RESEARCH ARTICLE

AN EXPERIMENTAL COMPARISON OF TRADITIONAL AND MACHINE LEARNING METHODS PREDICTION PERFORMANCES: A STUDY ON HEALTH OUTCOMES

Songül ÇINAROĞLU [*]

ABSTRACT

Machine learning techniques can identify the non-linear patterns in a dataset and can uncover hidden relationships. Random forest is one of the modern machine learning techniques that provides an alternative to traditional classification methods such as logistic regression. In this study it is aimed to compare the prediction performance of logistic regression with that of random forest and to identify the predicting factors of public health outcomes at a provincial level. The data representing 81 provinces of Turkey are taken from the Turkish Statistical Institute for the year 2013. Life expectancy at birth and mortality are chosen as the public health outcomes. Three different random forest models are constructed by determining the number of trees: 50, 100, and 150. The prediction results of different methods are recorded by changing the "k" parameter from 3 to 20 in k-fold cross validation. The Area Under the ROC Curve (AUC), sensitivity, and specificity are considered as performance measures. The study results reveal that the differences between the prediction model performances to predict health outcomes are statistically significant (p<0.000). Moreover, logistic regression outperformed random forest models. The decision tree graphs show that the most important predictor variables for mortality are the total number of beds and for life expectancy at birth, the percentage of higher education graduates. In the light of this study, it is highly recommended for health professionals to be more aware about increasing potential of modern prediction methods in health services research.

Keywords: *Machine learning, logistic regression, random forest, health outcomes*

ARTICLE INFO

[*] *Assist. Prof. Dr., Hacettepe University, FEAS, Department of Health Care Management, cinaroglus@hacettepe.edu.tr*

https://orcid.org/0000-0001-5699-8402

Recieved: 01.12.2018
Accepted: 07.01.2020

Cite This Paper:

Cinaroglu, S. (2020). An Experimental Comparison of Traditional and Machine Learning Methods Prediction Performances: A Study on Health Outcomes. Hacettepe Sağlık İdaresi Dergisi, 23(1): 23-40.

ARAŞTIRMA MAKALESİ

GELENEKSEL VE MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN TAHMİN PERFORMANSLARININ DENEYSEL KARŞILAŞTIRMASI: SAĞLIK SONUÇLARI ÜZERİNE BİR ÇALIŞMA

Songül ÇINAROĞLU [*]

ÖZ

Makine öğrenmesi teknikleri veri setinde doğrusal olmayan desenleri ve gizli ilişkileri tanımlayabilmektedir. Rastgele orman, modern makine öğrenmesi tekniklerinden birisi olarak lojistik regresyon gibi geleneksel sınıflama yöntemlerine alternatif oluşturmaktadır. Bu çalışmada il düzeyinde halk sağlığı sonuç göstergelerini tahmin etmek üzere lojistik regresyon ve rastgele orman tahmin performanslarının karşılaştırılması amaçlanmıştır. Veriler Türkiye genelinde 81 ili temsil etmek üzere 2013 yılı için Türkiye İstatistik Kurumu'ndan temin edilmiştir. Sağlık sonuç göstergesi olarak doğuşta beklenen yaşam süresi ve mortalite seçilmiştir. Ağaç sayısının 50, 100 ve 150 olarak belirlendiği üç farklı rastgele orman modeli oluşturulmuştur. Tahmin yöntemlerinin karşılaştırılmasında "k" parametresinin 3 ile 20 arasında belirlendiği k-kat çapraz geçerlilik yöntemi kullanılmıştır. Performans ölçüsü olarak ROC Eğrisi altında kalan alan, duyarlılık ve seçicilik kullanılmıştır. Çalışma sonuçları sağlık sonuçlarının tahmininde tahmin modeli performanslarının istatistiksel olarak farklı olduğunu ortaya koymaktadır (p<0,000). Ayrıca, lojistik regresyon yöntemi rastgele orman modellerine göre daha iyi performans sergilemektedir. Karar ağacı grafiği mortalitenin tahmininde en önemli değişkenin toplam yatak sayısı, doğuşta yaşam beklentisinin tahmininde yüksek öğrenim mezun yüzdesi olduğunu göstermektedir. Çalışma sonucunda sağlık profesyonellerine sağlık ile ilgili araştırmalarda modern tahmin yöntemlerinin artan potansiyeli konusundaki farkındalıklarını yükseltmeleri tavsiye edilmektedir.

Anahtar Kelimeler*: Makine öğrenmesi, lojistik regresyon, rastgele orman, sağlık sonuçları*

MAKALE HAKKINDA

[*] *Dr. Öğr. Üyesi., Hacettepe Üniversitesi, İİBF, Sağlık Yönetimi Bölümü, cinaroglus@hacettepe.edu.tr*

https://orcid.org/0000-0001-5699-8402

Gönderim Tarihi: 01.12.2018
Kabul Tarihi: 07.01.2020

Atıfta Bulunmak İçin:

Çınaroğlu, S. (2020). An Experimental Comparison of Traditional and Machine Learning Methods Prediction Performances: A Study on Health Outcomes. Hacettepe Sağlık İdaresi Dergisi, 23(1): 23-40.

I. INTRODUCTION

Outcome measures are important tools to determine the impact of health care and the quality of health services. The focus on these measures is to improve quality of life through prevention and treatment of diseases. The outcome information is used for research for the development of clinical practice and to bridge the gap between what is done and what is actually accomplished (Pereira et al., 2004). Policy makers use health outcomes for public health planning. The effect of public spending on health is usually measured by health outcome variables such as life expectancy at birth (LE) and mortality (M) (Gani, 2009). These measures are historically recognized in the literature as the best outcome measures in health and are proxies for health outcomes (Crémieux et al.,1999). LE is defined as an indicator of the number of years a newborn infant would live if the existing conditions of M at the time of its birth remain the same throughout its life span (Halicioglu, 2011). M rates are also used as an indicator of health outcomes, they are referring to the state of being subject to death (Pereira et al., 2004).

