International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Examining the Most Important Determinants of Health-Related Quality of Life (HRQoL). The Machine Learning Approach

Nino Kokashvili[a*], Youjun Shin[b]

[a,b]*University of Tartu, Liivi 4, Tartu, 51009, Estonia*

[a]*Email: nino.kokashvili@ut.ee*

[b]*Email: youjun.shin@ut.ee*

Abstract

There are various circumstances affecting the individual health-related quality of life (HRQoL). The aim of the paper is to understand which health determinants are the most crucial while designing the efficient health policy. Using the machine learning approach, authors examine 42 health status related factors. The paper incorporates 27 individual level and 15 regional level health state determinants in empirical investigation. Results show that in terms of factor weights, the subjective health is the most influential on individual level and medical labor force - on regional level. However, in terms of frequency, the hospital visiting plays the most important role on individual level and estate condition - on regional level. In addition, empirical results indicate that individual level factors have higher impact on health status than regional level factors. Based on empirical results of the paper, authors provide policy recommendations.

Keywords: Health Determinants; Health Policy; Machine Learning; Health-related quality of life (HRQoL); JEL Classification: I12; I18; I15.

---

* Corresponding author.

1. Introduction

The role of health is a central concern in society [31]. At the same time, the production of health is a question of the public policy [35, 8]. The biggest challenge in the process of improving the individual health state lies on the complexity to directly control all health state influential factors [22]. Understanding which decisions, both individual and public, result in change of health level is crucial [26]. Therefore, determinants of health and contribution of the health care take the leading area of research interests for health economists. Understanding which health determinates are the most crucial is an extremely important task to design the most efficient health policy [47]. The research on the production of health function requires estimation of the relationship between health inputs and outputs [44]. Health indicators, which are simultaneously meaningful and measurable are difficult to find. Researchers in the field have mainly focused on health-related quality of life (HRQoL) as a relevant health status indicator in last decades, while choosing between mortality rates and HRQoL [15]. Even though mortality rates are strictly accurate measures, still HRQoL is assumed to better capture the aspects of health status that are essential for economics and policy oriented research [11, 1, 3]. Health-related quality of life (HRQoL) is effected by various factors [13]. There exists a growing academic literature studying the determinants of health state on individual [4, 23] and regional [19, 43] level. However, the health state accounts the multidimensional circumstances and the empirical evidence on the matter is diverse. The main challenge in measuring the effects on HRQOL is related to both: the multidimensional nature of determinants [22] and the difficulty of mathematical techniques [2]. Even though, the large body of empirical literature has been devoted to resolve the problem by means of econometric models [21]. Still, researchers have faced difficulty to eliminate biases in the estimates [27, 36] . The investigation of the modern health production function requires advanced applications. The article contributes to the literature which tries to identify factors crucial for health status. The novelty of the study lies in its effort to use the machine learning methodology to estimate factors impacting on HRQoL. Machine learning approach provides a new perspective and methodology compared to traditional regression estimations. This article examines 42 health status related individual and regional factors and investigates their role in perceiving the health-related quality of life (HRQOL). The data used for empirical investigation is taken from national health and nutrition examination survey 2016 (NHANES) and statistics Korea. Results show that individual factors have higher impact on health status than regional factors. In terms of factor weights, the subjective health is the most influential on individual level and medical labor force - on regional level. However, in terms of frequency, the hospital visiting plays the most important role on individual level and regional development level - on regional level. Considering the empirical results of the paper, authors recommend policy makers to pay higher attention on individual aspects of the health policy. However, the regional indicators, such as medical facilities (capital and labor medical supply) are extremely important determinants of health-related quality of life (HRQOL). Moreover, the regional development indicators, such as, population, crime level, number of medical cases, have significant effect on health-related quality of life (HRQOL). Therefore, policy makers are recommended to pay high attention to medical facility supply, especially in low developed regions.The article is organized as follow. Section 2 presents the theoretical framework of the article. Section 3 reviews empirical literature about health status determinants. Section 4 presents methodology used for the analyses. Section 5 introduces the data and explains the choice of variables. The 6th section derives and discusses results. The last section concludes and presents the policy implications.

2. Theoretical framework

Health status is influenced by different factors [31, 12, 6]. The previous literature in the field has identified three main categories of health state determinants [13]:

- Social/Economic environment
- Physical environment
- Individual characteristics and behaviors

The role of social/economics environment, such as, economic development of the country or region of an individual or family income and social status, is enormous while discussing the individual health state [10]. Socio-economic environment provides resources and supporting systems which are extremely important for a sufficient health care. Higher income and social status is usually associated with better health [18]. However, the low education is linked to relatively poor health state [45]. Besides socio-economic atmosphere, also the physical environment, such as clean water and air, safety and healthy workspace, effect health state dramatically [34]. Healthier and well-maintained physical environment is related to improved health status [33] While physical and socio-economic environment indirectly influence our health level, the individual characteristics and behaviors, such as age, gender, smoking habits and obesity issues, are factors which directly effect the individual's health condition [46]. Among various theoretical frameworks about health determinants, the Dahlgren-Whitehead Rainbow model is assumed to provide the most consistent framework and has been widely used in health economics literature [32, 39]. Based on the Dahlgren-Whitehead Rainbow model authors construct own theoretical framework, which distinguishes between different layers of influential circumstances of health status and maps the relationship between them. The model identifies following four groups of factors impacting on health-related quality of life (HRQOL): Social and community networks, individual lifestyle, hereditary and general socio-economic, cultural and environmental factors (see *Figure 1: Theoretical framework: Determinants of health status*). The broader classification applied in the model is individual vs regional (country-level) determinants, in which social or community networks, individual lifestyle and hereditary factors are grouped together. The rest of the aspects, such as the general socio-economic, cultural and environmental factors, represent the regional (country level) indicators. Above mentioned classification of factors provide possibility for the article to explore not only the relative influence of the determinants, but also compare the regional and individual factors. Understanding the difference between them is an extremely important aspect in providing a relevant health policy. Therefore, the main scope of the article is investigating the health determinants with policy implications in mind.

Source: Authors own work based on the rainbow model of Dahlgren-Whitehead.

Social & Community Networks	**Individual Lifestyle Factors**	**Hereditary Factors**	**General Socio-Economic, cultural & Environmental Factors**
Education Occupation Income Level Public Aid House ownership Marriage Health Insurance Private Health Insurance Economic Activity Father's Education Mother's Education Subjective Health State Availability of Food Monthly family Income Impatient Experience Outpatient Experience	Walking Per Week Muscle Per Week Aerobic Exercise Obesity Working Hour Smoking Lifetime Daily smoking Drink Per Year	AGE SEX	Outpatient Regional Impatient Regional Number of Hospitals Number of Medical Employees Crime Number of Surgeries Hospital Staying Number of Medical Cases Population Number of Hospital Beds GRDP Number of Houses per 1000 Number of volunteer cases Unemployment Rate

Figure 1: Theoretical framework: Determinants of health status

3. Methodology

The paper uses the machine learning methodology to estimate factors impacting on Health Related Quality of Life (HRQoL). The machine learning is a branch of artificial intelligence, which builds algorithms to make decisions by learning from the surrounding circumstances. In its design, a computer studies and behaves like humans, drawing the solutions from given environment [5]. The main idea of machine learning approach is to provide generic algorithms to a computer instead of custom code so that it can find necessary information from the dataset [16]. In other words, if a computer learns relevant algorithms, which are suitable for a particular scenario, it generates logic to explain the situation[a].

In this article, authors analyze factors affecting quality of health status. Health status (HRQoL) is the response

a The machine learning is a method consisted of 3 factors: experience, task and, performance. Given the experience collected by the data $X=(x_1, x_2, x_3, x_4, \dots, x_n)$ $Y=(y_1, y_2, y_3, y_4, \dots y_n)$*, a task of machine learning is to find out a function which could explain the relationship between X and Y vectors, and improve the accuracy of the function (performance) by learning.*

variable. Determinant factors represent both: individual and regional variables as predictor variables. This empirical setting is a suitable environment for supervised learning[b]. In addition to the algorithms used for supervised learning, to analyze the effects on the quality of health status in Korea, authors also apply following 3 models:

- The random forest model[c] (see *3.1 The random forest model for details)*
- The gradient boosting regression (GBR) tree[d] (see *3.2 The gradient boosting regression (GBR) tree for details)*
- The artificial neural networks (Deep learning)[e] (see *3.3 Deep learning for details)*

The main methodological advancement of the paper lies in the effort to perform ensemble learning[f], which combines together the three models mentioned above (Deep learning, Random forest and GBM). Ensemble learning provides a possibility to achieve a higher predictive power compared to any of the single machine learning models [14]. It is an aggregation of strong learners such as bagging and boosting models and it merges various strong models through a meta-learning algorithm and makes it even the stronger model. Stacking method could lead to an asymptotically optimal performance for learning. The biggest challenge of using machine learning is the interpretation [29]. Relatively simple models such as GLM and decision tree are easily interpreted by coefficients and nodes. However, regarding relatively complicated models used in this article, it is hard to interpret and observe the process, which is called "black box". Many researchers have tried to open the "black box" for the better usage of machine learning model [9]. In this article, authors use local interpretable model-agnostic explanations (LIME)[g] to interpret the model. The steps in the method are as follows [37, 38]:

- Choosing observations for explanation, out of the black box
- Perturbing the observations and employing the predictions based on the new dataset obtained from the perturbation
- Putting weights data based on the proximity to the chosen observation
- Regressing a weighted value on the perturbed dataset

2. Machine learning is divided into two parts depending on the learning method: supervised learning and unsupervised learning. Supervised learning is an algorithm, which predicts inexperienced or not-known response values with response and predictor variables and it is applied mainly for classification or regression analysis. Support Vector Machine (SVM), Decision tree, K-Nearest Neighbor (KNN), Artificial neural networks (ANN), ridge/lasso regression fall into this category. Unsupervised learning techniques aim to calculate and to find out a pattern and relationship of a given data. We can list in these category algorithms such as Principal Component Analysis (PCA), k-means, Non-negative Matrix Factorization (NMF), and so forth. The criteria to choose between supervised and unsupervised learning is up to given data.

