.Kowal, G.Paliwoda-Pękosz, ICT forglobalcompetitivenessandeconomicgrowth in emergingeconomies: Economic, cultural, andsocialinnovationsforhumancapital in transitioneconomies,Information Systems Management, 34(10),2017, 304-307.

[3]. World Economic Forum, INSEAD andCornell University, The Global Information Technology Report 2016: Innovating in theDigitalEconomy, Geneva, FontainebleauandIthaca, 2016.

[4]. S.K.PurohitandA.K.Sharma,Developmentofdatam iningdrivensoftwaretooltoforecastthecustomerreq uirementforqualityfunctiondeployment, International Journal of Business Analytics, (4)(1),2017, 56-86.

[5]. R.Agarwal,M.Mittal, and S.Pareek,Lossprofitestimationusingtemporalassoc iationrulemining,International Journal of Business Analytics, 3(1), 2016, 45-57.

[6]. OrganisationForEconomicCooperationandDevelo pment(OECD).OsloManual:Guidelinesforcollecti ngandinterpretinginnovationdata (3rd ed.)(Paris, France: OECD Publishing, 2005).

[7]. C. Binz,and B.Truffera,Global innovationsystems - A conceptualframeworkforinnovationdynamics in transnationalcontexts,ResearchPolicy, 46, 2017,1284-1298.

[8]. A.Mataradzija,A.Rovcanin,andA.Mataradzija,Inn ovationandinnovativeperformanceintheEuropean Union,Proc.oftheManagement, Knowledge and Learning International Conference,Bangkok, Thailand, 2013, 77-82.

[9]. Cornell University, INSEAD & WIPO. The Global Innovation Index 2016: Winningwith Global Innovation. Ithaca, FontainebleauandGeneva, 2016.

[10]. U.Andersson,À. Dasì,R. Mudambi, andT. Pedersen, Technology, innovationandknowledge: Theimportance of ideasandinternationalconnectivity,Journal of World Business, 51, 2016, 153-162.

[11]. S.Amiri,andJ.M.Woodside,Emergingmarkets: Theimpact of ICT on theeconomyandsociety,DigitalPolicy, RegulationandGovernance, 19(5), 2017, 383-396.

[12]. K. Kononova, Someaspects of ICT measurement: Comparativeanalysis of e-indexes, Proc.ofthe 7th International Conference on Information andCommunication Technologies in Agriculture, Foodand Environment (HAICTA 2015), Kavala, Greece, 2015, 938-945.

[13]. A. Preda, D. A. Crișan, J. L. Stănică, and A. N. A. Samuel, Transectionalanalysisbetweeninnovationand ICT readinessfortheeuropeanunioncountries,Journal of Information Systems& Operations Management, 10(2), 2016, 393-403.

[14]. J.Zoroja,ImpactofICTsoninnovationactivities:Indi cationforselectedeuropeancountries,Naše gospoda rstvo/OurEconomy, 62(3), 2016, 39-51.

[15]. L. Rokachand O. Maimon, Data miningwithdecisiontrees: Theoryandapplications(2nd ed.)(5 TohTuck Link, Singapore:World Scientific Publishing Co. Pte. Ltd, 2015).

[16]. M. Guller, Bigdataanalyticswithspark: a practitioner'sguidetousingsparkforlargescale data analysis(New York: Apress, 2015).

[17]. D.S.Putler,andR.E.Krider,Customerandbusinessa nalytics:Applieddataminingforbusinessdecisionm akingusing R(Boca Raton, FL: CRS Press, 2015).

[18]. T. M. Khoshgoftaar, E. B. Allen, and J.Deng,Using regressiontreestoclassifyfault-pronesoftwaremodules, in D. ZhangandJ. J. P. Tsai(Eds.), Machine learningapplication in software engineering (5 TohTuck Link, Singapore: World Scientific Publishing Co. Pte. Ltd., 2005) 87-94.

[19]. G.Tutz,Regressionforcategoricaldata(32 Avenue of theAmericas, New York: Cambridge UniversityPress, 2012).

[20]. L.Yang,S.Liu,S.Tsoka, and L. G. Papageorgiou, Regressiontreeapproachusingmathematicalprogra mming, ExpertSystemsWith Applications, 78, 2017, 347-357.