There exists a huge difference between developed countries and developing ones in terms of the level of health outcomes. It is well known that industrialized or developed nations achieve a high level of economic and social development, they prioritize the health needs of their population and improve health coverage. However, the relationship between health outcomes and the development of the country is a controversial issue in the literature (Wright and Walley, 1998). Acemoglu and Johnson (2006) studied the effect of health outcomes on economic welfare by focusing on the effect of LE on economic growth. The results of their study suggest that there is no evidence that a large increase in LE causes significant increase in per capita economic growth. In line with this, Gilligan and Skrepnek (2015) suggest that, the level of LE differed between non-industrialized and industrialized nations. Non-industrialized, less developed nations were associated with adjusted life expectancies lower than their industrialized peers. Studies revealed that political and social instabilities are one of the major causes of inequalities, exasperating barriers to access healthcare services in developing countries (Kyriopoulos et al., 2014). The dynamic nature of economic, social, and population dynamics in developing countries makes it difficult to unravel predictors of health outcomes (Wagstaff, 2000). Moreover, there is a huge literature relating to the use of LE and M as outcome measures of disease level (Lee, 2019). However, to our knowledge there is a scarcity of knowledge about public health outcomes at a province level in a developing country.

As a long-term member of the Organization for Economic Co-operation and Development (OECD), Turkey has made considerable advances in improving the quality of life of its citizens over the last two decades. OECD better life statistics reports that LE at birth in Turkey is 77 years for the year 2016. This is three years lower than the OECD average of 80 years. Moreover, LE for women is 79 years, compared with 74 for men for the year 2016 (OECD, 2016). It is apparent that this developing trend in Turkey is parallel to the global improvement in accessibility of health care services. To support this notion, Hitiris and Posnett (1992) state that global improvements in the incidence of poverty, adult literacy, sanitation, nutrition, and access to safe drinking water are generally considered to have positively impacted LE since the 1990's. Studies that focus on Turkey report that through the 1990's there has been a significant decrease in M. Celik and Hotchkiss (2000) support the view that a decrease in purchasing poverty, developments in educational opportunities, and improvements in health care services, play a major role in this process. Furthermore, compared with other middle-income peers like Egypt, Lebanon, and Iran, Turkey has a declining trend in fertility and M as a result of intermediate socioeconomic trend since the 1990's (Omran and Roudi, 1993). Scholars suggest that improvement trend in health outcomes is especially prominent for Turkey. Lichtenberg et al., (2014) supports the view that there have been enormous gains in longevity in Turkey since the year 2000. The level of longevity growth in Turkey was also greater than other middle and low income European countries. Atun (2015) concurs with Lichtenberg et al. (2014) and reports that maternal and child mortality among rural and urban populations decreased significantly between 2003 and 2008. Free health care services and reduced cost sharing are possible reasons for these improvements. Notably, satisfaction from health care services grew from 39.5% in 2003 to 75.9% in 2011 (Atun, 2015).

Despite this positive trend, cardiovascular diseases and diabetes are the most common diseases in Turkey (MoH, 2017). Sozmen et al., (2015) state that the diabetes burden is a growing trend in recent years and that this trend will continue. In these circumstances, obesity and other diabetes risk factors need urgent action in Turkey. Baser et al. (2013) analyzed the increasing trend in cardiovascular diseases on health outcomes in Turkey. Study results show that increasing number of cardiac surgery patients, and the cost of angiograms to check for heart disease effect health outcomes. To fight against cardiovascular diseases and diabetes, Kilic et al., (2015) suggest the need for early detection and screening programs throughout Turkey. While this increasing trend of the general health status of Turkish people is encouraging (Atun, 2015), previous literature shows that countries with greater disparity between urban and rural areas have been shown to have worse overall population health (Cilingiroglu and Yardim, 2014). This increases the need for further studies to understand the determinants of health outcomes at provincial level. To the best of our knowledge, no study to date has been published on the determinants of public health outcomes in Turkey at province level.

Machine learning methods have proven their usefulness by discovering hidden information and patterns among recorded health outcomes (Rosset et al., 2010). By managing large quantities of data, health outcomes can be estimated more reliably and health policy makers will be able to manage health outcomes more effectively. Scholars argue that machine learning techniques such as random forest (RF) have better accuracy and lower error rates than the traditional classification methods do (Kurt et al., 2008; Sut and Simsek, 2011). RF is able to identify non-linear patterns in the data and can improve the predictive capability of commonly used linear methods (Sut and Simsek, 2011). Although there is literature about the predictors of health outcomes at the disease level (Kurt et al., 2008; Sut and Simsek, 2011), it provides scant information about the predictors of health outcomes at the community level through newer prediction methods such as RF. Based on this theoretical background and empirical findings, the aim of this study is to compare RF with the general linear model of logistic regression (LR). Moreover, this study will go a step further to seek a basis for future studies by comparing the use of the traditional and modern prediction techniques to explore the predictor factors of health outcomes at the provincial level in Turkey.

The remainder of this paper is divided into five sections. The next section presents a brief overview about the materials and the methods used in the study. Section three provides information about the study results. The fourth section discusses the study results and gives a brief overview about policy implications. Last, the fifth section summarizes the study results and makes recommendations for future studies.

II. METHODS

2.1. Study Design Process

The analysis process began with an interpretation of the summary statistics. During the preliminary analysis process, the correlations among and between the study variable groups were analyzed through the Spearman correlation coefficient and were presented on a correlogram. None of the correlations were higher than 0.75. Thus, all the selected variables were included in the analysis. Next, the median values as a cut-off point were used to dichotomize the continuous outcome variables, which are LE and M. After ensuring that the outcome variables were balanced, they were recoded as either 1 or 0.

Following this step, the LR and RF models were applied to the dataset by changing the “k” parameter from 3 to 20 in cross validation and building three models for RF by generating 50, 100, and 150 trees. The Receiver Operating Characteristics (ROC) curves were used to compare the sensitivity and the specificity of these classifiers. The Area under the ROC Curve (AUC), sensitivity, and specificity were computed and compared using the Kruskall-Wallis variance analysis. In this study, the AUC values are presented on a heatmap, which is a measure of the overall distinctive power of the prognostic variable. A value of 1 indicates perfect differentiation. A value of 0.5 equals random prediction, and a value lower than 0.5 indicates no discriminative power. Sensitivity measures how well the test identifies those with the disease. The specificity measures how well the test excludes those who do not have the disease (Sut and Simsek, 2011). Finally, the prediction results of the best performed prediction model were used to present the predictors of LE and M on a decision tree graph. Analysis was performed in R program.