[c] Random forest is a tree-based model suggested by Breiman (2001). It aggregates results from multiple decision trees through bootstrap and it complements the weak points of the basic decision tree model.

[d] GBM is one of machine learning techniques for regression and classification problem and typically it is also considered one tree-based model.

[e] Deep learning is one of modern machine learning model taking into account Neural Network theory. It is a mathematical model which copies how the structures of human neural networks work (Ripley, 1996: Titterington, 2010). Especially it takes into account the behavior of how neurons transfer and process signals. Deep learning is used widely in many different fields such as convolution deep neural networks, deep belief networks for computer visions, sound recognition, natural language process, signal pattern analysis and so forth.

[f] Ensemble learning is one of machine learning techniques that uses multiple machine learning models together.

[g] The method is from the paper: RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. Why should i trust you?: Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* ACM, 2016. p. 1135-1144.

- Presenting the prediction and interpreting the model out of the black box

3.1 The random forest model

The advantage of Random forest model over the decision tree model is that it applies randomness to the model [25]. The randomness of Random forest model works through bootstrap: the model selects N sample data and chooses m explanatory variables among total variables. Aggregated trees with randomness, resulting from Random forest, improve the predictive power, and also they are free from over-fitting problem according to Law of Large Numbers [42]. However, even though there are merits of Random forest over the decision tree model, as a demerit, it is not possible to observe how the process is developed. The process of random forest model is summarized as follows [25]:

- Select N sample data through bootstrap with replacement.
- Make decision trees with m explanatory variables among total P.
- Iterate 1 and 2 processes and make T number of decision trees.
- Aggregate results from T decision trees and calculate the mean of results (continuous variable) or select the majority (discrete variable) from T decision trees.
- Evaluate the model.

3.2 The gradient boosting regression (GBR) tree

The GBM (Gradient Boosting Machine)[h] model used in the article is suggested by the paper [17]. Unlike Random forest, which improves the performance of the model by aggregating a number of sing decision trees, GBM enhances the predictive power through taking into account result from previous decision trees and sequentially update the next decision tree [40]. Simply, GBM is to train a weak learner until it becomes a strong learner. GBM starts from loss function as other machine learning models do and the goal is to minimize the loss. The process of GBM is the following [17].

- Set Approximation function $F_0(x) = \text{argmin}_\alpha \sum_{i=1}^{n} L(y_i, \alpha)$
- Calculate pseudo – responses $r_t(x_i) = y_i - F_{t-1}(x_i)$
- Fit the decision tree h(x) using data from training set $\{(x_i, r_m(x_i))\}_{n=1}^{N}$
- Calculate the optimal weight value α_t with given approximation function $\alpha_t = \text{argmin}_\alpha \sum_{i=1}^{n} L(y_i, F_{t-1}(x) + \alpha h(x))$
- Update the function $F_t(x) = F_{t-1}(x) + \alpha_t h(x)$
- Finish the iteration when $F_t(x) = F_t(x)^*$, find the vest approximation function

3.3 Deep learning

Deep learning model consists of an input layer, hidden layers, and an output layer [24]. The whole process simply consists of the summation of weighted values, transformation via certain function and connection to the

next layer until the end of the calculation as follows [48]:

- From an input layer with given explanatory variables, the process begins with putting weights and summing them as follows:

$$f(x_1, x_2, x_3, \cdots, x_d) = \sum_{i=0}^{d} w_i x_i$$

- Transform those values via the activation function, given by:

$$\sigma(w_0 + w_1 x_1 + w_1 x_1 + \cdots + w_d x_d)$$

- Each layer goes from input layer to a number of hidden layers. Several activation functions exist and they are selected by researchers depending on the predictive power and design of the model. Activation functions could be:

$$\text{sigmoid function: } f(t) = \frac{1}{(1 + e^{-t})}$$

$$\text{hyperbolic tangent function: } f(t) = \frac{(e^{t} - e^{-t})}{(e^{t} + e^{-t})}$$

$$\text{absolute function: } f(t) = ||t||$$

$$\text{rectified linear unit: } f(t) = \max(0, t)$$

- Once the output layer values are obtained, the next process is to iterate the mentioned steps above again to find the optimal weight values.

1. The Data and descriptive statistics

In this article authors concentrate on the case of Korea and focus on the year of 2016. The data used for the empirical investigation is taken from the national health and nutrition examination survey 2016 (NHANES) and statistics Korea. Authors choose 42 health status-related factors including individual and regional levels and investigate which ones are most crucial factors for public health, HRQoL. The national health and nutrition examination survey 2016 (NHANES) is a national level survey in Korea to collect information related to citizens' health level, health related behavior and the socio-economic state. The survey is performed annually, however, in this article authors concentrate on the year of 2016. The regional development indicators are taken from Statistics Korea for the corresponding year, 2016. The main variable (the dependent variable) of our interest is Health-related quality of life (HRQoL). Authors use EQ-5D[i] as the measure of HRQoL. The dataset

[i] EQ-5D is one of tools to measure the HRQoL and it is introduced by the EuroQol Group. It consists of five multiple categories of the self-measurements such as mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. EQ-5D values are known that they are useful for investigating various topics such as health conditions, the efficiency of medical treatment, economic evaluation of healthcare

provides information of HRQoL for 7,879 observations from which 3625 are male and 4460 female respondents. The average value of HRQOL is 0.945. Average HRQoL for males is 0.960 and 0940 for females. The sample covers 7 age groups: <20; 20-29; 30-39; 40-49; 50-59; 60-69 and >70 (see Figure 2 in annex 1). The lowest average score of HRQoL is found for the >70 age group (see Figure 3 in annex 1). The dataset consists of 4 income level groups: High, Low, Middle, High-Low (see Figure 4 in annex 1). Observations are evenly distributed by income level groups (see Figure 5 in annex 1). Descriptive statistics indicate that the value of HRQoL (average scores) increases as the income level goes up (see Figure 4 in annex 1). The same trend is evident for the education level indicators. The sample consists of 4 education groups (the last education level): primary, middle school, high school and university (see Figure 6 in annex 1). The value of HRQoL (average scores) increases as the education level goes up (see Figure 6 in annex 1). The occupation of respondents is divided into 7 categories: specialist, officer, service, agriculture, mechanic, simple labor, student/unemployed. The highest HRQoL scores (average) are evident for following categories: specialist and officer followed by mechanic and service.

The relatively low HRQoL scores are evident for agriculture, simple-labor and unemployed categories (see Figure 7 in annex 1). The dataset covers 42 health-status related factors, from which 27 are individual level and 15 regional level factors. 21 individual level factors represent categorical variables (see table 1 in annex 2) and 6 – continues (see table 2 in annex 2). The regional level variables show information about the economic development and the existing health-related activities in the region. 1 regional level factors represent categorical variables and 14 – continues (see table 3 in annex 2).

The dataset covers 16 regions of Korea (see Table 4 in annex 2). Sample is representative as observations are almost equally distributed among regions (see figure 8 in annex 2). The highest average HRQoL values are found in Gangwon, Seoul and Ulsan regions. The lowest HRQoL score is evident in Geyonggi region (see figure 8 in annex 2).

