International Journal of Computational Economics and Econometrics

ISSN online: 1757-1189 - ISSN print: 1757-1170
https://www.inderscience.com/ijcee

The use of classification models to identify factors differentiating the competitiveness of the EU-15 and EU-13 countries

Agnieszka Kleszcz

DOI: 10.1504/IJCEE.2021.10043423

Article History:
Received: 14 April 2021
Accepted: 19 October 2021
Published online: 30 November 2022

The use of classification models to identify factors differentiating the competitiveness of the EU-15 and EU-13 countries

Agnieszka Kleszcz

Faculty of Natural Sciences,
Jan Kochanowski University of Kielce,
Kielce, Poland
Email: aakleszcz@gmail.com

Abstract: This paper reports on a study of the Global Competitiveness Index pillars, aiming to differentiate the European Union countries grouped by their accession year in terms of their competitiveness. A linear (regularised logistic regression) and nonlinear (random forests) classifiers are proposed, to model the relationship between multidimensional economic condition indicators and the country's group. The key discriminators of the competitiveness of the EU-15 (accession before 2004) and the EU-13 (accession in or after 2004) are obtained by analysis of feature importance in classification models. Upon study of 12 competitive indicators from the World Economic Reports (2007–2017 edition) we conclude that the highest disparities between the groups of countries can be observed in infrastructure. Innovation, market size and institutions are the next three most important differentiating factors. A major methodological contribution of the paper is the use of explainable statistical models for identifying key features differentiating groups of countries.

Keywords: logistic regression; random forest; European Union; Global Competitiveness Index; GCI; feature importance.

Reference to this paper should be made as follows: Kleszcz, A. (2023) 'The use of classification models to identify factors differentiating the competitiveness of the EU-15 and EU-13 countries', *Int. J. Computational Economics and Econometrics*, Vol. 13, No. 1, pp.110–128.

Biographical notes: Agnieszka Kleszcz is currently a PhD student at the Jan Kochanowski University. Additionally, over the last five years, she has been working on R&D projects related to ICT at the AGH University of Science and Technology. In 2011, she was awarded an MSc in Environmental Engineering from the University of Agriculture in Krakow. Her research interests are focused on EU economic policies, ICT and environmental engineering.

1 Introduction

As many economic sources indicate, competitiveness is a multidimensional concept and has many different interpretations. Olczyk (2016) consolidated the state of the art of academic research on international competitiveness, based on a bibliometric study of the economics literature published over the past 70 years. The importance of the topic was confirmed by the number of publications (1,174 publications by 1,970 authors in 457 journals) and still it remains an unexplained issue. A multitude of factors have been identified in the literature as factors that determine the competitiveness of countries. Simionescu et al. (2021) highlighted the role of innovation, foreign direct investment (FDI), and human capital in supporting competitive European economies. Alternatively, the World Bank identified the factors of: institutions, infrastructure, macroeconomic environment, health and primary education, technological readiness, and market size. There are various frameworks, models and analytical tools presented in literature that can be used in studying relationships between some key factors and national competitiveness. The causal analysis by Neffati (2015) is one example based on robust statistical theory, however the method is designed for time-series, and lacks flexibility to compare groups of countries. The values of linear regression coefficients are also used as a proxy for assessing the impact of variables (Palei, 2015). While this approach in general can lead to inaccurate estimates in the presence of a correlated predictor it can be made rigorous by calculating the *feature importance* for the model. The main contribution of this paper is a proposition to use a classification model with feature importance to identify key factors differentiating countries' competitiveness among EU-13 and EU-15.

Classification techniques are an essential part of statistical learning and data mining applications. When building a statistical classification model, the question of which variables to include often arises. Practitioners have now at their disposal a wide range of technologies to solve this issue (e.g., different categories of algorithms like test-based, penalty-based, screening-based). Feature importance describes how important the feature was for the classification performance of the model. More precisely feature importance is a quantification of the individual contribution of the corresponding feature towards the effectiveness of a particular classifier (Saarela and Jauhiainen, 2021; Desboulets, 2018). The methods span from simple randomisation (Altmann et al., 2010) to sophisticated interpretation of a model's internals (Guyon et al., 2002). Some classifiers, e.g., tree-based models, have natural means of measuring importance, while others can remove unimportant features through feature selection, e.g., logistic regression with Lasso.

In Europe, the enlargement process of the EU led to significant heterogeneity, which affected development and competitiveness under the EU policy and created high discrepancies (Simionescu et al., 2021). Those discrepancies are still observed, especially between EU-15 and EU-13, despite one of the EU policies priority is economic and social cohesion. Such cohesion can only be achieved when the sources of disparities are identified and removed. Literature studies identify that the directions of development of research on competitiveness are related to the search for an answer to the question of what constitutes the source of the advantage of one economy over another. Despite many publications and approaches presented in the literature, analysis of multiple studies leads to the conclusion that this phenomenon still needs additional research. It is also difficult to find in the literature studies showing which factors

determine the difference in competitiveness between EU-13 and EU-15. This paper attempts to fill the gap in this respect by utilising classification models. In this particular study we experiment with the method by finding the importance of 12 competitiveness pillars from the GCI index in order to differentiate between two groups of European Union (EU) countries: EU-15 (countries which entered the EU before 2004) and EU-13 (countries that joined the EU in and after 2004).

The motivation and contributions of this study are three-fold:

1 we empirically test the classification performance of the linear and nonlinear classifier
2 through using logistic regression with Lasso regularisation and random forest and the concept of feature importance we compare the explanations provided by these different classifiers
3 we propose such an approach as a valid methodology for identifying differences between groups of countries, that may be easily extended beyond EU or competitiveness analysis.

Our analysis focuses on identifying needs at the national level which should be strengthened to decrease the disparities between EU-15 (characterised by the stronger competitive position) and EU-13 countries.

The paper is organised as follows. Section 2 includes a literature review and describes competitiveness measures, (in particular GCI) along with its associated dataset and some basic data analysis. Research methodology is described in Section 3, while the main part of the paper describes the results of the analysis in Section 4. Section 5 includes a comparison of our results with other authors and Section 6 summarises the main conclusions drawn from the analysis.

2 Literature review

Competitiveness has been the topic of economic research and analysis since the second half of the 20th century among scientists, economic politicians and business. The competitiveness can be measured in various ways: analysing one or several factors of competitiveness, using theoretical models of competitiveness or creating composite indices. It is widely believed to be a complex phenomenon, hence its discussion requires the use of various criteria and methods of measurement. Roszko-Wójtowicz and Grzelak (2020) presented assessment of the competitiveness within the EU-15 and EU-13 groups as well as the Visegrad group economies in 2005–2018 based on a selected set of diagnostic variables referring to the concept of macroeconomic stabilisation pentagon. A linear ordering of objects was proposed using the reference Hellwig method. Results present comparative assessment between individual EU countries based on diagnostic variables in selected years. Author emphasised that the EU-15 countries dominate the top of competitiveness rankings, irrespectively of the unit of time under consideration. Besides dynamics of changes in the synthetic measure of competitiveness of the EU-28, and especially the EU-15, is definitely lower than the dynamics of changes in the EU-13 and the V4 countries. Noticeable greatest advancement in the competitiveness of the EU-13 countries was observed in 2018. Despite of that, a comparison of the values

of individual measures indicates the existence of significant differences between EU-13 and EU-15. However, the presented assessment of the competitive position of the EU-15 and EU-13 groups does not exhaust the complexity of the issue, and it is only one of its threads that make up the entire assessment system.