2.2. Dataset

Data representing 81 provinces of Turkey are taken from Turkish Statistical Institute (TurkStat) statistics for the year 2013. LE, and M for total population are determined as dependent variables. Existing literature suggest that utilization of health services, poverty, satisfaction from health status, technical capacity of health institutions such as number of beds, the level of education are predictors of health outcomes (Wagstaff, 2000; Crisp et al., 2000; Berkman et al., 2011; Fenton et al., 2012). In the light of present literature: the number of applications per doctor, satisfaction rate with health status, total number of beds, percentage of households in middle or higher income groups, percentage of higher education graduates are determined to be proxy variables.

2.3. Predictive Methods and Applications

Two different machine learning methods were adopted and compared to construct predictors of public health outcomes which are LE and M.

2.3.1. Logistic Regression Model

LR is a suitable classifier and traditional predictor method, when the relationship between input and output variables is linear and the data balanced between groups. In medical research, LR is one of the popular methods to predict health outcomes when the health outcome is dichotomous, as with M (Muchlinksi et al., 2016).

The likelihood (odds ratio) represents the ratio between the probability *p* that the dependent variable *Y* is 1 and the probability *1-p* that the dependent variable *Y* is 0. The natural logarithm of odds (Logit) is a linear function of the explanatory variables X_1, X_2, ……..., X_n and takes values from -∞ to +∞ (Van Den Eeckhaut et al., 2006).

$$Logit\,(p) = \; In\,\frac{p}{1-p} = \; \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n \qquad \textbf{(1)}$$

β_1, β_2……β_n are the coefficients that measure the contribution of the independent variables X_1, X_2 …….X_n to the dependent variable. If the coefficient β is positive, e^{β} > 1, and the factor has a direct correlation with the dependent variable. If β is negative, then, e^{β} is between 0 and 1 (Trigila et al., 2015).

2.3.2. Random Forest

RF offers an alternative approach for increasing predictive accuracy (Muchlinksi et al., 2016). A widely used machine learning model, RF is based on decision theory developed by Breiman, Friedman, Olshen and Stone (1984). RF uses Classification and Regression Tree (CART) algorithm to generate trees. If the response variable is a factor, RF performs classification, but if the response is continuous the RF performs regression (Liaw and Wiener, 2002). When the response variable is dichotomous, RF grows a forest of classification trees (Grömping, 2009). CART first grows a very large tree and then prunes it. Grömping (2009) state that pruning a large tree instead of growing only a small number of trees improves the prediction performance of RF.

Unlike LR, RF provides predictive models for classification and regression. RF builds a forest of classification trees for the dataset. The classification tree for a binary outcome variable uses a training sample of "n" cases. Case *i* has a vector of covariates x_i used to build a tree-structured classification rule. Repetitive partitioning splits the training sample into increasingly homogeneous groups by inducing a partition on the explanatory space. Three types of splits that use the vector of input *x* include (Siroky, 2009; Muchlinski et al., 2016):

1. Univariate split: Is $x_{i} \leq t$?
2. Linear combination split: Is $\sum_{i=1}^{p}(w_i\, x_i) \leq t$?
3. Categorical split: Is x_i ε S.

The split searches for the separation that best differentiates the cases in the training sample into two maximally homogenous groups. These have been defined in various ways in the literature. Formally, the tree is grown by using equation (2): where the regions R_J and the coefficients β_j are estimated from the data. The R_J are usually disjointed, and the β_j is the average of the Y values in the R_J.

$$Y = \sum_{j=1}^{r} \beta_j I\left(x\, \varepsilon\, R_J\right) + \varepsilon \qquad \textbf{(2)}$$

2.3.3. Comparison of Logistic Regression with Random Forest

There are several methods to compare prediction performances of different prediction methods. One of these performance evaluation methods is a ROC graph. This graph illustrates the performance of a binary classifier (Muchlinksi et al., 2016). The ROC graph is easily summarized by a single metric called the AUC. This is one way to visualize the predictive performance of a binary classifier based on its performance (Sut and Simsek, 2011). In other words, the AUC is a univariate description of the ROC Curve. The larger the AUC score, the better the model's predictive performance (Muchlinksi et al., 2016). An additional way to improve performance of a dichotomous classifier is cross-validation. This is a widely used strategy because of its simplicity and universality. k-fold is one of the most well-known types of cross-validation. When k is large, the number of training instances become large in each iteration (Hastie et al., 2009). In k-fold cross-validation, the data is first partitioned into equally sized folds and cross-validation procedure is applied. This procedure involves breaking data into different folds. A number of these folds are used to train the model, while a separate fold is held out to test the predictions made by the model in the training data (Hastie et al., 2009). Tenfold cross validation is one of the popular cross validation techniques. In tenfold cross validation, firstly the data set is randomly divided into ten equal parts. Then a tree is built based on 90% of the data (named as "training set") and tested using the remaining 10% of the data (named as "testing set"). After that, another tree is generated similarly based on different training and testing data. This process is running ten times using different training and testing datasets (Li et al., 2001). This has been widely used to assess the relative predictive performance of statistical models in many disciplines and a popular strategy for algorithm selection (Hastie et al., 2009).

There is a previous knowledge about comparison of LR performance results with machine learning methods to predict health outcomes. One of these studies Maroco et al. (2011) compare LR model performance with RF to predict dementia. Study results state that RF performs well compared to other methods. Camdeviren et al. (2007) compare LR model and classification tree performances to evaluate the diagnosis of postpartum depression data. The study found that classification tree methods gave more information with greater detail on diagnosis by evaluating the number of risk factors together, unlike the LR model. Despite the existence of literature comparing LR with machine learning methods to predict health outcomes at disease level, there are hardly studies assessing the predictors of public health outcomes at province level. This study aims to fill this gap and explore predictors of public health outcomes at province level, while comparing LR and RF prediction performance results.

III. RESULTS

3.1. Summary Statistics

The median scores of the variables for health outcome indicators belongs to 81 provinces of Turkey are as follows; LE [median 78.10; min. 75; max. 81], M [median 2822; min. 510; max. 54766]. Summary statistics for predictor variables are reported in Table 1. Number of applications per doctor [median 5787; min. 2763; max. 8067], satisfaction rate with health status (%) [median 72; min. 59.20; max. 80.80], total number of beds [median 1338; min. 150; max. 33581], percentage of households in middle or higher income groups (%) [median 34.1; min. 16.30; max. 58.90], percentage of higher education graduates (%) [median 12.9; min. 8.60; max. 22.70].