2. Empirical Results

This section presents empirical results based on machine learning exercise discussed in previous chapter (see *section 3. Methodology*). The aim of the article is to understand the relative importance of health determinant factors. In this paper, authors have initially performed three different machine learning algorithms: the random forest, the gradient boosting model (GBM), and the deep learning. Then, we have aggregated the best performing parts of each application based a stacked ensemble model. According to the result of the calculation, based on the mean squared error (MSE), a stacked ensemble model shows the highest accuracy rate (0.003), followed by random forest model (0.005), deep learning model (0.006), and gradient boosting model (0.008) (see below the table 5: *Estimation of model accuracy*, for additional information about hyper-parameters, please see Table 5 in annex 3)

policy, and so on (EuroQol 1990). EuroQol, G. (1990). EuroQol--a new facility for the measurement of health-related quality of life. Health policy (Amsterdam, Netherlands), 16(3), 199

Table 5: Estimation of model accuracy

Type	MAE	MSE	RMSE	RMSLE
Gradient boosting model	0.052	0.009	0.094	0.055
Random forest	0.043	0.006	0.076	0.044
Deep learning	0.047	0.007	0.082	0.048
Stacked model	0.036	0.003	0.059	0.034

Source: authors own measurements. Based on the results of the accuracy of the models, the empirical investigation proceeds with the stacked ensemble model. Authors use the local interpretable model-agnostic explanations (LIME) for analyses (see section 3. *Methodology*). Results are described based on two approaches: the factor frequency and the factor weights. In other words, regarding the interpretation of results through machine learning and LIME, we present results with two perspectives: how frequently variables are used for the prediction of HRQoL and how large are variables weighted.The frequency of factors shows how often are the variables used in the calculation while predicting the health-related quality of life (HRQOL). The weight of variable shows the value of contribution (weight) each factor adds to the indicator of the health-related quality of life (HRQOL). Figure 9 (above) presents results for the total frequency of variables, which indicates that there is a dramatic difference in frequency between individual and regional factors. Individual variables are more important to be taken into account compared to regional factors. The figure 10 (below) presents results for the total frequency of variables by levels (individual and regional). Most important individual factors based on frequency are hospital visiting experience, existence of private insurance, subjective health state and age. Relatively less important individual factors based on frequency are house ownership and working type as well as various behavior factors, such as, smoking or drinking habits. In terms of regional variables, the most important health determinant factors based on frequency are regional development level, such as, supply of houses, regional GDP level (GRDP) and employment level followed by healthcare facilities. Relatively less important regional health determinants based on frequency are crime level and hospital visiting experience (for more detailed information on frequency of variables, see annex 3, figures 11 and 12).The figure 13 (below) presents result for the most important variables by weight. In case of factor weights, similarly to frequency pattern, individual variables on average have higher importance than regional factors while predicting the health-related quality of life (HRQOL). Figure 14 (below) presents results for the total weight of variables by levels (individual and regional). Most important individual factors based on factor weight are subjective health state, lifestyle indicators (occupation, datary, exercise behaviors) and parents (mother and father) education levels. Relatively less important individual factors based on factor weights are private or public insurance, income level and hospital visiting experience. Most important regional factors based on factor weight are medical facility levels (capital and labor medical supply) and regional development (population, crime level, number of medical cases). Relatively less important regional factors based on factor weights are volunteer cases and number of houses per 1000.

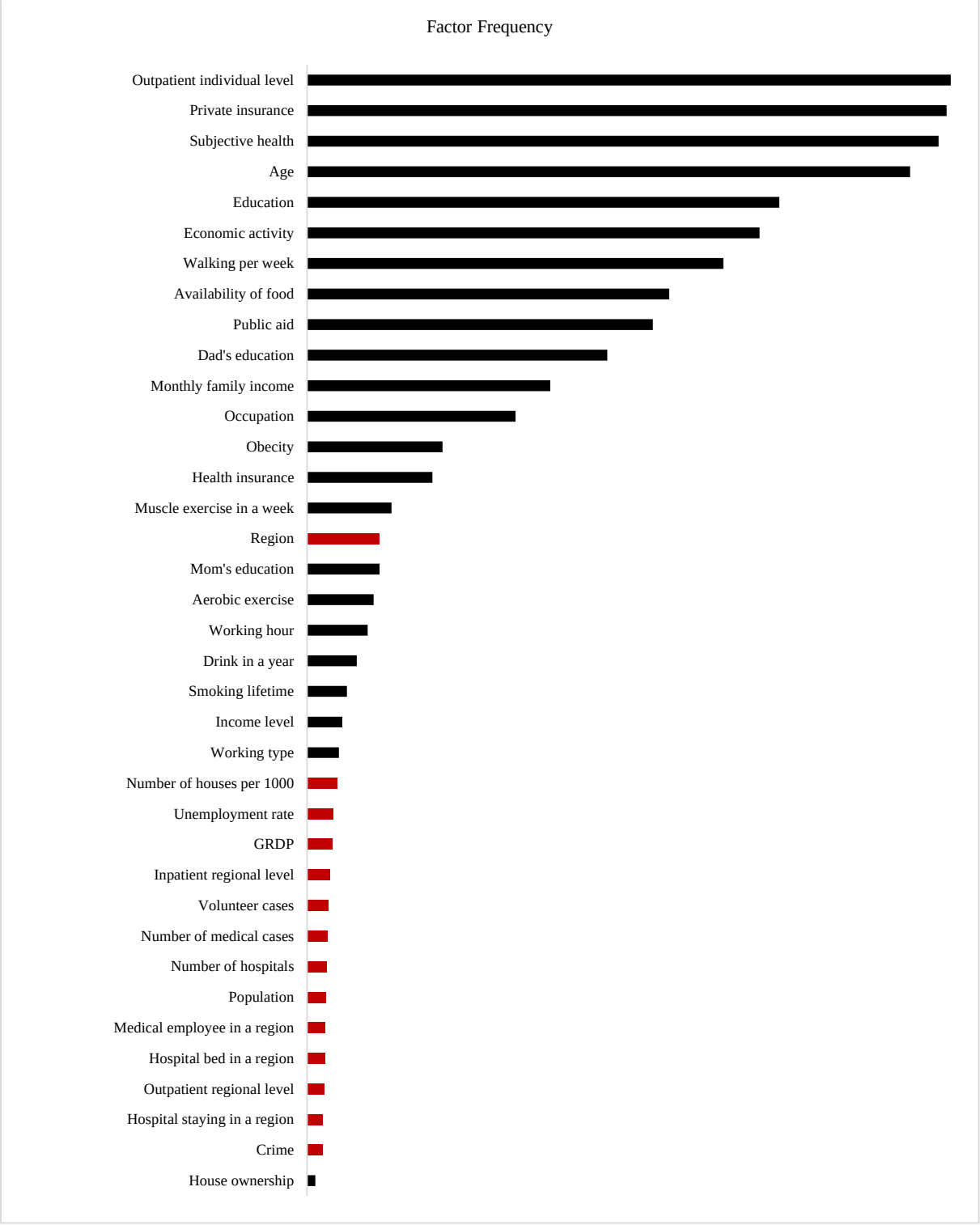

Figure 9: Total frequency of variables

Source: compiled by authors. Red: regional level factors, Black: individual level factors