Maintaining global competitiveness has been one of the main challenges facing countries around the world in recent years. Measurement of competitiveness can be approached from several points of view (Širá et al., 2020). Economic literature presents several competitiveness indexes which measure competitiveness at the country or a region level. Those competitiveness indices use, i.e., the different number of key factors, weighting them differently and covering a different number of countries and data sources. The Global Competitiveness Index (GCI) (published by the World Economic Forum) and the World Competitiveness Yearbook (by the Institute for Management Development) are the best-known indices. On the other hand, the regional competitiveness (NUTS-2 level) across the EU has been measured over the past ten years by the regional competitiveness index (RCI) (Annoni and Dijkstra, 2019).

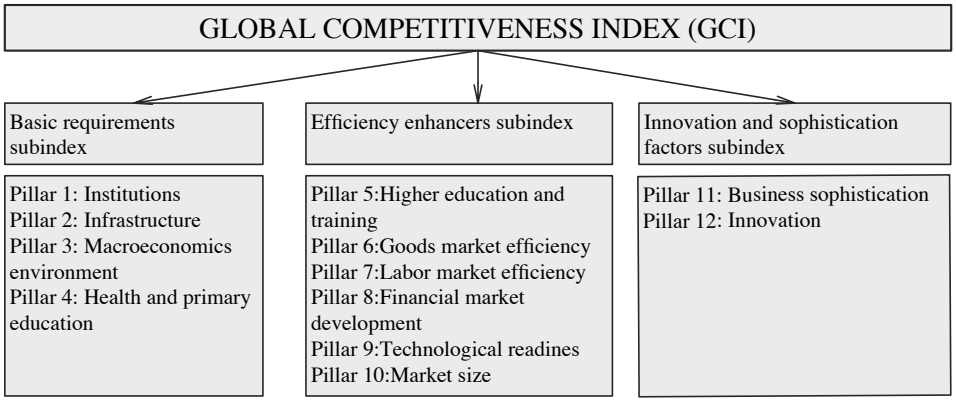
Multiple studies analyse global competitiveness. Many authors are focused on the overall GCI score evaluation and imply various statistical methods to suggest recommendations to improve the current competitive position of countries. The linear regression models to analyse GCI are popular methods presented in many papers (Bucher, 2018; Kiselakova et al., 2019; Marčeta and Bojnec, 2020; Pérez-Moreno et al., 2015; Ivanova and Cepel, 2018). Others analyse the evolution of the GCI by country. Some studies analyse also the dependence of socioeconomic inequality on regional differences of individual countries of Europe (Marčeta and Bojnec, 2020). Although a wide range of policy actions have been taken, both at the European and national levels, to improve the economic resilience of EU economies, many challenges still remain. Significant regional differences remain and, more importantly, are not improving in some EU member states (Annoni and Dijkstra, 2019).

In numerous studies, many factors have been shown to have significant effects on competitiveness most of this work has been incorporated into the GCI. The GCI of the World Economic Forum (WEF) is one of the best-known competitiveness indices widely used among academics, policy-makers and business leaders, which measures the microeconomic and macroeconomic foundations of national competitiveness. The forum defines national competitiveness as the set of institutions, policies and factors that determine the level of productivity of a country. The GCI includes statistical data from internationally recognised organisations like the International Monetary Fund (IMF), International Telecommunication Union, World Bank, World Health Organization and various United Nations' specialist agencies, including UNESCO. The pillars expressing the basic requirements reflect the general characteristics of factor-driven economies, while efficiency-enhancers indicate the general characteristics of efficiency-driven economies. Furthermore, innovation and sophistication factors show the intensity of entrepreneurship and innovation in each country (Petrarca and Terzi, 2018; Şener and Delican, 2019; Schwab, 2017) (Figure 1).

Competitiveness in this case depends primarily on good-functioning public and private institutions (pillar 1); a good-developed infrastructure (pillar 2); a stable macroeconomic environment (pillar 3); and a healthy workforce and primary education (pillar 4). As a country becomes more competitive (with higher productivity and wages), they will then move into the efficiency-driven stage of development. At this point, competitiveness is increasingly powered by higher education and training (pillar 5);

efficient goods markets (pillar 6); good-functioning labour market efficiency (pillar 7); developed financial markets (pillar 8); the ability to use the benefits of existing technologies (pillar 9); and a large domestic or foreign market (pillar 10). Finally, as countries move into the innovation and sophistication-driven stage, companies must compete by producing new and different goods using the most sophisticated production processes (pillar 11), and by innovating new ones (pillar 12) (Nallari and Griffith, 2013). Pillars are measured using normalised scale from 0 to 7 (Schwab, 2017). Since 2018 WEF introduced new GCI 4.0 with slightly different pillars and a different scoring regime. The GCI 4.0 introduces a new progress score ranging from 0 to 100 (Schwab, 2018). Because of this, to keep data coherent our analysis is up to 2017 (edition 2017–2018).

Figure 1 Sub-indices and pillars of GCI



Source: Own elaboration based on Schwab et al. (2013)

3 Research methodology

International comparison requires us to indicate factors that affect the success of developed economies in an international comparison. This paper determines which factors determine the competitiveness position in EU countries based on the 12 pillars of the GCI. We use two classifiers: linear (logistic regression) and nonlinear (random forests), for problem classification and feature importance estimation. All calculations were performed with the use of Python software.