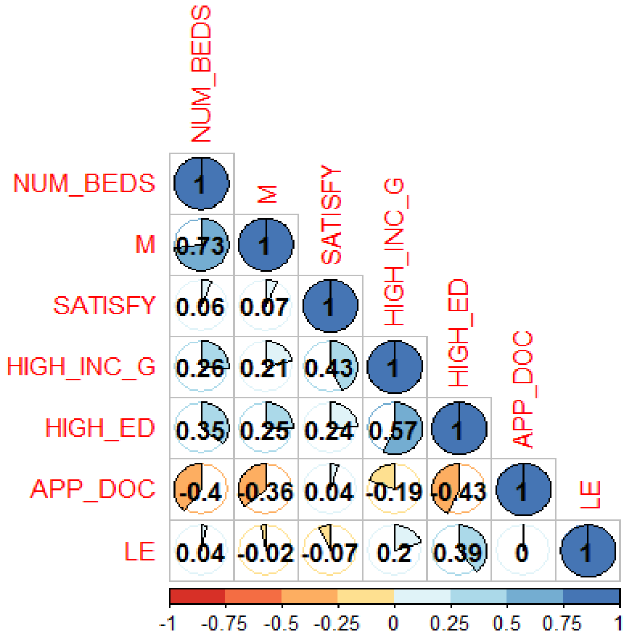
Table 1. Summary Statistics

Health Outcome Indicators	**Short Name**	**N**	**Median**	**Min.**	**Max.**
Life expectancy at birth	LE	81	78.10	75	81
Mortality	M	81	2822	510	54766
Predictor Variables	**Short Name**	**N**	**Median**	**Min.**	**Max.**
Number of applications per doctor	APP_DOC	81	5787	2763	8067
Satisfaction rate with health status (%)	SATISFY	81	72	59.20	80.80
Total number of beds	NUM_BEDS	81	1338	150	33581
Percentage of households in middle or higher ıncome groups (%)	HIGH_INC_G	81	34.1	16.30	58.90
Percentage of higher education graduates (%)	HIGH_ED	81	12.9	8.60	22.70

3.2. Correlations Among Study Variables

Correlations among study variables are provided in Figure 1. It is seen that all correlation coefficients are lower than 0.75. It is clear to say that, there is little fear for multicollinearity. Thus, it is decided to consider all variables in the analysis.

Figure 1. Correlation Coefficients



See Table 1 for labels. In the above figure, correlations with p-value > 0.05 are considered as insignificant and they are not presented on the correlogram.

3.3. Binary Coding of Health Outcome Indicators

In this study continuous health outcome indicators were dichotomized by using the median values as cut-off points. It is seen that, the number of observations for both of two binary variable groups are balanced for LE and M, respectively. Table 2 shows cut-off points and binary coding for health outcome indicators. All outcome indicators are recoded as 0 and 1 (e.g. LE <78.10 was recoded as 0).

Table 2. Cut-Off Points and Binary Coding for Health Outcome Indicators

Health Outcome Indicators	N	Median	Cut-off Points	n	%	Recoded As
Life Expectancy at Birth (LE)	81	78.10	LE ≥ 78.10	39	48.1	1
			LE < 78.10	42	51.9	0
Mortality (M)	81	2822	M ≥ 2822	40	49.4	1
			M < 2822	41	50.6	0

3.4. Descriptive Statistics of Prediction Model Performances

Prediction performance results of LR and RF methods to predict LE and M were recorded by using AUC, sensitivity and specificity values, “k” parameter changed from 3 to 20 in the cross validation. Three different RF models are constructed by generating 50, 100 and 150 trees in the forest. Thus, 18 different AUC, sensitivity and specificity values were recorded for four different applications (LR, RF 50, RF 100 and RF 150). Table 3 shows mean values and standard deviations of AUC, sensitivity and specificity values of different prediction model applications. It is seen that mean values for LR is high and it has achieved superior performance to others. Moreover, the RF model generated by using 50 trees yields better prediction results among other RF models.

Table 3. Descriptive Statistics of Prediction Model Performances

Health Outcome Indicators	Performance Measures	Logistic Regression			Random Forest_50			Random Forest_100			Random Forest_150		
		N	**Mean**	**Std. Dev.**	**N**	**Mean**	**Std. Dev.**	**N**	**Mean**	**Std. Dev.**	**N**	**Mean**	**Std. Dev.**
Life Expectancy at Birth (LE)	**AUC**	**18**	0.6690	0.0267	**18**	0.5827	0.0395	**18**	0.5810	0.0455	**18**	0.5814	0.0427
	Sen.	**18**	0.7090	0.0208	**18**	0.6534	0.0722	**18**	0.6362	0.0642	**18**	0.6256	0.0570
	Spec.	**18**	0.4914	0.0295	**18**	0.4700	0.0403	**18**	0.4572	0.0441	**18**	0.4658	0.0433
Mortality (M)	**AUC**	**18**	0.9832	0.0098	**18**	0.9712	0.0136	**18**	0.9709	0.0125	**18**	0.9698	0.0161
	Sen.	**18**	0.9505	0.0236	**18**	0.9445	0.0112	**18**	0.9444	0.0140	**18**	0.9444	0.0112
	Spec.	**18**	0.9638	0.0300	**18**	0.9625	0.0183	**18**	0.9555	0.0153	**18**	0.9566	0.429

Abbreviations: AUC: Area Under the ROC curve; **Sen.:** sensitivity; **Spec.:** Specificity; **Std. Dev.:** Standart deviation

3.5. Statistical Differences of Prediction Model Performances

Statistical differences of AUC, sensitivity and specificity values of prediction models (LE, RF_50, RF_100 and RF_150), produced by changing "k" parameter in cross validation from 3 to 20, compared using Kruskall-Wallis variance analysis (see Table 4). Study results ascertain that; prediction model performances are statistically significant in terms of both of two health outcome indicators. Some of the descriptive statistics for predicting LE are; AUC (X^2= 36.919, p<0.001), sensitivity (X^2= 20.536, p<0.001); specificity (X^2= 7.537, p<0.05). M are; AUC (X^2= 23.141, p<0.001), sensitivity (X^2= 37.771, p<0.001) and specificity (X^2= 43.511, p<0.001).