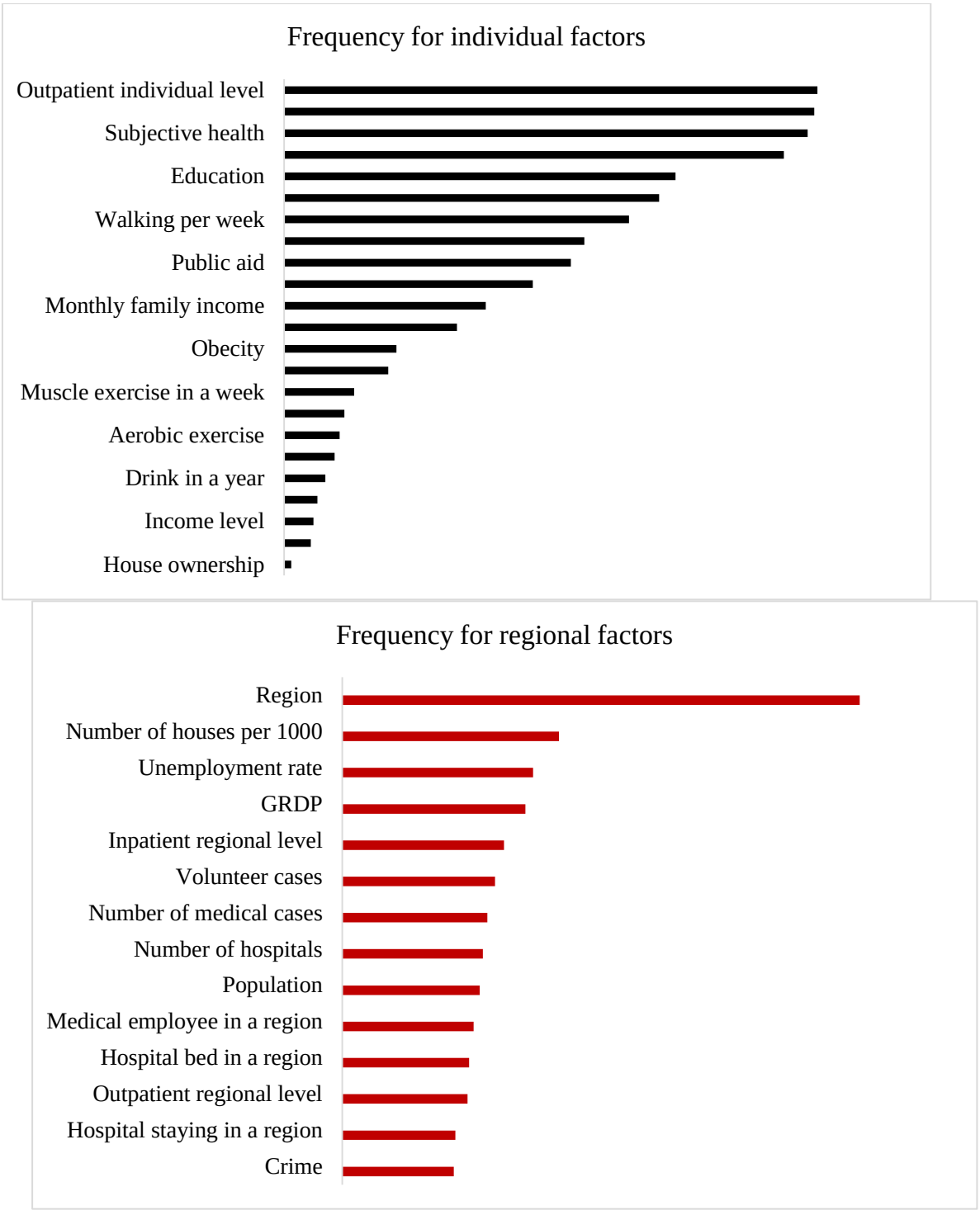

Figure 10: Total frequency of variables by levels (individual and region)

Source: compiled by authors

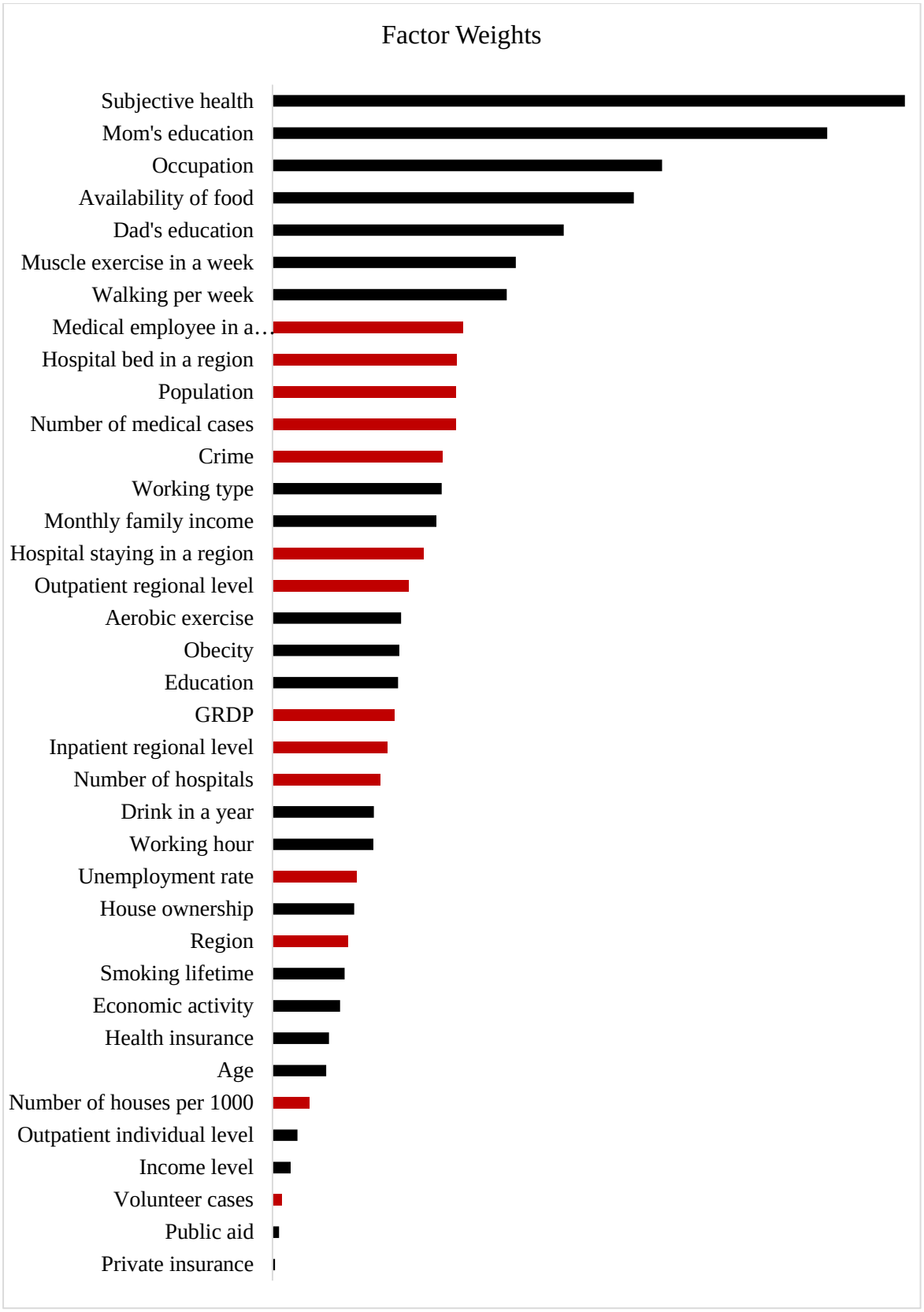

Figure 13: Most important variables by weight

Source: compiled by authors. Red: regional level factors, Black: individual level factors