3.1 Dataset

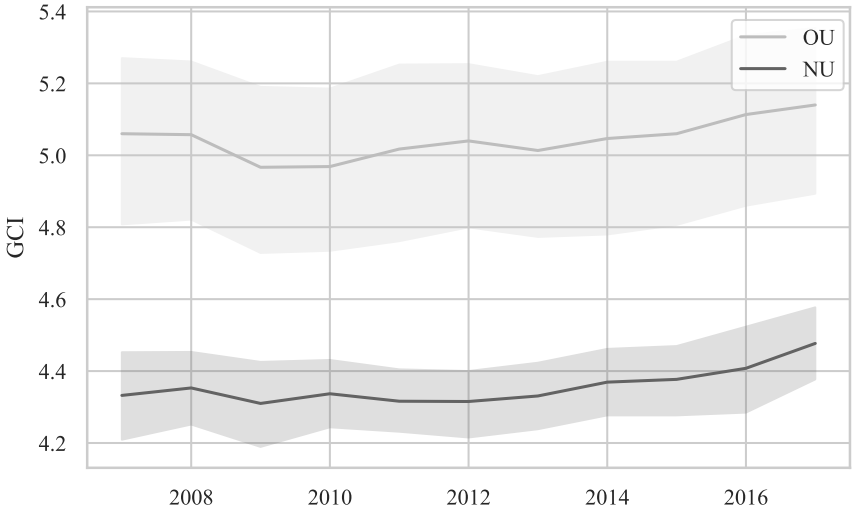
A dataset has been created that characterises the level of development of 28 EU countries based on economic indicators from the World Economic Reports (Schwab and Porter, 2008; Schwab, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017). Data are from 2007 (2007–2008 edition) to 2017 (2017–2018 edition). For each year (edition) there were 28 observations, representing respectively each of the 28 EU countries, with a total of 308 observations collected. From 28 EU countries, EU-15 countries are labelled as the 'old union' (OU), countries which entered the EU before 2004:

Germany, Netherlands, Finland, Sweden, UK, Denmark, Austria, Belgium, France, Ireland, Luxembourg, Spain, Portugal, Italy, Greece.

The remaining EU-13 countries are labeled as the 'new union' (NU), countries that joined the EU in and after 2004: Czech Republic, Estonia, Lithuania, Malta, Poland, Bulgaria, Slovenia, Latvia, Cyprus, Hungary, Romania, Slovakia, Croatia. Accordingly, OU class consists of 165 observations and NU: 143. The dataset was organised into a matrix of the shape of 308 observations and 13 columns. Out of the 13 columns, 12 contain the pillar values (features) and remaining one contains binary information about the class: OU or NU. The 12 GCI pillars represented individual competitiveness features: Market size, Business sophistication, Innovation, Institutions, Infrastructure, Macroeconomic environment, Health and primary education, Higher education and training, Goods market efficiency, Labour market efficiency, Financial market development, Technological readiness.

Figure 2 shows time evolution of GCI since 2007 in two classes (OU and NU). We observe the average and its 95% confidence interval and consequently observe high disparity between countries.

Figure 2 Evolution of average GCI index between 2007–2017 (with 95% confidence intervals for average) grouped into the old and new union

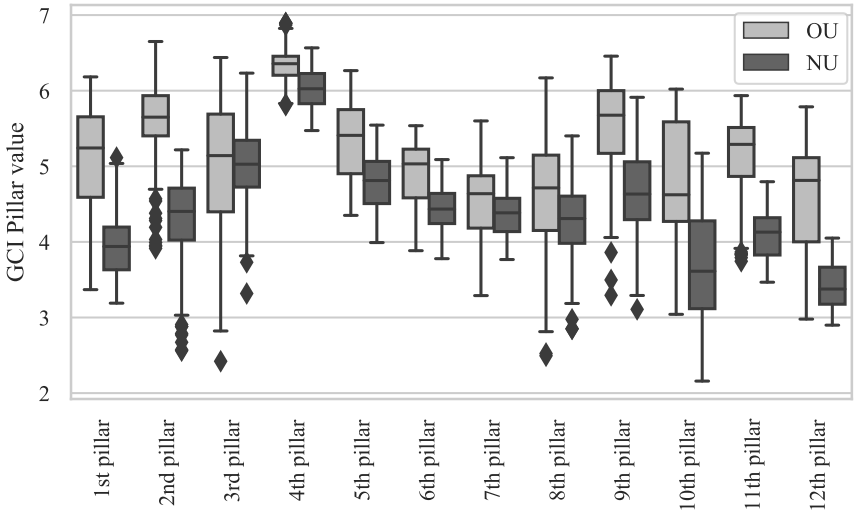


Notes: OU – old EU, NU – new EU.

Source: Own calculations based on the GCI edition 2007–2008 through 2017–2018

Figure 3 depicts a basic statistical visualisation based on a box-plot for 12 pillars divided into two classes (OU and NU). We can observe differences in each country's competitiveness expressed by the 12 GCI pillars depending on the class (i.e., OU or NU). In most cases we can observe differences between EU-15 and EU-13, with only a few pillars overlapping with each other: i.e., 3rd, 7th and 8th. Finding their importance is of primary interest.

Figure 3 Distribution of pillars for two classes (OU and EU) based on box-plot



Notes: 1st pillar: institutions, 2nd pillar: infrastructure, 3rd pillar: macroeconomic environment, 4th pillar: health and primary education, 5th pillar: higher education and training, 6th pillar: goods market efficiency, 7th pillar: labour market efficiency, 8th pillar: financial market development, 9th pillar: technological readiness, 10th pillar: market size, 11th pillar: business sophistication, 12th pillar: innovation.

Source: Own calculations based on the GCI edition 2007–2008 through 2017–2018

Many studies indicate that the pillars of GCI are potentially interrelated. Pillars are not independent they tend to strengthen each other, and a weakness in one area could have a negative effect on other pillars (Schwab, 2014). Figure 4 shows a correlation matrix which is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between the two variables. We can observe that some pillar variables are highly correlated with each other. The correlation is mostly positive.

3.2 Multiple logistic regression

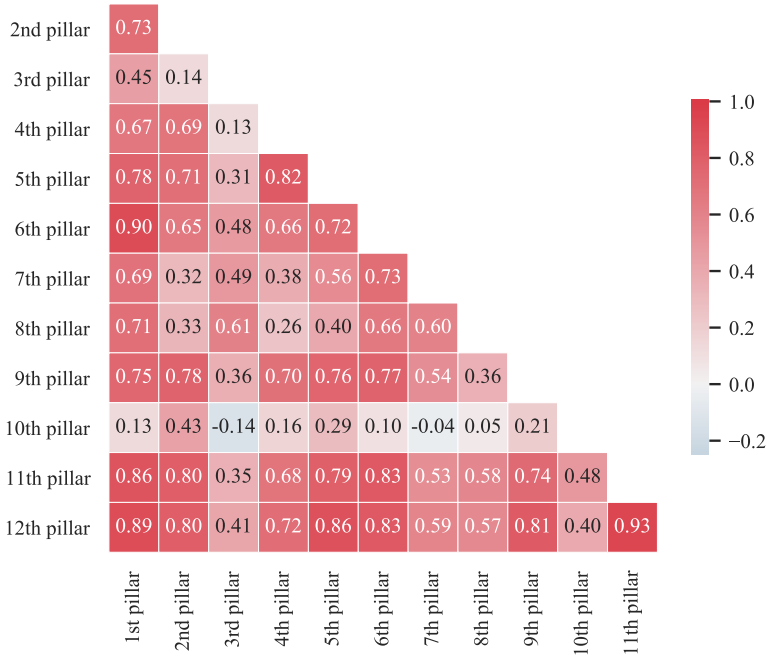
Logistics regression is one of the most common and useful methods for solving the classification problem (James et al., 2013; Molnar, 2019). In its simplest form logistic regression is designed for two-class problems and models the relationship between one dependent binary variable and independent variables. The model produces results in a binary format that is used to predict the outcome of a categorical dependent variable. Such a binary response is typically coded as 0 and 1. Attention then focuses on estimating the conditional probability $\Pr(Y = 1|\, X = x)$.