[21]. L. Parziale, O. Benke.,W. Favero, R. Kumar, S. Lafalce., C. Madera, andS. Muszytowski,Enablereal-time analyticsonIBMzsystemsplatform.(Redbook,2016 ,Retrievedfromhttp://www.redbooks.ibm.com/red books/pdfs/sg248272.pdf)

[22]. G. Ceyhan,Üniversite öğrencilerinin yansıtıcı düşünme düzeyleri ve araştırmaya yönelik kaygılarının çeşitli değişkenler açısından CART analizi ile incelenmesi, masterdiss.,University of T.C. Yüzüncü Yıl, Van, TR, 2014.

[23]. C. Kuzey,Veri madenciliğinde destek vektör makinaları ve karar ağaçları yöntemlerini kullanarak bilgi çalışanlarının kurum performansı üzerine etkisinin ölçülmesi ve bir uygulama, doctoraldiss.,University of T.C. Istanbul, Istanbul, TR, 2012.

[24]. Y.YohannesandP.Webb, Classificationandregressiontrees, CART: A user manuel foridentifyingindicators ofvulnerabiltytofamineandchronicfoodinsecurity(Washington,DC:InternationalFoodPolicyResearc hInstitute, 1999).

[25]. [25]S. Sumathiand S. Paneerselvam, Computationalintelligenceparadigms: Theory&ap

plicationsusing MATLAB(Boca Raton, FL: CRS Press, 2010).
[26]. K.KimandJ.Hong,Ahybriddecisiontreealgorithmformixednumericandcategoricaldatainregressionanalysis,PatternRecognitionLetters, 98, 2017, 39-45.
[27]. K. P. Soman, S. Diwakar, and V. Ajay, Data mining: Theoryandpractice (PatparganjIndustrialArea, Delhi: PHI Learning Private Limited, 2009).
[28]. R.A.Berk,Statistical learningfrom a regressionperspective, (2nd ed.) (Cham, Switzerland: Springer International Publishing, 2016).
[29]. C.D. Sutton, Classificationandregressiontrees, bagging, andboosting,in C. R. Rao, E. J. Wegman, andJ. L. Solka (Eds), Data miningand data visualization, (Amsterdam, TheNetherlands: Elsevier B.V., 2005) 303-328.
[30]. J.H.ChoandP.U.Kurup,Decisiontreeapproachforclassificationanddimensionalityreductionofelectronicnosedata,SensorsandActuators B: Chemical, 160, 2011, 542-548.
[31]. O.MaimonandL.Rokach,Decisiontree,inO.MaimonandL.Rokach(Eds.),Thedataminingandknowledgediscoveryhandbook(NewYork:SpringerScience+Business Media, Inc., 2005)165-192.

APPENDIX 1

Table 3. NRI indicators

A. Environment Sub-Index	B. Readiness Sub-Index	C. UsageSub-Index	D. ImpactSub-Index
1st pillar: PoliticalandRegulatory Environment	**3rd pillar: Infrastructure**	**6th pillar: IndividualUsage**	**9th pillar: EconomicImpacts**
1.01 Effectiveness of law-makingbodies	3.01 Electricityproduction, kWh/capita	6.01 Mobile phonesubscriptions/100 pop.	9.01 Impact of ICTs onbusinessmodels
1.02 LawsrelatingtoICTs	3.02 Mobile network coverage, % pop.	6.02 Individualsusing Internet, %	9.02 ICT PCT patents, applications/million pop.
1.03 Judicialindependence	3.03 International Internet bandwidth, kb/s peruser	6.03 Households w/ personalcomputer, %	9.03 Impact of ICTs on organizationalmodels
1.04 Efficiency of legal system in settlingdisputes	3.04 Secure Internet servers/million pop.	6.04 Households w/ Internet access, %	9.04 Knowledge-intensivejobs, % workforce
1.05 Efficiency of legal system in challengingregs	**4th pillar: Affordability**	6.05 Fixedbroadband Internet subs/100 pop.	**10th pillar: SocialImpacts**
1.06 Intellectualpropertyprotection	4.01 Prepaidmobile cellulartariffs, PPP $/min	6.06 Mobile broadbandsubs/100 pop.	10.01 Impact of ICTs on accesstobasicservices
1.07 Software piracy rate, % softwareinstalled	4.02 Fixedbroadband Internet tariffs, PPP $/month	6.07 Use of virtualsocialnetworks	10.02 Internet access in schools
1.08 No. procedurestoenforce a contract	4.03 Internet &telephonycompetition, 0–2 (best)	**7th pillar: Business Usage**	10.03 ICT use&gov'tefficiency
1.09 No. daystoenforce a contract	**5th pillar: Skills**	7.01 Firm-leveltechnologyabsorption	10.04 E-Participation Index, 0–1 (best)
2nd pillar: Business andInnovation Environment	5.01 Quality of educationsystem	7.02 Capacityforinnovation	
2.01 Availability of latesttechnologies	5.02 Quality of math&scienceeducation	7.03 PCT patents, applications/million pop.	
2.02 Venturecapitalavailability	5.03 Secondaryeducationgrossenrollment rate, %	7.04 ICT useforbusiness-to-businesstransactions	
2.03 Total tax rate, % profits	5.04 Adultliteracy rate, %	7.05 Business-to-consumer Internet use	
2.04 No. daystostart a business		7.06 Extent of staffraining	
2.05 No. procedurestostart a business		**8th pillar: GovernmentUsage**	
2.06 Intensity of localcompetition		8.01 Importance of ICTstogov'tvision	
2.07 Tertiaryeducationgrossenrollment rate, %		8.02 Government Online Service Index, 0–1 (best)	
2.08 Quality of managementschools		8.03 Gov'tsuccess in ICT promotion	
2.09 Gov'tprocurement of advancedtech			