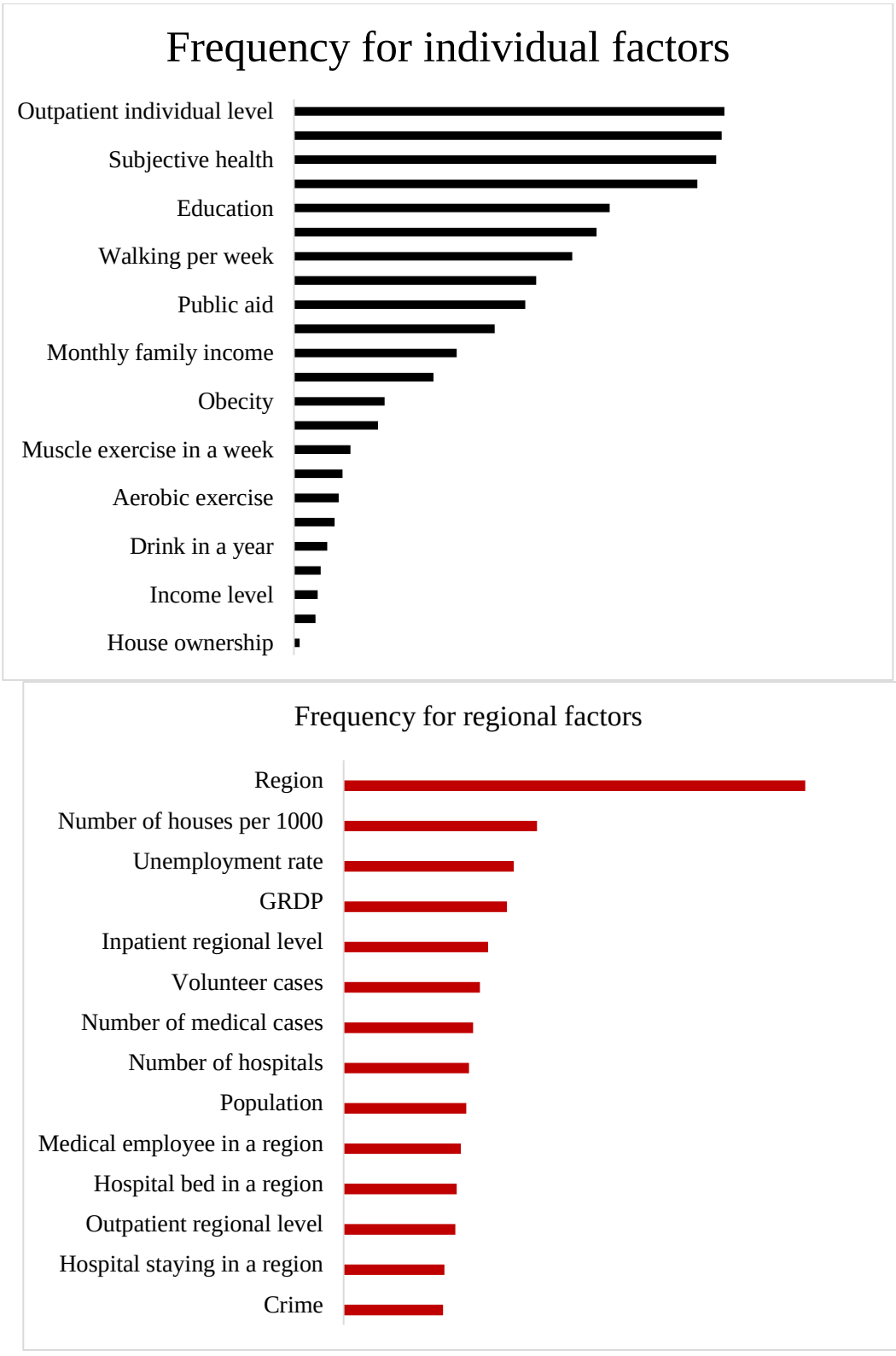

Figure 14: Most important variables by level

Source: compiled by authors

4. Conclusions

The aim of the paper is to evaluate the importance of various health related factors. Authors have used the novel approach, the machine learning technique, for empirical analyses. The approach provides the most accurate results compared to previously used methodologies in the literature. Empirical findings of the paper show that individual factors have higher impact on health status compared to regional factors. Authors employ to ways of measurement: 1. according to factor weights and 2. according to factor frequency. In terms of factor weights, result indicate that subjective health is the most important determinant from the individual level factors. On the other hand, the medical labor force plays the most significant role in health status from the regional level factors. In terms of factor frequency, results show that the hospital visiting has the highest impact on the health state from individual level determinants, the estate condition – from regional level factors. Based on the empirical results (both factor weights and frequency) from machine learning exercise, authors provide policy recommendations for policy makers. Higher attention is recommended to be devoted on individual aspects of health policy. However, certain regional indicators, such as medical facilities (capital and labor medical supply), should be maintained efficiently as those factors are significant determinates of health state. In addition, the regional economic and development factors, such as population, crime level, number of medical cases, are recommended to be elaborated while preparing and implementing the health policy in a specific region. The main limitation of the study lies on the methodology and data used for the empirical investigation. Authors have only focused on the data from Korea for the single year of 2016. Authors recommend for farther research in the field to extend the work by applying dynamic level the cross-country study, which will provide additional insight into the problem.

References

[1]. Andresen, E.M., Vahle, V.J. and Lollar, D., 2001. Proxy reliability: health-related quality of life (HRQoL) measures for people with disability. Quality of Life Research, 10(7), pp.609-619.

[2]. Arden-Close, E., Pacey, A. and Eiser, C., 2010. Health-related quality of life in survivors of lymphoma: a systematic review and methodological critique. Leukemia & lymphoma, 51(4), pp.628-640.

[3]. Bansback, N., Czoski-Murray, C., Carlton, J., Lewis, G., Hughes, L., Espallargues, M., Brand, C. and Brazier, J., 2007. Determinants of health related quality of life and health state utility in patients with age related macular degeneration: the association of contrast sensitivity and visual acuity. Quality of Life Research, 16(3), p.533.

[4]. Bier, J.A.B., Prince, A., Tremont, M. and Msall, M., 2005. Medical, functional, and social determinants of health-related quality of life in individuals with myelomeningocele. Developmental Medicine and Child Neurology, 47(9), pp.609-612.

[5]. Bies, R.R., Muldoon, M.F., Pollock, B.G., Manuck, S., Smith, G. and Sale, M.E., 2006. A genetic algorithm-based, hybrid machine learning approach to model selection. Journal of Pharmacokinetics and Pharmacodynamics, 33(2), pp.195-221.

[6]. Braveman, P., Egerter, S. and Williams, D.R., 2011. The social determinants of health: coming of age. Annual review of public health, 32, pp.381-398.

[7]. Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32.

[8]. Brownson, R.C., Chriqui, J.F. and Stamatakis, K.A., 2009. Understanding evidence-based public health policy. American journal of public health, 99(9), pp.1576-1583.

[9]. Burrell, J., 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), p.2053951715622512.

[10]. Celik, Y. and Hotchkiss, D.R., 2000. The socio-economic determinants of maternal health care utilization in Turkey. Social science & medicine, 50(12), pp.1797-1806.

[11]. Cieza, A. and Stucki, G., 2005. Content comparison of health-related quality of life (HRQOL) instruments based on the international classification of functioning, disability and health (ICF). Quality of Life Research, 14(5), pp.1225-1237.

[12]. Currie, C., Zanotti, C., Morgan, A., Currie, D., De Looze, M., Roberts, C., Samdal, O., Smith, O.R. and Barnekow, V., 2009. Social determinants of health and well-being among young people. Health Behaviour in School-aged Children (HBSC) study: international report from the, 2010, p.271.

[13]. Degroote, S., Vogelaers, D.P., Vermeir, P., Mariman, A., De Rick, A., Van Der Gucht, B., Pelgrom, J., Van Wanzeele, F., Verhofstede, C. and Vandijck, D.M., 2013. Socio-economic, behavioural,(neuro) psychological and clinical determinants of HRQoL in people living with HIV in Belgium: a pilot study. Journal of the International AIDS Society, 16(1), p.18643.

[14]. Dietterich, T.G., 2002. Ensemble learning. The handbook of brain theory and neural networks, 2, pp.110-125.

[15]. Efficace, F., Bottomley, A., Osoba, D., Gotay, C., Flechtner, H., D'haese, S. and Zurlo, A., 2003. Beyond the development of health-related quality-of-life (HRQOL) measures: a checklist for evaluating HRQOL outcomes in cancer clinical trials—does HRQOL evaluation in prostate cancer research inform clinical decision making?. Journal of Clinical Oncology, 21(18), pp.3502-3511.

[16]. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M. and Hutter, F., 2015. Efficient and robust automated machine learning. In Advances in neural information processing systems (pp. 2962-2970).

[17]. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232

[18]. Gangadharan, L. and Valenzuela, M.R., 2001. Interrelationships between income, health and the environment: extending the Environmental Kuznets Curve hypothesis. Ecological Economics, 36(3), pp.513-531.

[19]. Hoi, L.V., Chuc, N.T. and Lindholm, L., 2010. Health-related quality of life, and its determinants, among older people in rural Vietnam. BMC public health, 10(1), p.549.

[20]. Huynh-Thu, V.A., Saeys, Y., Wehenkel, L. and Geurts, P., 2012. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. Bioinformatics, 28(13), pp.1766-1774.

[21]. Jones, A.M., 2000. Health econometrics. In Handbook of health economics (Vol. 1, pp. 265-344). Elsevier.

[22]. Kirigia, J.M., Seddoh, A., Gatwiri, D., Muthuri, L.H. and Seddoh, J., 2005. E-health: determinants, opportunities, challenges and the way forward for countries in the WHO African Region. BMC Public

Health, 5(1), p.137.

[23]. Kivits, J., Erpelding, M.L. and Guillemin, F., 2013. Social determinants of health-related quality of life. Revue d'epidemiologie et de sante publique, 61, pp.S189-S194.

[24]. LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. nature, 521(7553), p.436.

[25]. Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), pp.18-22.

[26]. Macinko, J. and Starfield, B., 2001. The utility of social capital in research on health determinants. The Milbank Quarterly, 79(3), pp.387-427.

[27]. Mantzavinis, G.D., Pappas, N., Dimoliatis, I.D. and Ioannidis, J.P., 2005. Multivariate models of self-reported health often neglected essential candidate determinants and methodological issues. Journal of clinical epidemiology, 58(5), pp.436-443.

[28]. Mark J van der Laan, Eric C Polley, and Alan E Hubbard, "Super Learner" Journal of the American Statistical Applications in Genetics and Molecular Biology, VOlume6, Issue1. (September 2007)

[29]. Marmot, M., 2005. Social determinants of health inequalities. The lancet, 365(9464), pp.1099-1104.

[30]. Marmot, M., Allen, J., Bell, R., Bloomer, E. and Goldblatt, P., 2012. WHO European review of social determinants of health and the health divide. The lancet, 380(9846), pp.1011-1029.

[31]. Marmot, M., Allen, J., Goldblatt, P., Boyce, T., McNeish, D., Grady, M. and Geddes, I., 2008. Fair society, healthy lives. The Marmot Review. London2010.

[32]. Modranka, E. and Suchecka, J., 2014. The determinants of population health spatial disparities. Comparative Economic Research, 17(4), pp.173-185.

[33]. Nieuwenhuijsen, M.J., Gomez-Perales, J.E. and Colvile, R.N., 2007. Levels of particulate air pollution, its elemental composition, determinants and health effects in metro systems. Atmospheric environment, 41(37), pp.7995-8006.

[34]. Owen, N., Leslie, E., Salmon, J. and Fotheringham, M.J., 2000. Environmental determinants of physical activity and sedentary behavior. Exerc Sport Sci Rev, 28(4), pp.153-158.

[35]. Patrick, D.L. and Chiang, Y.P., 2000. Measurement of health outcomes in treatment effectiveness evaluations: conceptual and methodological challenges. Medical care, 38(9), pp.II-14.

[36]. Patrick, Donald L., and Pennifer Erickson. "Health status and health policy: quality of life in health care evaluation and resource allocation." (1993).

[37]. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.