With the binary response coded in the form $Y \in \{0, 1\}$, the linear logistic model is often used, it models the log-likelihood ratio as the linear combination of predictors:

$$\log \frac{\Pr(Y = 1|\, X = x)}{\Pr(Y = 0|\, X = x)} = \beta_0 + \beta^T \mathbf{x} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \tag{1}$$

where $\mathbf{x} = (X_1, X_2, \ldots X_p)$ is a vector of predictors, $\beta_0 \in \mathbb{R}$ is an intercept term, and $\beta \in \mathbb{R}^p$ is a vector of regression coefficients (Bishop, 2016; Hastie et al., 2015). Those coefficients are estimated by using the maximum likelihood estimation (MLE) approach. Simplicity and a plethora of theoretical results make this model highly useful and popular.

Figure 4 Correlation matrix between 12 GCI pillars (see online version for colours)



Note: As in Figure 3.

Source: Own calculations based on the GCI edition 2007–2008 through 2017–2018

3.3 The Lasso

Lasso (least absolute shrinkage and selection operator) is a regularisation technique commonly used with linear models, e.g., logistic regression for feature selection. As the regularisation, it weakens the model by setting constraints on the weights. The less powerful model is less prone to overfitting while still being able to model high dimensional dependencies. In the case of the Lasso regularisation, a particular constraint is the L_1 norm $|\beta|_1$ (scaled by a coefficient λ) of the weights added to the negative log-likelihood function ℓ:

$$\text{loss} = \ell + \lambda |\beta|_1 . \tag{2}$$

The model is fitted by minimising the loss. The remarkable property of this regularisation is the ability to learn sparse models, namely some of the weights are set to zeros during the training. This can be interpreted as feature selection since the predictor

with zero weight is not used in the model and can be removed totally from the dataset. The parameter λ (and its inverse $C = 1/\lambda$) used for scaling the L1 norm also has an interesting interpretation which is that when it is set to zero, the Lasso regularisation vanishes. On the other hand, a large value of the coefficient λ makes the L1 penalty component dominant in the loss function which results in most weights being set to zeros and only a few features used for prediction. Clearly, we can use this parameter to control the flexibility of the model. Small value – highly flexible model, large value – inflexible model. Typically, the best value of λ for a given dataset is determined from cross-validation.

In this paper, however, we propose to deliberately use the regularisation coefficient to construct a family of linear models by continuously varying their flexibility and show how to use feature selection for obtaining feature importance. This family is used to estimate feature importance by gradually adding new features to the model (Kolassa, 2017; James et al., 2013; Murphy, 2012).

3.4 Random forest

Random Forest is an ensemble machine learning algorithm based on multiple decision trees (DTs) for making decisions. A single DT is an algorithm making decisions, based on multiple questions about predictors just like in the '21 questions game'. In the random forest method, each node in the DT uses only a random subset of features to calculate the output. This procedure decorrelates trees. The random forest then combines the output of individual DT to generate the final output.

Training of the random forest is a multistep process. At each step, the original training data is randomly sampled-with-replacement forming bootstrapped samples. These bootstrap samples are then fed as training data to multiple DTs. Each of the trees is trained individually on these bootstrap samples. The final result of the ensemble classification model is determined by a majority vote from all the tree classifiers. This concept is known as bagging or bootstrap aggregation. Since each DT takes a different set of training data as input, the variations in the original training dataset do not affect the final result obtained from the aggregation of DTs. Therefore, bagging reduces variance without changing the bias of the complete ensemble.

The random forest has two useful features that made them highly popular before the deep learning revolution. The first one is the out-of-bag (OOB) validation which uses OOB observations, i.e., the remaining observations not used to fit a given DT. OOB samples are perfect to estimate generalisation error without wasting precious samples. Additionally, the OOB score is calculated using only a subset of DTs not containing the OOB sample in their bootstrap training dataset.

The second, useful feature of random forest is predictor (feature) importance assessment. During the training, each tree in the forest selects a predictor to split on, the order of predictors depends on the quality of data split measured by an impurity function (e.g., Gini index). The impurity function measures how well two sets of labelled objects are separated. The lower the impurity, the better the separation is. The feature importance is measured as impurity decrease weighted by the probability of reaching a given node in the tree. The higher the impurity decrease, the more important feature is (Kolassa, 2017; James et al., 2013; Palczewska et al., 2014).

3.5 Bootstrap

Bootstrap is a statistical technique used for confidence interval calculations and statistical hypothesis testing. It is a resampling based method, so the distribution of a statistic is estimated by multiple re-computations of the statistic from randomly sampled named bootstrap samples.

Given a dataset on n observations the bootstrap confidence interval calculation can be performed using the following steps:

1 select n random (with repetition) observations to construct a bootstrap sample $X^{(*1)}$

2 recompute the statistics of interest $\hat{\theta}^{*1}$ on the new dataset $X^{(*1)}$

3 repeat 1 and 2 (get i^{th} estimate) B times (99+).

The final estimator value is given by:

$$\hat{\theta} = \frac{1}{B}\sum_{i} \hat{\theta}^{*i}, \tag{3}$$

where $\hat{\theta}$ is the estimate. In this work: the feature importance. The confidence interval can be obtained from the standard deviation of $\hat{\theta}^{*i}$. If the distribution of the statistics is skew, a quantile-based confidence interval is a better choice. In this case, the interval is defined as $(q_{0,025}, q_{0,975})$ where q_p is a $p-$quantile of the bootstrapped estimates $\{\hat{\theta}^{*i}\}$.

4 Results

Below we present results from two models: linear and nonlinear. The first model utilised Logistic Regression and the second used random forests. Firstly, for the classification problem and secondly for feature importance estimation.

4.1 Logistic regression and random forests model for classification of countries in the OU and NU

With logistic regression and random forests, we found very similar results for the classifications of 'old' or 'new' EU. For the classification problem, logistic regression and random forests were evaluated through 10-fold cross-validation.

The logistic regression model gives the probability of a country belonging to the OU class (strength of membership to OU class) as a following formula:

$$\Pr\left(Y = \text{OU} \mid X = \text{Pillars}\right) = \frac{e^{\beta_0+\beta_1\text{Pillar}_1+\cdots+\beta_{12}\text{Pillar}_{12}}}{1+e^{\beta_0+\beta_1\text{Pillar}_1+\cdots+\beta_{12}\text{Pillar}_{12}}}, \tag{4}$$

where X_i is the i^{th} pillars. If the probability is above 0.5 or equivalently

$$\beta_0 + \beta_1\text{Pillar}_1 + \cdots + \beta_{12}\text{Pillar}_{12} \ > 0,$$

the country is classified as old union, otherwise the country is classified as new union.