Table 4. Statistical Differences of Prediction Model Performances

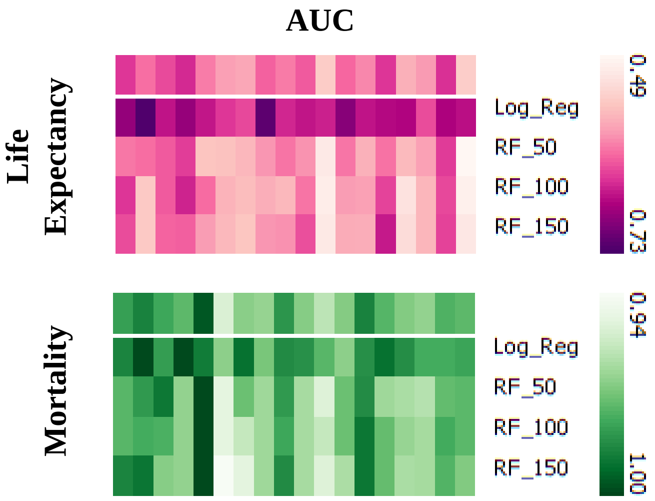
Prediction Models	Life Expectancy											
	AUC				Sensitivity				Specificity			
	N	Mean Rank	X^2	p	N	Mean Rank	X^2	p	N	Mean Rank	X^2	p
LR	18	62.39	36.919	<0.001	18	53.53	20.536	<0.001	18	46.95	7.537	<0.05
RF_50	18	26.11			18	37.32			18	35.88		
RF_100	18	28.64			18	29.22			18	29.56		
RF_150	18	28.86			18	25.03			18	33		
Prediction Models	**Mortality**											
	AUC				**Sensitivity**				**Specificity**			
	N	**Mean Rank**	X^2	**p**	**N**	**Mean Rank**	X^2	**p**	**N**	**Mean Rank**	X^2	**p**
LR	18	51.36	23.141	<0.001	18	13.89	37.771	<0.001	18	10.92	43.511	<0.001
RF_50	18	31.83			18	45.21			18	41.35		
RF_100	18	31.94			18	44.75			18	47.53		
RF_150	18	30.86			18	43.89			18	47.89		

Abbreviations: X^2: Chi-square; **RF:** Random Forest; **LR;** Logistic Regression; **AUC:** Area Under the ROC Curve

3.6. Heatmap of AUC for Different Classifiers

AUC values of prediction models, which are produced by changing k” parameter in cross validation from 3 to 20, to predict LE and M are represented on a heatmap. The dark colors in color bar of the heatmap represent for high AUC values, whereas light ones representing low values, ranged from 0 to 1. Study results obtained from the heatmap confirm our priori results and visualize that AUC values for logistic regression are more intense with darker purple and green for predicting LE and M, respectively (Figure 2).

Figure 2. Heatmap of AUC for Different Classifiers

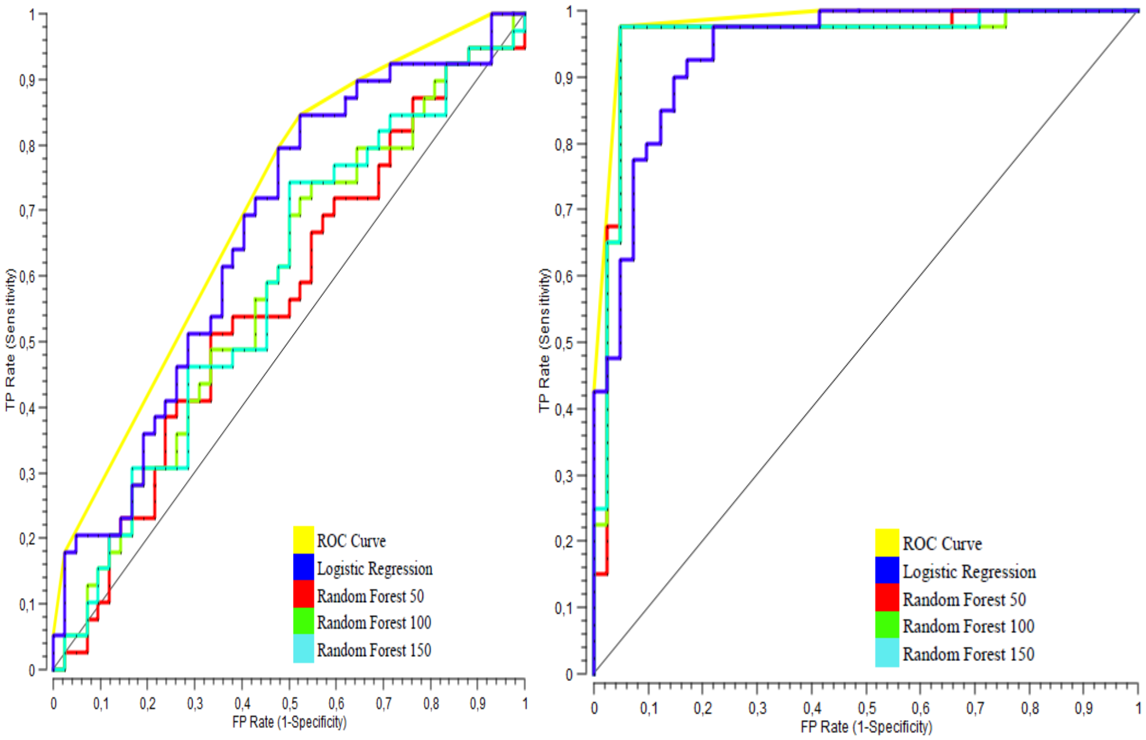


Abbreviations: RF: Random Forest, **Log_Reg:** Logistic Regression, **AUC:** Area Under the ROC Curve

3.7. ROC Curves for Logistic Regression and Random Forest Models

A comparative ROC curves of the prediction models for the prediction of LE and M are shown in Figure 3 (a) and (b), respectively. The AUC of the LR, represented with dark blue color was superior to three different RF models and LR is much closer to the ROC curve (represented with yellow) than RF models.