[38]. Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.

[39]. Rice, L. and Sara, R., 2018. Updating the determinants of health model in the Information Age. Health promotion international.

[40]. Ridgeway, G., 2007. Generalized Boosted Models: A guide to the gbm package. Update, 1(1), p.2007.

[41]. Ripley, B.D., Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press, (1996).

[42]. Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M. and Rigol-Sanchez, J.P., 2012. An

assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, pp.93-104.

[43]. Romero, M., Vivas-Consuelo, D. and Alvis-Guzman, N., 2013. Is Health Related Quality of Life (HRQoL) a valid indicator for health systems evaluation?. SpringerPlus, 2(1), p.664.

[44]. Rosenzweig, M.R. and Schultz, T.P., 1983. Estimating a household production function: Heterogeneity, the demand for health inputs, and their effects on birth weight. Journal of political economy, 91(5), pp.723-746.

[45]. Ross, C.E. and Wu, C.L., 1995. The links between education and health. American sociological review, pp.719-745.

[46]. Salive, M.E., Cornoni-Huntley, J., Guralnik, J.M., Phillips, C.L., Wallace, R.B., Ostfeld, A.M. and Cohen, H.J., 1992. Anemia and hemoglobin levels in older persons: relationship with age, gender, and health status. Journal of the American Geriatrics Society, 40(5), pp.489-496.

[47]. Starfield, B. and Shi, L., 2002. Policy relevant determinants of health: an international perspective. Health policy, 60(3), pp.201-218.

[48]. Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural networks, 61, pp.85-117.

[49]. Titterington, M., "Neural Networks." Wiley Interdisciplinary Reviews: Computational Statistics, Vol.2, No.1, (2010), pp.1–8.