The most straightforward indicator of classification accuracy is the ratio of the number of correct predictions to the total number of predictions (or observations). We evaluate the model using model evaluation metrics such as accuracy, precision and recall. Both methods show high efficiency (above 97%) but the logistic regression model gives slightly better results for these evaluation metrics. The most common metrics used for classifiers are in Table 1.

Table 1 Assessment of classification models logistic regression and random forests

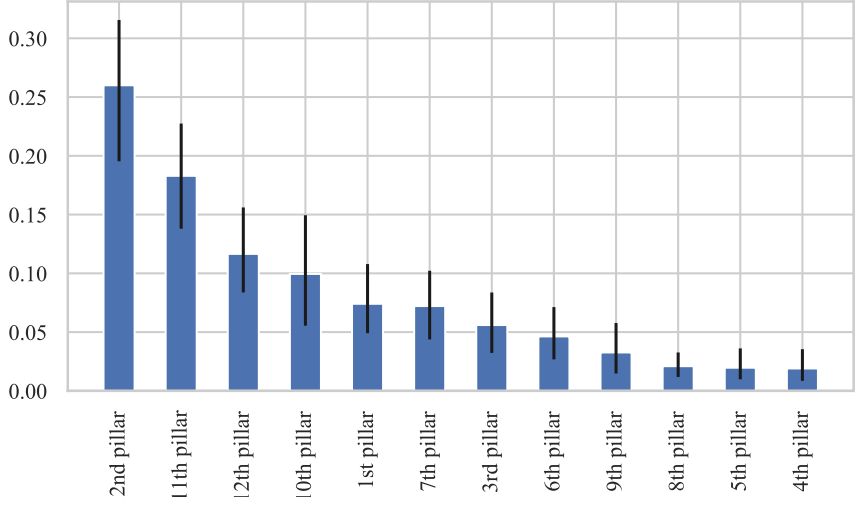
Model evaluation metrics	*Logistic regression*	*Random forest*
Precision	0.995	0.975
Recall	0.992	0.980
Accuracy	0.994	0.977

Source: Own calculations

4.2 Determining feature importance with random forest

A random forest classifier gives a unique chance to estimate the importance of the predictors (feature importance). The feature (pillars) importance's are visualised in Figure 5. The blue bars represent the feature importances, namely the average value of 299 bootstrap samples. The vertical lines represent a 95% confidence interval calculated as quantiles of the bootstrap samples.

Figure 5 Feature importance's plot for random forests (see online version for colours)



Note: As in Figure 3.

Source: Own calculations based on the GCI edition 2007–2008 through 2017–2018

The most important feature is the 2nd pillar – infrastructure with a value 0.26, as it was expected from the distribution analysis (box-plot) in Figure 3. The other highly important features values also presenting noticeable gap are: 0.18 – 11th pillar: business

sophistication, 0.11 – 12th pillar: innovation, 0.10 – 10th pillar: market size. Remaining features are less informative.

The most interesting results are those for the 3rd, 7th and 8th pillars as their distributions overlap (see Figure 3). The random forest model predicts the following order of importance: 7th, 3rd, and 8th. However, the confidence intervals are quite high and overlap each other so pillars 7 and 3 can be considered equally important. Pillar 8 is noticeably less important, and it belongs to a group of least import features containing pillar 5 and pillar 4 as well.

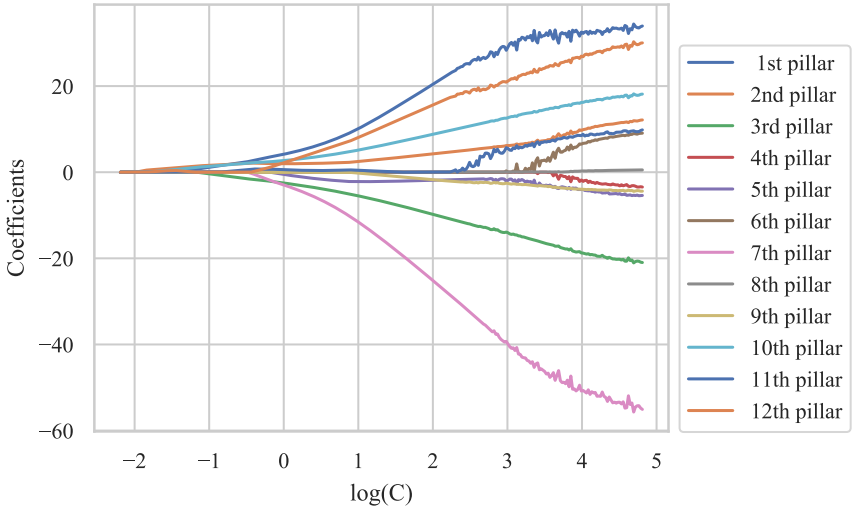
4.3 Logistic regression with Lasso regularisation

The logistic regression model with Lasso regularisation was trained. In Figure 6, we can see the example paths of coefficients as a function of C. The lower the value of parameter C, the more coefficients are shrunk to zero. Having fewer features included makes the model more simple and interpretable. The magnitude of feature coefficients can be interpreted as the importance of that feature, a larger coefficient meaning the feature had more relevance in the classification. In addition, the direction of the coefficient tells whether the feature increases or decreases the probability of belonging to a certain class. On the left-hand side of the figure (strong regularisation), all the coefficients are 0. When regularisation gets progressively looser, coefficients can get non-zero values one after the other. Thus, for highly regularised models only a small fraction/part of the feature has non-zero coefficients. These are the most important features of our model. When regularisation strength is decreased, more pillars get non-zero coefficients. In the experiment, we test 500 regularisation path coefficients logarithmically spaced in the range 10^{-5}–10^{2}. The importance of the feature is defined as a fraction/part of the cases, the feature had a non-zero coefficient value. If the feature is important its importance will be close to one. On the other hand, a feature that appears only in weakly regularised models will have importance closer to zero.

In Figure 7, we present feature importances with confidence intervals obtained for 299 bootstrap samples. The most important feature is again the 2nd pillar as it was expected from the distribution analysis in Figure 3. The other highly important features are 10th pillar: market size, 1st pillar: institutions, 3rd pillar: macroeconomic environment and 12th pillar: innovation (see Figure 7). An interesting result is for the 3rd pillar macroeconomic environment. From the *box-plot* (see Figure 3) we may expect this pillar to be not important for differentiating EU countries, however, its coefficient is mostly non-zero in the logistic regularisation path. Remaining features are less informative. Furthermore, we can observe that the 11th pillar: business sophistication, 5th pillar: higher education and training, 9th pillar: technological readiness, 8th pillar: financial market development, 4th pillar: health and primary education, 6th pillar: goods market efficiency have high uncertainty. It means that bootstrap samples for those pillars are highly scattered around the mean.

Comparing feature importance obtained from logistic regression with Lasso regularisation and random forests we notice different importance characteristics. Random forest importances are convex-like which means they have high resolution for the most important features. On the other hand, Lasso importance features are concave-like which results in increased resolution for the least average important features.