Figure 3. ROC Curves for Logistic Regression and Random Forest Model for Predicting Health Outcomes



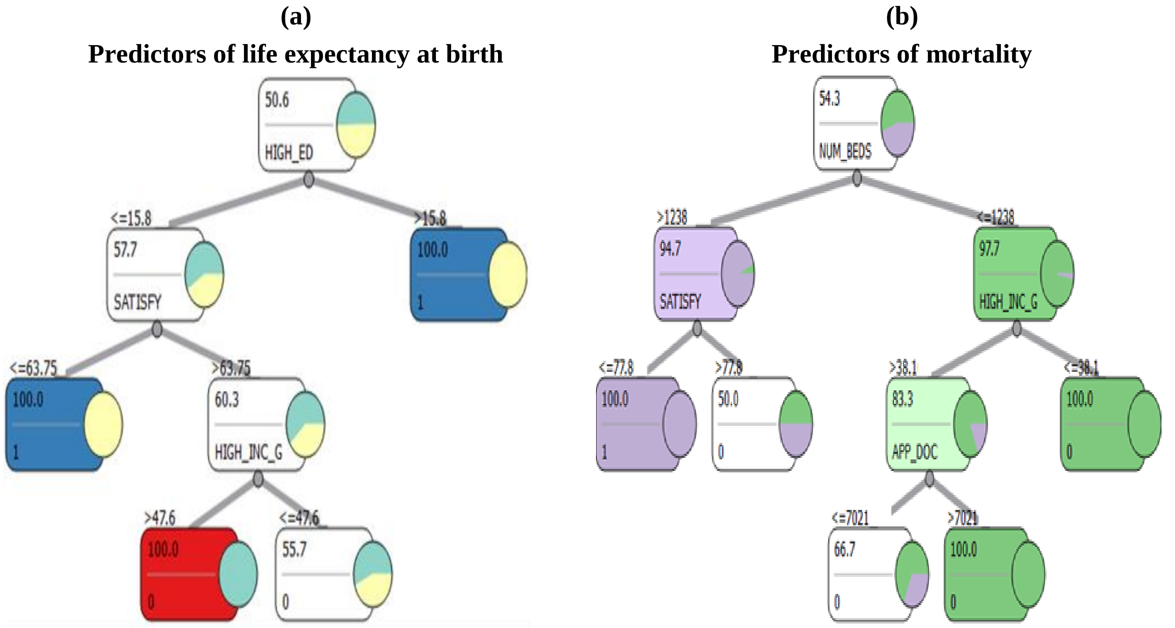
(a)
Outcome variable: Life expectancy at birth

(b)
Outcome variable: Mortality

3.8. Predictors of Life Expectancy at Birth and Mortality

Figure 4 (a) presents the results of the decision tree generated by using RF to found out predictors of LE for total population, number of trees determined as 50 which has a superior prediction performance (mean AUC=0.5827) compared with 100 and 150 trees. The decision tree graph shows that percentage of higher education graduates is the most important predictor variable of LE at province level in Turkey. Furthermore, there exists four groups in the decision tree. First group consists of percentage of higher education graduates >15.8%. The second group consists of percentage of households in middle or higher income groups ≤47.6%, satisfaction rate with health status >63.8% and percentage of higher education graduates ≤15.8%. The third group composed of percentage of households in middle or higher income groups >47.6%, satisfaction rate with health status >63.8% and percentage of higher education graduates ≤15.8%. Lastly, the last group consists of satisfaction rate with health status ≤63.8% and percentage of higher education graduates ≤15.8%. Figure 4 (b) figures out the results of the decision tree generated by RF to found out predictors of M for total population, number of trees determined as 50 which has a superior prediction performance (mean AUC= 0.9712) compared with 100 and 150 trees. The decision tree graph shows that number of beds is the most important variable in the prediction of total M at province level in Turkey. The decision tree consists of five groups. The first group consist of percentage of households in middle or high-income groups ≤38.1% and total number of beds ≤1238. The second group consists of total number of applications per doctor >7021, percentage of households in middle or high-income groups >38.1%, total number of beds ≤1238. The third group composed of total number of applications per doctor ≤7021 percentage of households in middle or high-income groups >38.1%, total number of beds ≤1238. The fourth group consists of satisfaction rate with health status >77.8% and total number of beds >1238. Finally, the last group of provinces for prediction of total mortality consists of satisfaction rate with health status ≤77.8% and total number of beds >1238.

Figure 4. Predictors of Life Expectancy at Birth and Mortality



(a)
Predictors of life expectancy at birth

(b)
Predictors of mortality

Labels: HIGH_ED: Percentage of higher education graduates (%), **SATISFY:** Satisfaction rate with health status (%), **HIGH_INC_G:** Pecentage of households in middle or higher income groups (%), **NUM_BEDS:** Total number of beds, **APP_DOC:** Number of applications per doctor

IV. DISCUSSION

Machine learning techniques are advantageous to statistical models for classification and regression problems (Khalilia et al., 2011). In statistics, modern and traditional methods belong to two different cultures. Breiman (2001) defines the differences between these two statistical modeling perspectives. Traditional models, such as LR, not only aim at "predicting," but also aim at "explaining" effects. In reality, prediction accuracy is the primary focus of both cultures. Modern prediction models are progressively used to analyze large medical datasets and clinical records (Khalilia et al., 2011).

In recent years, RF and other modern machine learning techniques have been applied successfully to various areas of health research, including genetic epidemiology and microbiology (Samant and Agarwal, 2018). Easy interpretability, robustness to outliers, and other influential observations are the key reasons why RF has increased its popularity (Hastie et al., 2009). Moreover, this new data analysis tool provides an alternative method to analyze "unusual" settings such as extreme cases, big data problems, and rare events (Hegelich, 2016).

Previous studies have compared RF and LR by using various healthcare datasets such as predicting individual health expenditure and disease risks for patients (Khalilia et al., 2011). These studies reported that RF had higher accuracy and lower error rates than traditional prediction methods did (Sut and Simsek, 2011; Couronne et al., 2017). Another benefit of RF is it can identify non-linear patterns in the dataset and can improve upon the predictive capability of traditional methods (Khalilia et al., 2011).

Although machine learning methods are commonly applied to studies at the disease level, there are still ways to improve our understanding of the predictors of public health outcomes. This study fills that gap by comparing the LR and RF prediction performances to predict LE and M at the provincial level in Turkey. According to several benchmark performance study results, RF is a promising algorithm in AUC, sensitivity, and specificity (Couronne et al., 2017). However, our study results differ from previous research. More clearly, our expectations were unmet when our results showed that LR outperformed the RF models. Computational scientists argue that the performance of the prediction methods may depend on dataset characteristics such as sample size, correlations between variables, and meeting the assumption of normal distribution (Couronne et al., 2017). In the light of previous concerns about the effect of dataset, relationships between the variables and the distinguishable features of study models on analysis results, we discuss the potential effects below.