Annex 1

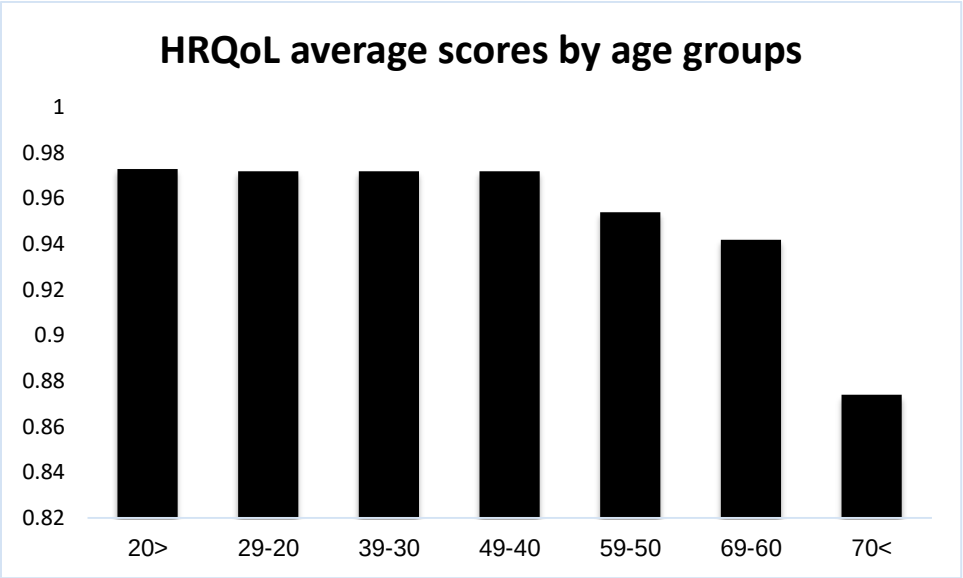

Figure 2: HRQoL average scores by age groups

Source: Authors own calculations

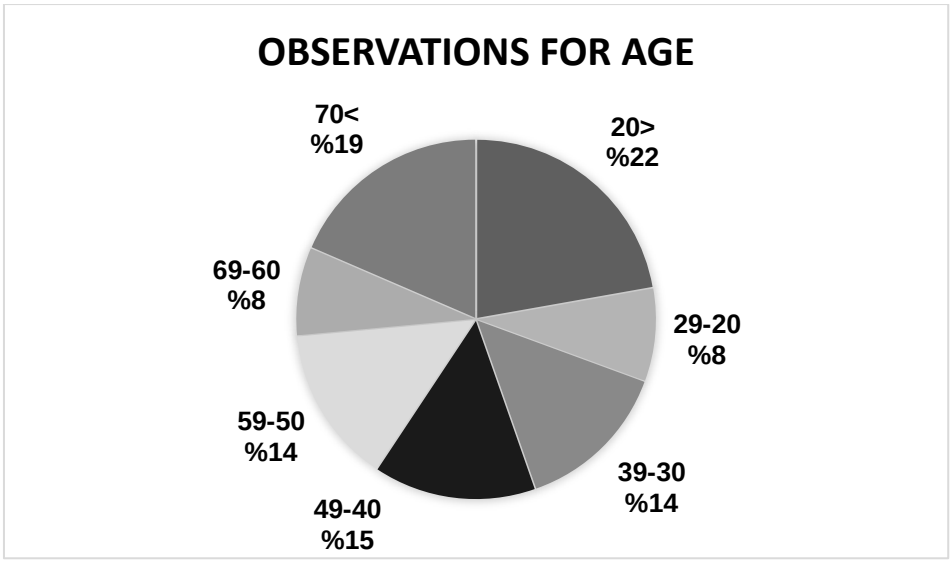

Figure 3: Count of observations for age groups

Source: Authors own calculations

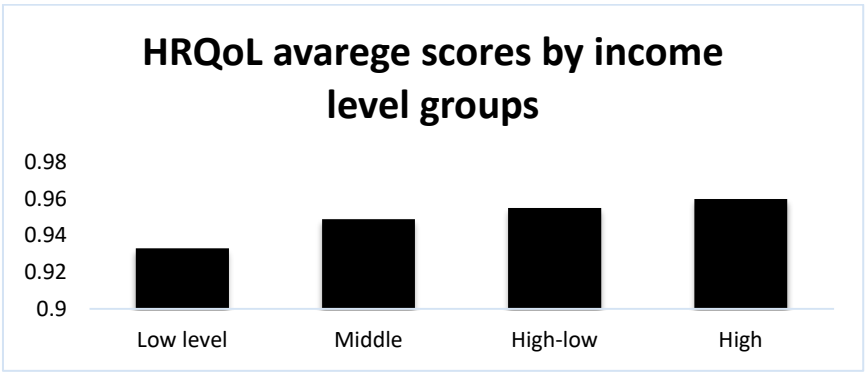

Figure 4: HRQoL average scores by income level groups

Source: Authors own calculations

Figure 5: Count of observations for income level groups

Source: Authors own calculations

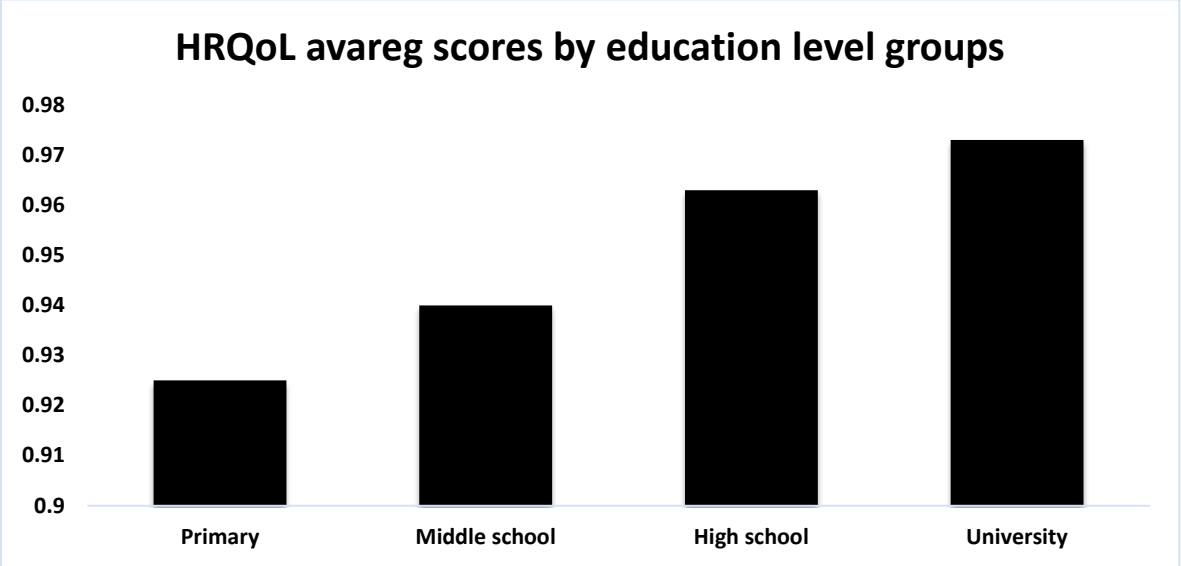

Figure 6: HRQoL average scores by education level groups

Source: Authors own calculations

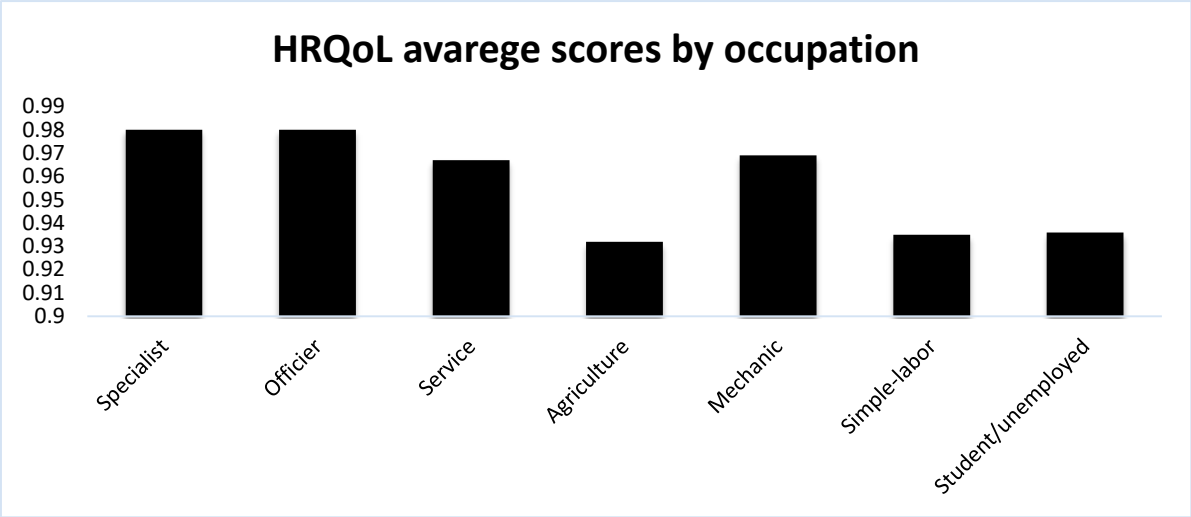

Figure 7: HRQoL average scores by occupation groups

Source: Authors own calculations

Annex 2

Table 1: Descriptive statistics for individual variables (categorical)

Variable	Categories	Freq.	Perc.	Variable	Categories	Freq.	Perc.
Marry	1.Yes	5,215	66.19	Muscle work per week	1.one-day	218	2.77
	2.No	2,664	33.81		2.Two-days	308	3.91
Health insurance	1.Company	5,339	67.76		3.Three-days	323	4.1
	2.Local insurance	2,187	27.76		4.Four-days	141	1.79
	3.Medical aid	278	3.53		5.Five-days	416	5.28
	4.Missing	75	0.95		6.Not-target	1,071	13.59
Private health insurance	1.Yes	6,267	79.54		7.Never	4,778	60.64
	2.No	1,553	19.71		8.Missing	624	7.92
	3.Missing	59	0.75	Aerobic exercise	1.Yes	3,269	41.49
Subjective health state	1.Very good	678	8.61		2.No	2,905	36.87
	2.Good	2,100	26.65		3.Missing	1,705	21.64
	3.Normal	3,304	41.93	Obesity	1.First-level	1,519	19.28
	4.Bad	954	12.11		2.Second-level	3,905	49.56
	5.Very bad	257	3.26		3.Third-level	2,130	27.03
	6.Missing	586	7.44		4.Missing	325	4.12
Economic activity	1.Yes	3,382	42.92	Availability of food	1.Enough food	3,595	45.63
	2.Not need	1,385	17.58		2.A lot but less kind	2,979	37.81
	3.No	2,554	32.42		3.Skip sometimes	185	2.35
	4.Missing	558	7.08		4.Skip many times	30	0.38
Work type	1.Daily-work	3,173	40.27		5.Missing	1,090	13.83
	2.Day-evening shift	118	1.5	Smoke life time	1.Never	3,538	44.9
	3.Day-night shift	35	0.44		2.Less than 5 packs	113	1.43
	4.Divided work	30	0.38		3.More than 5 packs	2,196	27.87
	5.Evening work	406	5.15		4.Not-target	1,679	21.31
	6.Night work	60	0.76		5.Missing	353	4.48
	7.Irregular work	17	0.22	Drink per year	1.Not recently	990	12.57
	8.Not working	3,473	44.08		2.Less than once month	1,114	14.14
	9.Others	5	0.06		3.Once month	591	7.5
	10.Missing	562	7.13		4.Two to four time a month	1,310	16.63
					5.Two to three a week	877	11.13
					6.More than four times a week	412	5.23
					7.Not target	2,232	28.33
					8.Missing	353	4.48

Variable	Categories	Freq.	Perc.	Variable	Categories	Freq.	Perc.
Sex	1:Female	4,351	55.22	Father education	1.Primary school	2,413	30.63
	2:Male	3,528	44.78		2.Middle school	770	9.77
Income level	1:High	1,969	24.99		3.High school	1,041	13.21
	2:Low	1,989	25.24		4.College	131	1.66
	3:Mid-high	1,943	24.66		5.University	479	6.08
	4:Mid-low	1,978	25.1		6.Graduated-university	95	1.21
Education	1.Primary	2,617	33.21		7.Not-target	1,612	20.46
	2.Middle school	825	10.47		8.Missing	1,338	16.98
	3.High school	1,824	23.15	Mother education	1.Primary school	3,137	39.81
	4.University	2,052	26.04		2.Middle school	656	8.33
	5.Missing	561	7.12		3.High school	965	12.25
Occupation	1.Agriculture	226	2.87		4.College	76	0.96
	2.Mechanic	569	7.22		5.University	220	2.79
	3.Officier	593	7.53		6.Graduated-university	26	0.33
	4.Service	754	9.57		7.Not-target	1,612	20.46
	5.Simple-labor	490	6.22		8.Missing	1,187	15.07
	6.Specialist	745	9.46	Walking per week	1.One day	384	4.87
	7.Student/unemployed	4,111	52.18		2.Two-days	617	7.83
	8.Missing	391	4.96		3.Three-days	786	9.98
Public aid	1.Yes	498	6.32		4.Four-days	474	6.02
	2.No	7,374	93.59		5.Five-days	711	9.02
	3.Missing	7	0.09		6.Six-days	361	4.58
House ownership	1.No house	2,507	31.82		7.Every-day	1,716	21.78
	2.One house	4,297	54.54		8.Never	1,124	14.27
	3.More than one house	1,068	13.56		9.Not-target	1,071	13.59
	4.Missing	7	0.09		10.Missing	635	8.06

Table 2: Descriptive statistics for individual variables

Variables	Mean	Std	Min	Max
Age	41.95	22.85	1	80
Monthly family income	424.1	310.6	17	1,500
Inpatient experience	0.14	0.509	0	19
Outpatient experience	0.556	1.272	0	26
Working hour	19.46	23.53	0	112
Daily smoking	1.862	5.433	0	50

Table 3: Descriptive statistic for Regional level variables

Variables	Mean	Std	Min	Max
Outpatient regional	968,171	655,453	93,592	1.71E+06
Inpatient regional	7.90E+06	5.82E+06	794,890	1.44E+07
Number of hospitals	11,095	8,176	1,096	21,786
Number of medical employees	48,870	36,275	4,809	101,142
Crime	222,733	173,297	36,885	466,970
Number of surgeries	206,611	139,418	21,561	398,348
Hospital staying	1.16E+06	722,792	135,075	2.17E+06
Number of medical cases	7.94E+06	4.14E+06	968,029	1.30E+07
Population	6.35E+06	4.67E+06	641,597	1.27E+07
Number of hospital beds	70,127	40,262	5,055	129,372
GRDP	31,877	7,728	20,183	61,778
Number of house per 1000	405.7	26.67	368.3	453.6
Number of volunteer cases	109,968	66,595	23,664	219,194
Unemployment rate	3.67	0.645	2.1	4.9

Table 4: Regions

Regions	Frequency	Percentage
Seoul	1,584	20.1
Busan	514	6.52
Daegu	379	4.81
Incheon	433	5.5
Gwangju	254	3.22
Daejeon	271	3.44
Ulsan	161	2.04
Jeju	170	2.16
Gyeonggi	1,896	24.06
Gyeongbuk	402	5.1
Gyeongnam	468	5.94
Chungbuk	267	3.39
Chungnam	325	4.12
Jeonbuk	262	3.33
Jeonnam	272	3.45
Gangwon	221	2.8