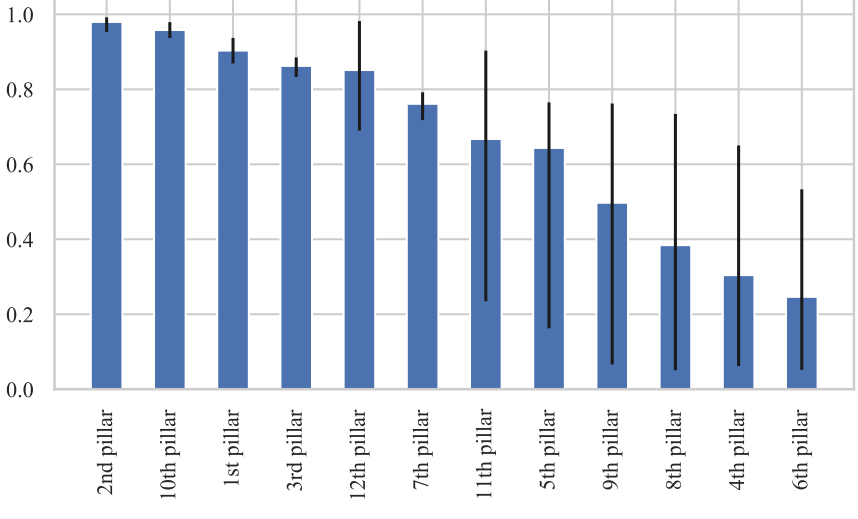
Figure 6 Coefficient paths for L1-regularised logistic regression (see online version for colours)



Note: As in Figure 3.

Source: Own calculations based on the GCI edition 2007–2008 through 2017–2018

Figure 7 Feature importance's plot for logistic regression with Lasso regularisation (see online version for colours)

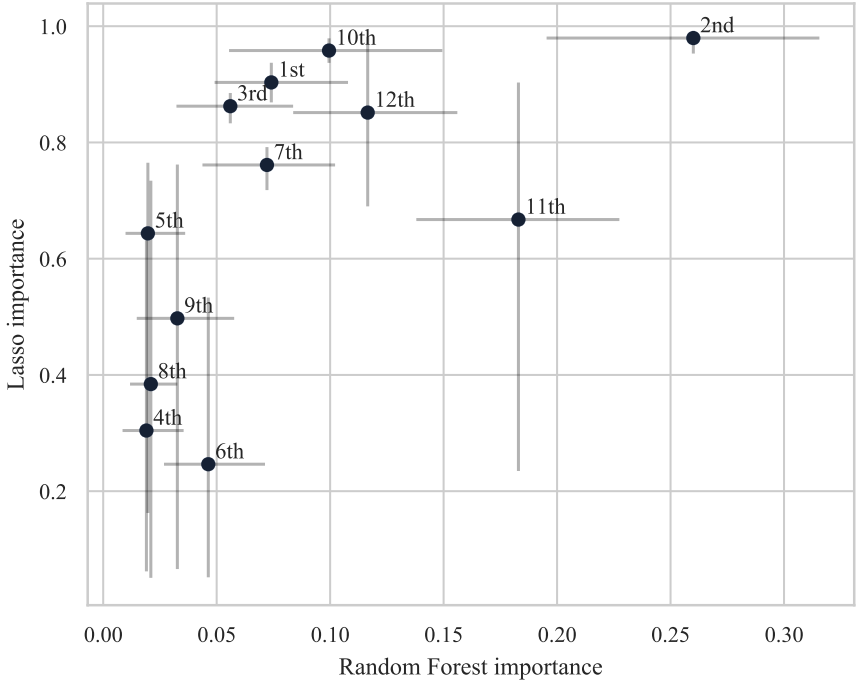


Note: As in Figure 3.

Source: Own calculations based on the GCI edition 2007–2008 through 2017–2018

Figure 8 shows feature importances for the two methods in the scatterplot. Both feature importance estimators indicate that 2nd pillar is the most important. Other pillars are ordered differently, however, the overall importance is quite similar.

Figure 8 Scatterplot of feature importance



Note: As in Figure 3.

Source: Own calculations based on the GCI edition 2007–2008 through 2017–2018

The plot simplifies the interpretation of the results. For every pillar, the pillars on the right and above are estimated to be more important by both models (the feature is dominated by all of them). For example, the average importance of the 7th pillar is dominated by pillars: 1st, 2nd, 10th and 12th. In the second example, the 2nd pillar is more important than the 11th (2nd it is located above and to the right of 11th. However, when confidence intervals overlap, we cannot make a clear statement about the dominance.

Besides this, from comparison of two models we can identify two groups of features. Each member of a group of pillars containing the 1st, 3rd, 7th, 10th, 12th is more important than any pillar from the group containing the 4th, 5th, 6th, 8th, 9th. Within the groups, the confidence intervals are so high, that we cannot make any significant claims and only the averages can be used for rough estimates. Finally, the figure shows how Lasso importance and random forest importance have different resolutions for different levels of importance. Random forest has the highest resolution for the most important features, while Lasso is best at differentiating the least important ones.

5 Discussion

The two methods for feature importance studied in this paper have different properties that can make one more preferable than the other. As the pros and cons can be problem-dependent, there is no clear winner here. The concept of feature importance for random forest is deeper and more technical when compared to Lasso, where the feature importance in the method emerges from the feature selection ability. For example, the Lasso-based method requires the range of regularisation coefficients as a parameter. Furthermore, if the underlying problem is heavily nonlinear, the random forest-based model could be easier to design as the method is nonlinear. Having said that we must point out that random forest could be more computationally demanding compared to Lasso. Their branching behaviour does not benefit well from modern accelerators like GPU. On the other hand, fitting a linear model involves dense operations that are suitable for such devices. It is worth noting that other feature importance methods like the one based on support vector machine (Guyon et al., 2002) or simple permutation-based algorithms can also be applicable and even preferable for some problems. For the particular problem of EU membership, we found a similar performance of both methods, however, they offer slightly different resolutions at the extreme ends of the importance. Regularised logistic regression differentiates the least important factors whereas random forest makes the most important features highly distinctive.

According to results from our models (random forest and logistic regression with Lasso regularisation) significant factors (confirmed by both methods) for increasing EU countries' competitiveness are as follows: 1st – institutions, 2nd – infrastructure, 10th – market size and 12th – innovation. We found in the literature; the importance of these pillars was highlighted by other authors as well.