Our benchmarking study considered a small sample size at the provincial level and we achieved high predictive accuracy for both of the models. In addition, our prediction accuracy results were close to each other. Previous evidence agrees with our study results and indicates that a small sample size can provide better insight into the model (Zhao et al., 2001). Moreover, as the number of observations in the dataset increases, the difference between RF and LR also increases slightly (Couronne et al., 2017). Our study results, based on a small sample dataset at the provincial level, support these findings and show that RF and LR prediction results are close to each other. On the other hand, the LR is robust to the violation of the normality assumption (Begueria and Lorente, 2002). The study results support this earlier finding because nonnormal distributed independent variables exist in our dataset. Previous studies agree with our results and emphasize the superior performance of LR compared to other data mining methods. Moreover, previous research has found that LR is a superior algorithm and can often be improved by adopting techniques such as cross validation, shrinking the parameters, and imposing a margin constraint in the separable case, or various forms of averaging (Ng and Jordan, 2002).

Another finding of this study is that the percentage of higher education graduates for LE at birth and the total number of beds for M are the most important predictors of public health outcomes. Other predictor variables include satisfaction rate with health status, percentage of households in middle or higher income groups, and the number of applications per doctor. To summarize, capacity indicators, satisfaction from health status, position in a high-income group, and the number of doctor applications are predictors of public health outcomes. While there is significant interest in improving health outcomes at the disease level, the assessment of health outcomes at the public level requires more attention. To ensure advantageous outcomes for low socio-economic groups and to achieve egalitarian health services, health policy makers in Turkey must understand the determinants of health benefits.

To the best of our knowledge this study is one of the first to compare traditional and modern prediction models for predicting public health outcomes at the provincial level in a developing country. However, several study limitations deserve consideration. First, this study is based on data at the provincial level. Further studies should examine public health outcomes at the household level. Such studies will allow the examination of household socio-demographic characteristics and rural-urban differences in health outcomes. A further limitation of this study is the use of a cross-sectional design. To clarify determinates of public health outcomes over time, future longitudinal studies are required. The number of the selected predictive factors is another limitation of the present study. Future studies will enrich the current study results by considering other indicators such as social protection, lifestyle, and health behaviors. Moreover, in spite of the fact that technical capacity indicators are strong predictors of health outcomes, this study used the total number of beds only. Further studies should add more technical capacity indicators into the model. Additionally, there exist other machine learning methods in the literature, such as neural networks, which are preferred over LR and have better classification results (Zhanga et al., 1999). To examine the predictive performance of the model, ensemble learning methods such as bagging, boosting, and variants could also be applied on RF. Finally, the results of this study show that increasing the number of trees in the forest does not make big differences in prediction model performance. Supportive evidence from the literature emphasizes that the number of trees did not significantly influence the classification rules (Khalilia et al., 2011). Future studies should compare the prediction performances of study models while changing tuning parameters.

V. CONCLUSIONS

Machine learning methods are alternatives to the traditional prediction methods and constitute one side of the “two cultures” of statistical modeling (Breiman, 2001). Whereas much is known about predicting health outcomes through machine learning methods at the disease level, there is a scarcity of knowledge on the comparison of traditional prediction models with machine learning techniques to predict public health outcomes at the provincial level in a developing country. This study examined the LR and RF prediction performances to predict public health outcomes which are; LE and M. The study compared the LR and RF models and found that, in terms of AUC, sensitivity, and specificity measures, LR had superior prediction performance and statistical significance. The differences in the study results are attributed to the sample size, non-linear relationships in the dataset, and the cross-validation process. In future research, the use of big data at the household level is suggested to confirm the study results in a big public health dataset. From a different perspective, this study explores public health outcomes by a comparison of the two cultures of statistical learning. It may provide inspiration for future studies to incorporate other machine learning methods, such as neural network or support vector machine into the model. Health policy and decision makers should be aware and vigilant about high potential of new prediction methods to better orchestrate and distribute scarce health resources in developing countries.

REFERENCES

Acemoglu, D., & Johnson, S. (2006). *Disease and development: the effect of life expectancy on economic growth*. NBER Working Paper Series, Working Paper, No. 12269. http://www.nber.org/papers/w12269. (25.11.2017).

Atun, R. (2015). Transforming turkey’s health system-lessons for universal coverage. *The New England Journal of Medicine*, 373(14), 1285-1289.

Baser, O., Burkan, A., Baser, E., Koselerli, R., Ertugay, E., & Altinbas. A. (2013). High cost patients for cardiac surgery and hospital quality in turkey. *Health Policy*, 109(2), 143-149.

Begueria, S., & Lorente, A. (2002). *Landslide hazard mapping by multivariate statistics: comparison of methods and case study in the Spanish pyrenees*. Technical report, Instituto Pirenaico de Ecologia, Zaragoza, Spain.

Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J., & Crotty, K. (2011). Low health literacy and health outcomes: an updated systematic review. *Annals of Internal Medicine*, 155(2): 97-107.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science*, 16(3): 199-231.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees,* Chapman and Hall/CRC, Taylor and Francis Group, Boca Raton.

Camdeviren, H., Yazici, A.C., Akkus, Z., Bugdayci, R., & Sungur, M. A. (2007). Comparison of logistic regression model and classification tree: an application to postpartum depression data. *Expert Systems with Applications*, 32(4), 987-994.

Celik, Y., & Hotchkiss, D. R. (2000). The socio-economic determinants of maternal health care utilization in Turkey. *Social Science and Medicine*, 50(12), 1797-1806.

Cilingiroglu, N., & Yardim, M. S. (2014). Approaching socioeconomic inequalities in Turkey by using self-assessed health. *European Journal of Public Health, 24*(2), 25-26.

Couronne, R., Probst, P., & Boulesteix, A. L. (2017). *Random forest versus logistic regression: a large scale benchmark experiment*. Technical Report Number 205, University of Munich, Department of Statistics, http://www.stat.uni-muenchen.de, (29.5.2018).

Crémieux, P. Y., Ouellette, P., & Pilon, C. (1999). Health care spending as determinants of health outcomes. *Health Economics*, 8(7), 627-639.

Crisp, B. R., Swerissen, H., & Duckett, S. J. (2000). Four approaches to capacity building in health: consequences for measurement and accountability. *Health Promotion International*, 15(2), 99-107.

Fenton, J. J., Jerant, A. F., Bertakis, K. D., & Franks, P. (2012). The cost of satisfaction a national study of patient satisfaction, health care utilization, expenditures and mortality. *Archives of Internal Medicine*, 172(5), 405-411.