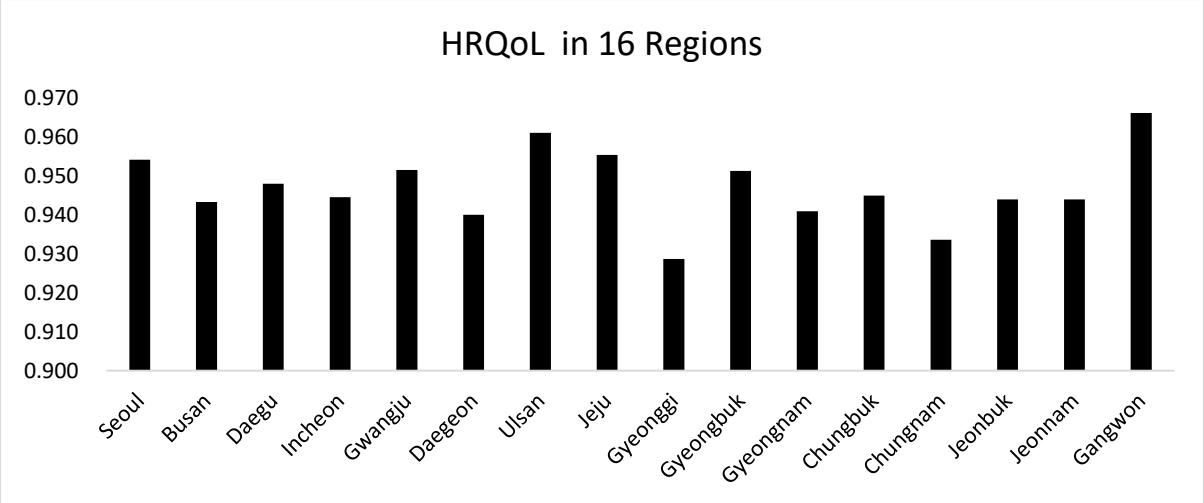

Figure 8: HRQoL average scores by regions

Source: Authors own calculations

Annex 3

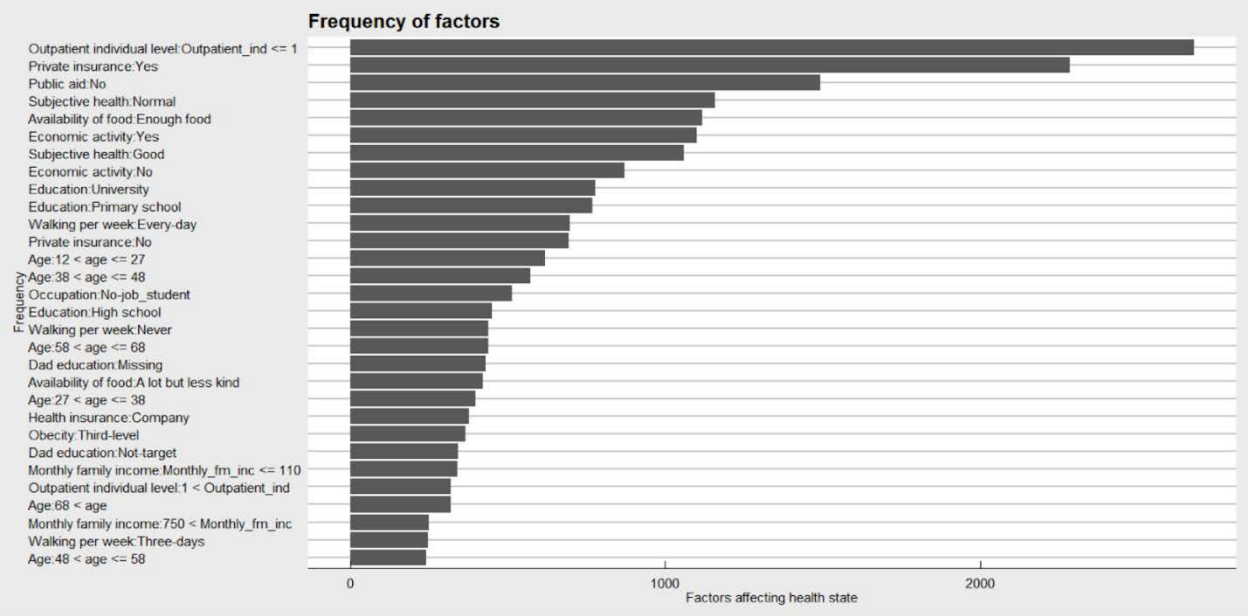

Figure 11: Most frequent factors

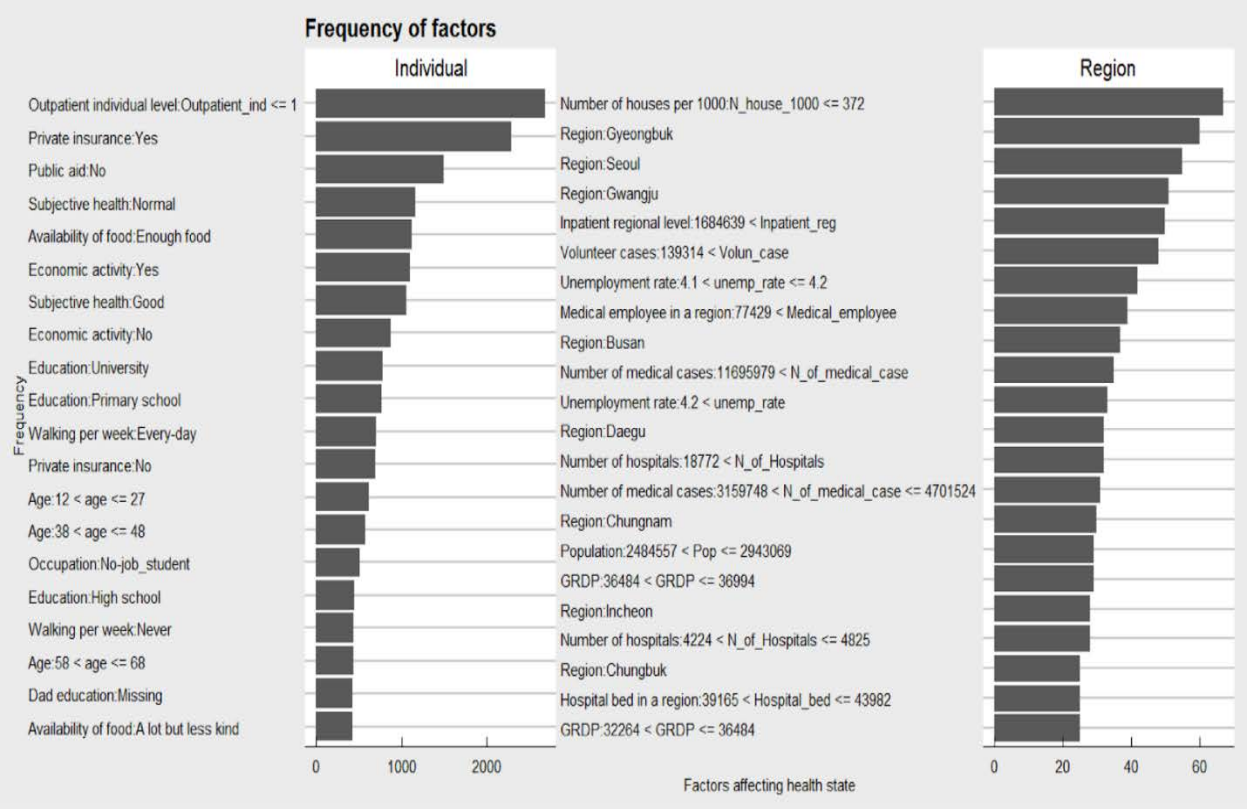

Figure 12: Most frequent factors by level

Table 5: Hyper-parameter description

Type	MAE	MSE	RMSE	RMSLE	Hyper-parameter
Gradient boosting model	0.052	0.009	0.094	0.055	Number of trees=10000, maximum depth of trees=11, minimum number of rows=4, column sample rate=0.73, column sample rate change per level=0.91, column sample rate per tree=0.97, learning rate=0.64, learning rate annealing=0.99, stopping rounds=5, stopping tolerance=1e-4, stopping metric = RMSE
Random forest	0.043	0.006	0.076	0.044	Number of trees=550, sample rate=0.61, maximum depth of trees=7, minimum number of rows=1, column sample rate change per level=0.9, column sample rate per tree=0.94, stopping rounds = 5, stopping tolerance = 1e-4, stopping metric = RMSE
Deep learning	0.047	0.007	0.082	0.048	Activation function=hyperbolic tangent function, hidden layers=20 and 20 cells, input dropout ratio=0, learning rate=0.01, learning rate annealing=1.0E-8, epochs=5, stopping metric=RMSE, stopping tolerance=1e-2, stopping rounds=2, constraint from lasso=7.4E-5, constraint from ridge=5.9E-5
Stacked model	0.036	0.003	0.059	0.034	Base models: gradient boosting model, random forest, deep learning, meta learner algorithm: generalized linear model

Evaluation values

MAE: $\frac{\sum_{i=1}^{n}|\hat{y}_i - y|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}$

MSE: $\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n} = \frac{\sum_{i=1}^{n}(e_i)^2}{n}$

RMSE: $\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(e_i)^2}{n}}$

RMSLE