- 1st – institutions: Investment, production and societal distribution of wealth and benefits is strongly influenced by legally binding rules, laws, and constitutions of institutions: these legal constraints affect the quality of a country's public institutions and the way they interact with individuals, government, and other businesses, which in turn affects investment decisions, competitiveness, and growth (Schwab, 2013).
- 2nd pillar – infrastructure: Extensive and efficient infrastructure is critical for ensuring the effective functioning of the economy: infrastructure refers to the basic physical infrastructure consisting of transport infrastructure (including high quality roads, railroads, ports, and air transport), telecommunication infrastructure and energy infrastructure. This infrastructure creates benefits for a large number of users. The efficient infrastructure supports economic growth, improves quality of life, and it is important for national security (Palei, 2015; Schwab, 2016).
- 10th pillar – market size: The size of the market influences productivity since large markets allow firms to utilise economies of scale. Nowadays large markets allow firms to benefit from economies of scale, international markets have become a substitute for domestic markets. Exports are the substitute for domestic demand and thereby serve as determinants of the market size of firms in a foreign country (Schwab, 2017; Ekici et al., 2019).

- 12th pillar – innovation: The role of innovation and its role in growth has been discussed in the literature (Ivanova and Cepel, 2018; Simionescu et al., 2021). Innovation is the main determinant of growth and performance in the global economy. It gives origin to new technologies and new products, transforming the conditions of production of goods and provision of services, it boosts productivity, creates jobs and improves the global competitiveness of the nation and the standard of life of citizens. Also, it is confirmed by a number of empirical studies applied to some countries (LeBel, 2008). Authors highlighted the role of innovation, that these are new production technologies or new products, in economic growth (Neffati, 2015; Razavi et al., 2012; Sofrankova et al., 2017). In Global Innovation Index 2019 rankings we can observe that 'old' EU countries have a better score then 'new' EU (except Greece – which is on the last 41st position). All EU countries in the ranking belong to the 41 most innovative countries out of 129 world economies (Global Innovation Index 2019, 2019).

It is worth mentioning that pillars of GCI are interrelated. Hence, when we identified important pillars they are in correlation with others and they tend to strengthen each other and vice-versa.

6 Conclusions

Understanding existing socio-economic countries' differentiation should be considered when building a national competitive strategy. Many studies analyse countries' competitiveness from different perspectives to find various dependencies. Literature studies show that comparative analyses of national competitiveness are mainly based on very broad composite indices such as the GCI, where a large number of variables is combined to produce a single composite competitiveness measure. Several publications analyse country competitiveness measured by the GCI in context to analyse factors that determine competitiveness and to identify the relation and mutual impact with other factors (i.e., human development index, gross domestic product by purchasing power parity) (Bucher, 2018; Ivanova and Cepel, 2018; Kiselakova et al., 2019; Marčeta and Bojnec, 2020; Pérez-Moreno et al., 2015).

This study aimed to develop explainable classification models for labelling countries as EU-15 or EU-13 using 12 GCI pillars as predictors. The data was collected from eleven years from the World Economic Reports (editions 2007–2017). Two statistical learning models: Logistic Regression and Random Forest, fit the data with high prediction accuracy. However, the main interest is not in the model itself, but rather in finding the pillars of the GCI differentiating EU-15 and EU-13 countries' economies the most.

In Random Forest, we can estimate feature importance, which helps in selecting the most contributing features for the classifier. The complementary approach based on Lasso regularisation was proposed for linear regression. The obtained pillar importance slightly depends on the method, however, both models support the observation that the most important differentiating factor is the 2nd pillar – infrastructure. Further important pillars differentiating EU countries according to Random Forest are Business sophistication, Innovation, and Market size. While the Logistic Regression model orders pillars as follows: Infrastructure, Market size, Institutions, Macroeconomic environment. On the other side, both models found the following group of pillars: 4th – health and

primary education, 5th – higher education and training, 6th Goods market efficiency, 8th – financial market development, 9th- Technological readiness, as having the least important features.

Because estimates of the importance have high variance, one cannot rely on the average value only. Detailed analysis of confidence intervals revealed that we can distinguish groups of features with significantly different importance. The within group variance of the importances is high so they cannot be compared with high significance. In view of the above, the limitation of our results is a high dispersion of importances for some pillars that cause difficulties in arranging them in ascending order, therefore, we cannot make a significant conclusion for those groups of pillars. This may be seen as a limitation of the utilised method and the results can be seen more as guideline instead of an exact rule for deducing differences between two groups of EU countries. Further research could be extended by using different feature importance estimation techniques or introducing causal modelling. However, from the analysis, we can conclude that EU-13 countries could match competitiveness of EU-15 countries by boosting the following pillars: 1st – institutions, 2nd – infrastructure, 10th – market size and 12th – innovation.

Overall, significant national disparities exist, with Northern and Western Europe performing strongly compared to a lagging Central and Eastern Europe. Results from the analysis performed identify the important most key features which divide more competitive EU-15 with a lagging EU-13. Indicating which areas in EU-13 should be strengthening to decrease disparities could be a functional insight and valuable information for policymakers. Such policymakers' knowledge could help define the right policies such as future initiatives, strategies or funding programmes to reduce disparities between the member states. Also, the results could be valuable for policymakers working on EU Cohesion Policy which aims to strengthen economic, social and territorial cohesion in the European Union to correct imbalances between countries and regions. It is worth mentioning that the method proposed in this paper itself is general enough to be broadly applied in econometrics beyond the topics reported in the paper, e.g., the government can use this method to identified key disparities between regions.

References

Altmann, A., Toloşi, L., Sander, O. and Lengauer, T. (2010) 'Permutation importance: a corrected feature importance measure', *Bioinformatics*, Vol. 26, No. 10, pp.1340–1347.

Annoni, P. and Dijkstra, L. (2019) *The EU Regional Competitiveness Index 2019*, pp.1–42, European Commission.

Bishop, C. (2016) *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer New York.

Bucher, S. (2018) 'The Global Competitiveness Index as an indicator of sustainable development', *Herald of the Russian Academy of Sciences*, Vol. 88, No. 1, pp.44–57.

Desboulets, L.D.D. (2018) 'A review on variable selection in regression analysis', *Econometrics*, Vol. 6, No. 4, pp.1–27.

Ekici, Ş.Ö., Kabak, Ö. and Ülengin, F. (2019) 'Improving logistics performance by reforming the pillars of Global Competitiveness Index', *Transport Policy*, Vol. 81, pp.197–207 [online] https://www.sciencedirect.com/science/article/pii/S0967070X18305456.

Global Innovation Index 2019 (2019) *Global Innovation Index 2019: Creating Healthy Lives – The Future of Medical Innovation*, World Intellectual Property Organization (WIPO), Geneva, ISBN-13: 979-1095870142.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) 'Gene selection for cancer classification using support vector machines', *Machine Learning*, Vol. 46, No. 1, pp.389–422.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015) *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press, University of California, Berkeley, USA.

Ivanova, E. and Cepel, M. (2018) 'The impact of innovation performance on the competitiveness of the visegrad 4 countries', *Journal of Competitiveness*, Vol. 10, No. 1, pp.54–72.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) *An Introduction to Statistical Learning: With Applications in R*, Springer Texts in Statistics, Springer, New York.