Gani, A. (2009). Health care financing and health outcomes in pacific island countries. *Health Policy and Planning*, 24(1), 72-81.

Gilligan, A. M., & Skrepnek, G. H. (2015). Determinants of life expectancy in the eastern mediterranean region. *Health Policy and Planning*, 30(5), 624-637.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4), 308-319.

Halicioglu, F. (2011). Modeling life expectancy in Turkey. *Economic Modelling*, 28(5), 2075-2082.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Data Mining, Inference and Prediction. (2nd ed.). Springer.

Hegelich, S. (2016). Decision trees and random forests: machine learning techniques to classify rare events. *European Policy Analysis*, 2(1), 98-120.

Hitiris, T., & Posnett, J. (1992). The determinants and effects of health expenditure in developed countries. *Journal of Health Economics*, 11(2), 173-181.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, 11(51), 1-13.

Kilic, B., Kalaca, S., Unal, B., Phillimore, P., & Zaman, S. (2015). Health policy analysis for prevention and control of cardiovascular diseases in diabetes mellitus in Turkey. *International Journal of Public Health*, 60(1), 47-53.

Kurt, I., Ture, M., & Kurum, T. A. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, 34(1), 366-374.

Kyriopoulos, I. I., Zavras, D., Skroumpelos, A., Mylona, K., Athanasakis, K., & Kyriopoulos, J. (2014). Barriers in access to healthcare services for chronic patients in time of austerity: an empirical approach in Greece. *International Journal for Equity in Health*, 13(54), 1-7.

Lee, R. (2019). *Mortality forecasts and linear life expectancy trends*. In: Bengtsson T., Keilman N. (eds) Old and New Perspectives on Mortality Forecasting. Demographic Research Monographs (A Series of the Max Planck Institute for Demographic Research). Springer, Cham.

Lehr, S., Liu, H., Klinglesmit, S., Konyha, A., Robaszewska, N., & Medinilla, J. (2016). *Use educational data mining to predict undergraduate retention.* IEEE 16th International Conference on Advanced Learning Technologies, Austin, TX, USA. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=andarnumber=7757015, (28.5.2018).

Li, X. B., Sweigart, J., Teng, J., Donohue, J., & Thombs, L. (2001). A dynamic programming based pruning method for decision trees. *INFORMS Journal on Computing*, 13(4), 332-344.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2/3*, S18-S22.

Lichtenberg, F. R., Tatar, M., & Caliskan, Z. (2014). The effect of pharmaceutical innovation on longevity, hospitalization and medical expenditure in Turkey. *Health Policy*, 117(3), 361-373.

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & Mendonça, A. (2011). Data mining methods in the prediction of dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*, 4(299), 1-14.

Muchlinksi, D., Siroky, D., Jingrui, H., & Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 24(1), 87-103.

Ng, A. Y., & Jordan, M. I. (2002). *On discriminative vs. Generative classifiers: a comparison of logistic regression and naive bayes*, Advances in Neural Information Processing Systems 14 (NIPS 2001), Vancouver, British Columbia, Canada. https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf, (28.5.2018).

Omran, A. R., & Roudi, F. (1993). The middle east population puzzle. *Population Bulletin*, 48(1), 1-40. https://www.ncbi.nlm.nih.gov/pubmed/12318382.

Organization for Economic Cooperation and Development. (OECD) (2016). *Better life index*. http://www.oecdbetterlifeindex.org/countries/turkey/. (07.06.2016).

Pereira, C., Murphy K., & Herndon, D. (2004). Outcome measures in burn care. Is mortality dead? *Burns*, *30*(8), 761-771.

Republic of Turkey Ministry of Health (MOH). (2017). *Health statistics year book-2017*. https://dosyasb.saglik.gov.tr/Eklenti/30148,ingilizcesiydijiv1pdf.pdf?0. (10.9.2019).

Rosset, S., Perlich, C., Swirszcz, G., Melville, P., & Liu, Y. (2010). Medical data mining: insights from winning two competitions, *Data Mining and Knowledge Discovery*, *20*(3), 439-468.

Samant, P., & Agarwal, R. (2018). Machine learning techniques for medical diagnosis for diabetes using iris images. *Computer Methods and Programs in Biomedicine*, 157, 121-128.

Siroky, D. S. (2009). Navigating random forests and related advances in algorithmic modeling. *Statistics Surveys*, *3*, 147-163.

Sozmen, K., Unal, B., Capewell, S., Critchley, J., & O'Flaherty, M. (2015). Estimating diabetes prevalence in turkey in 2025 with and without possible interventions to reduce obesity and smoking prevalence, using a modelling approach. *International Journal of Public Health*, 60(1), 13-21.

Sut, N., & Simsek, O. (2011). Comparison of regression tree data mining methods for prediction of mortality in head injury. *Expert Systems with Applications*, 38(12), 15534-15539.

Trigila, A., Iadanza, C., Esposito, C., & Scarascia-Mugnozza, G. (2015). Comparison of logistic regression and random forest techniques for shallow landslide susceptibility assessment in giampilieri (NE Sicily, Italy). *Geomorphology*, 249(15), 119-136.

Turkish Statistical Institute (TurkStat). http://www.turkstat.gov.tr/UstMenu.do?metod=istgosterge. (07.06.2018).

Van den Eeckhaut, V. D., Vanwalleghem, M. T., Poesen, J., Govers, G., Verstraeten, G., & Vandekerckhove, L. (2006). Prediction of landslide susceptibility using rare events logistic regression: a case-study in the flemish ardennes (Belgium). *Geomorphology,* 76(3-4), 392–410.

Wagstaff, A. (2000). *Research on equity, poverty and health outcomes: lessons for the developing world*, HNP Discussion Paper, 28908. http://siteresources.worldbank.org/HEALTHNUTRITIONANDPOPULATION/Resources/281627-1095698140167/Wagstaff-ResearchOn-whole.pdf. (10.10.2017).

Wright, J., & Walley, J. (1998). Assessing health needs in developing countries. *British Medical Journal*, 316(7147), 1819-1823.

Zhanga, G., Hu, M. Y., Patuwob, B. E., & Indrob, D. C. (1999). Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis, *European Journal of Operational Research*, 116(1), 16-32.

Zhao, L., Chen, Y., & Schaffner, D. W. (2001). Comparison of logistic regression and linear regression in modeling percentage data. *Applied and Environmental Microbiology*, 67(5), 2129-2135.