Table 3. NRI indicators

A. Environment Sub-Index	**B. Readiness Sub-Index**	**C. UsageSub-Index**	**D. ImpactSub-Index**
1st pillar: PoliticalandRegulatory Environment	**3rd pillar: Infrastructure**	**6th pillar: IndividualUsage**	**9th pillar: EconomicImpacts**
1.01 Effectiveness of law-makingbodies	3.01 Electricityproduction, kWh/capita	6.01 Mobile phonesubscriptions/100 pop.	9.01 Impact of ICTs on businessmodels
1.02 LawsrelatingtoICTs	3.02 Mobile network coverage, % pop.	6.02 Individualsusing Internet, %	9.02 ICT PCT patents, applications/million pop.
1.03 Judicialindependence	3.03 International Internet bandwidth, kb/s peruser	6.03 Households w/ personalcomputer, %	9.03 Impact of ICTs on organizationalmodels
1.04 Efficiency of legal system in settlingdisputes	3.04 Secure Internet servers/million pop.	6.04 Households w/ Internet access, %	9.04 Knowledge-intensivejobs, % workforce
1.05 Efficiency of legal system in challengingregs	**4th pillar: Affordability**	6.05 Fixedbroadband Internet subs/100 pop.	**10th pillar: SocialImpacts**
1.06 Intellectualpropertyprotection	4.01 Prepaidmobile cellulartariffs, PPP $/min	6.06 Mobile broadbandsubs/100 pop.	10.01 Impact of ICTs on accesstobasicservices
1.07 Software piracy rate, % softwareinstalled	4.02 Fixedbroadband Internet tariffs, PPP $/month	6.07 Use of virtualsocialnetworks	10.02 Internet access in schools
1.08 No. procedurestoenforce a contract	4.03 Internet &telephonycompetition, 0–2 (best)	**7th pillar: Business Usage**	10.03 ICT use&gov'tefficiency
1.09 No. daystoenforce a contract	**5th pillar: Skills**	7.01 Firm-leveltechnologyabsorption	10.04 E-Participation Index, 0–1 (best)
2nd pillar: Business andInnovation Environment	5.01 Quality of educationsystem	7.02 Capacityforinnovation	
2.01 Availability of latesttechnologies	5.02 Quality of math&scienceeducation	7.03 PCT patents, applications/million pop.	
2.02 Venturecapitalavailability	5.03 Secondaryeducationgrossenrollment rate, %	7.04 ICT useforbusiness-to-businesstransactions	
2.03 Total tax rate, % profits	5.04 Adultliteracy rate, %	7.05 Business-to-consumer Internet use	
2.04 No. daystostart a business		7.06 Extent of staffraining	
2.05 No. procedurestostart a business		**8th pillar: GovernmentUsage**	
2.06 Intensity of localcompetition		8.01 Importance of ICTstogov'tvision	
2.07 Tertiaryeducationgrossenrollment rate, %		8.02 Government Online Service Index, 0–1 (best)	
2.08 Quality of managementschools		8.03 Gov'tsuccessin ICT promotion	
2.09 Gov'tprocurement of advancedtech			