Kiselakova, D., Sofrankova, B., Onuferová, E. and Čabinová, V. (2019) 'The evaluation of competitive position of EU-28 economies with using global multi-criteria indices', *Equilibrium*, Vol. 14, No. 3, pp.441–462.

Kolassa, S. (2017) 'Statistical Learning with Sparsity. The Lasso and Generalizations, Vol. 143, Trevor Hastie, Robert Tibshirani, Martin Wainwright, in: Monographs on Statistics and Applied Probability CRC Press (2015), 351, ISBN: 978-1-4987-1216-3', *International Journal of Forecasting*, Vol. 33, No. 3, pp.743–744.

LeBel, P. (2008) 'The role of creative innovation in economic growth: Some international comparisons', *Journal of Asian Economics*, Vol. 19, No. 4, pp.334–347.

Marčeta, M. and Bojnec, S. (2020) 'Drivers of Global Competitiveness in the European Union Countries in 2014 and 2017', *Organizacija*, Vol. 53, No. 1, pp.37–52.

Molnar, C. (2019) *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*, 1st ed., 24 March, Lulu.

Murphy, K.P. (2012) *Machine Learning, a Probabilistic Perspective*, Massachusetts Institute of Technology, The MIT Press Cambridge, Massachusetts, London, UK.

Nallari, R. and Griffith, B. (2013) *Clusters of Competitiveness*, The World Bank, Washington DC.

Neffati, M. (2015) 'The implications of ICT-development and innovation on competitiveness: case of Euro-Mediterranean countries', *International Journal of Development Research*, Vol. 5, No. 2, pp.3482–3488.

Olczyk, M. (2016) 'International competitiveness in the economics literature: a bibliometric study', *Athens Journal of Business & Economics*, Vol. 2, No. 4, pp.375–388.

Palczewska, A., Palczewski, J., Robinson, R.M. and Neagu, D. (2014) 'Interpreting random forest classification models using a feature contribution method', *Advances in Intelligent Systems and Computing*, Vol. 263, pp.193–218, Springer, Cham.

Palei, T. (2015) 'Assessing the impact of infrastructure on economic growth and global competitiveness', *2nd Global Conference on Business, Economics, Management and Tourism: Procedia Economics and Finance*, Vol. 23, pp.168–175.

Petrarca, F. and Terzi, S. (2018) 'The Global Competitiveness Index: an alternative measure with endogenously derived weights', *Quality & Quantity*, Vol. 52, No. 5, pp.2197–2219.

Pérez-Moreno, S., Rodríguez, B. and Luque, M. (2015) 'Assessing global competitiveness under multi-criteria perspective', *Economic Modelling*, Vol. 53, No. 2016, pp.398–408.

Razavi, S., Abdollahi, B., Ghasemi, R. and Shafie, H. (2012) 'Relationship between 'innovation' and 'business sophistication': a secondary analysis of countries global competitiveness', *European Journal of Scientific Research*, Vol. 79, No. 1, pp.29–39.

Roszko-Wójtowicz, E. and Grzelak, M.M. (2020) 'Macroeconomic stability and the level of competitiveness in EU member states: a comparative dynamic approach', *Oeconomia Copernicana*, Vol. 11, No. 4, pp.657–688.

Saarela, M. and Jauhiainen, S. (2021) 'Comparison of feature importance measures as explanations for classification models', *SN Applied Sciences*, Vol. 3, No. 2, p.272.

Schwab, K. (2009) *The Global Competitiveness Report 2009–2010*, Technical report, World Economic Forum [online] http://www3.weforum.org/docs/WEF_GlobalCompetitivenessReport_2009-10.pdf (accessed 8 January 2021).

Schwab, K. (2010) *The Global Competitiveness Report 2010–2011*, Technical report, World Economic Forum [online] http://www3.weforum.org/docs/WEF_GlobalCompetitivenessReport_2010-11.pdf (accessed 7 January 2021).

Schwab, K. (2011) *The Global Competitiveness Report 2011–2012*, Technical report, World Economic Forum [online] http://www3.weforum.org/docs/WEF_GCR_Report_2011-12.pdf (accessed 10 January 2021).

Schwab, K. (2012) *The Global Competitiveness Report 2012–2013*, Technical report, World Economic Forum [online] http://www3.weforum.org/docs/WEF_GlobalCompetitivenessReport_2012-13.pdf (accessed 13 January 2021).

Schwab, K. (2013) *The Global Competitiveness Report 2013–2014*, Technical report, World Economic Forum [online] http://www3.weforum.org/docs/WEF_GlobalCompetitivenessReport_2013-14.pdf (accessed 19 January 2021).

Schwab, K. (2014) *The Global Competitiveness Report 2014–2015*, Technical report, World Economic Forum [online] http://www3.weforum.org/docs/WEF_GlobalCompetitivenessReport_2014-15.pdf (accessed 19 January 2021).

Schwab, K. (2015) *The Global Competitiveness Report 2015–2016*, Technical report, World Economic Forum [online] http://www3.weforum.org/docs/gcr/2015-2016/Global_Competitiveness_Report_2015-2016.pdf (accessed 20 January 2021).

Schwab, K. (2016) *The Global Competitiveness Report 2016–2017*, Technical report, World Economic Forum [online] http://www3.weforum.org/docs/GCR2016-2017/05FullReport/TheGlobalCompetitivenessReport2016-2017_FINAL.pdf (accessed 25 January 2021).

Schwab, K. (2017) *The Global Competitiveness Report 2017–2018*, Technical report, World Economic Forum [online] http://www3.weforum.org/docs/GCR2017-2018/05FullReport/TheGlobalCompetitivenessReport2017-2018.pdf (accessed 26 January 2021).

Schwab, K. (2018) *The Global Competitiveness Report 2018*, Technical report, World Economic Forum [online] http://www3.weforum.org/docs/GCR2018/05FullReport/TheGlobalCompetitivenessReport2018.pdf (accessed 22 January 2021).

Schwab, K. and Porter, M. (2008) *The Global Competitiveness Report 2008–2009*, Technical report, World Economic Forum [online] http://www3.weforum.org/docs/WEF_GlobalCompetitivenessReport_2008-09.pdf (accessed 27 January 2021).

Şener, S. and Delican, D. (2019) 'The causal relationship between innovation, competitiveness and foreign trade in developed and developing countries', *Procedia Computer Science*, Vol. 158, pp.533–540.

Simionescu, M., Pelinescu, E., Khouri, S. and Bilan, S. (2021) 'The main drivers of competitiveness in the EU-28 countries', *Journal of Competitiveness*, Vol. 13, No. 1, pp.129–145.

Širá, E., Vavrek, R., Kravčáková Vozárová, I. and Kotulič, R. (2020) 'Knowledge economy indicators and their impact on the sustainable competitiveness of the EU countries', *Sustainability*, Vol. 12, No. 10, p.4172.

Sofrankova, B., Kiselakova, D. and Čabinová, V. (2017) 'Innovation as a source of country's global competitiveness growth', *SHS Web of Conferences*, Vol. 39, 12pp, 01026.