APPENDIX 2

Country	GII	Predicted_GII	Differences
Albania	28,4	26,32857143	2,071428571
Algeria	24,5	26,32857143	-1,828571429
Argentina	30,2	35,168	-4,968
Armenia	35,1	35,168	-0,068
Australia	53,1	57,3	-4,2
Austria	52,6	57,3	-4,7
Azerbaijan	29,6	35,168	-5,568
Bahrain	35,5	35,168	0,332
Bangladesh	22,9	26,32857143	-3,428571429
Belgium	52	57,3	-5,3
Benin	22,2	26,32857143	-4,128571429
Bhutan	27,9	26,32857143	1,571428571
Bolivia	25,2	26,32857143	-1,128571429
BosniaandHerzegovina	29,6	35,168	-5,568
Botswana	29	26,32857143	2,671428571
Brazil	33,2	35,168	-1,968
Bulgaria	41,4	35,168	6,232
Burundi	20,9	26,32857143	-5,428571429
Cambodia	27,9	26,32857143	1,571428571
Cameroon	22,8	26,32857143	-3,528571429
Canada	54,7	57,3	-2,6
Chile	38,4	35,168	3,232
China	50,6	46,57777778	4,022222222
Colombia	34,2	35,168	-0,968

Cameroon	22,8	26,32857143	-3,528571429
Canada	54,7	57,3	-2,6
Chile	38,4	35,168	3,232
China	50,6	46,57777778	4,022222222
Colombia	34,2	35,168	-0,968
Costa Rica	38,4	35,168	3,232
CôtedIvoire	25,8	26,32857143	-0,528571429
Croatia	38,3	35,168	3,132
Cyprus	46,3	46,57777778	-0,277777778
CzechRepublic	49,4	46,57777778	2,822222222
Denmark	58,5	57,3	1,2
DominicanRepublic	30,6	26,32857143	4,271428571
Ecuador	27,1	26,32857143	0,771428571
Egypt	26	35,168	-9,168
El Salvador	26,6	26,32857143	0,271428571
Estonia	51,7	46,57777778	5,122222222
Ethiopia	24,8	26,32857143	-1,528571429
Finland	59,9	57,3	2,6
France	54	57,3	-3,3
Georgia	33,9	35,168	-1,268
Germany	57,9	57,3	0,6
Ghana	26,7	26,32857143	0,371428571
Greece	39,8	35,168	4,632
Guatemala	27,3	26,32857143	0,971428571
Guinea	17,2	26,32857143	-9,128571429
Honduras	26,9	26,32857143	0,571428571
Hong Kong SAR	55,7	57,3	-1,6
Hungary	44,7	46,57777778	-1,877777778
Iceland	56	57,3	-1,3
India	33,6	35,168	-1,568
Indonesia	29,1	26,32857143	2,771428571
Iran, IslamicRep.	30,5	26,32857143	4,171428571
Ireland	59	57,3	1,7
Israel	52,3	57,3	-5
Italy	47,2	46,57777778	0,622222222
Jamaica	29	35,168	-6,168
Japan	54,5	57,3	-2,8
Jordan	30	35,168	-5,168
Kazakhstan	31,5	35,168	-3,668
Kenya	30,4	26,32857143	4,071428571
Korea, Rep.	57,1	57,3	-0,2
Kuwait	33,6	35,168	-1,568
KyrgyzRepublic	26,6	26,32857143	0,271428571
Latvia	44,3	46,57777778	-2,277777778
Lebanon	32,7	35,168	-2,468
Lithuania	41,8	46,57777778	-4,777777778
Luxembourg	57,1	57,3	-0,2
Macedonia, FYR	35,4	35,168	0,232
Madagascar	24,8	26,32857143	-1,528571429
Malawi	27,3	26,32857143	0,971428571
Malaysia	43,4	46,57777778	-3,177777778
Mali	24,8	26,32857143	-1,528571429
Malta	50,4	46,57777778	3,822222222
Mauritius	35,9	35,168	0,732
Mexico	34,6	35,168	-0,568
Moldova	38,4	35,168	3,232
Mongolia	35,7	35,168	0,532
Montenegro	37,4	35,168	2,232
Morocco	32,3	35,168	-2,868
Mozambique	29,8	26,32857143	3,471428571
Namibia	28,2	26,32857143	1,871428571
Nepal	23,1	26,32857143	-3,228571429
Netherlands	58,3	57,3	1
New Zealand	54,2	57,3	-3,1
Nicaragua	23,1	26,32857143	-3,228571429
Nigeria	23,1	26,32857143	-3,228571429
Norway	52	57,3	-5,3
Oman	32,2	35,168	-2,968
Pakistan	22,6	26,32857143	-3,728571429

Merve Doğruel Anuşlu "Regressiontree Analysis Withcartalgorithm: GII Predictionusing NRI Indicators "International Journal of Engineering Research and Applications (IJERA) , vol. 8, no.6, 2018, pp.61-74