The Classification of Countries' Human Development Index Level Under Economic Inequality by Using Data Mining Classification Algorithms

Esra POLAT (espolat@hacettepe.edu.tr)
Department of Statistics, Faculty of Science, Hacettepe University, 06800, Ankara, Turkey

ABSTRACT

The goal of the study is finding the best data mining classification method, in determining four levels of Human Development Index (HDI) for 100 countries for the data set of 2018. Hence, Naïve Bayes, IBK, KStar, J48, RandomForest, RandomTree, REPTree, SMO, Simple Logistic, Logistic, Multilayer Perceptron methods are implemented on to the data set by WEKA data mining software for classifying 100 countries in terms of HDI levels (very high, high, medium and low) using the explanatory variables as GDP per capita, poverty rate, Net Income Gini index and Wealth Gini index. The results show that best classification method for this data set is Multilayer Perceptron with highest accuracy rate of 88%. Moreover, GDP per capita US$ is found as the most effective variable on determining the HDI levels of countries.

Keywords: *Data Mining, Economic Inequality, Human Development Index, WEKA*

JEL Classification: *C38, I31, I32*

1. INTRODUCTION

Since its introduction in 1990, HDI was aimed to measure human development. It was proved that economic growth and income growth are not the only factors to contribute to well-being (Gallardo, 2009; Andrei and Georgescu, 2018; Georgescu and Kinnunen, 2019; Georgescu et al., 2020a). HDI is created as a composite index having in the component three dimensions of well-being: education, per capita income and life expectancy indicators. HDI is utilized for ranking the countries into four classes of human development. In 2014, United Nations presented fixed cutoff points (COP) in order to determine the four classes of HDI. COPs have been computed using the quartiles of

the distributions of the indicators in the composition of HDI averaged during 2004-2013. These COPs are: low human development- below 0.550; medium human development- 0.550-0.699; high human development-0.700-0.799; very high human development-0.800 and above (Georgescu et al., 2020b). Gross domestic product (GDP) per capita is a measurement of a country's economic output that accounting for population. Although GDP shows you how wealthy a country feels to each of its residents, it is not an exact measure of economic wellbeing. HDI with its three elements income, education and health has turned into the major option to GDP (Georgescu et al., 2020a).

Wealth, income and consumption are three major measures of economic inequality. The Gini index is the most extensively used inequality measurement. Corrado Gini (1912) developed Gini index (also called as Gini coefficient) for measuring the distribution of income across citizens. This index is generally used for gauging income distribution or less ordinarily for measuring wealth distribution between citizens, hence, it is treated as of economic inequality. This index values are between 0 (or 0%) and 1 (or 100%) that 0 means excellent equality, while, 1 means excellent inequality. In case of negative wealth or income values, greater than 1 are theoretically attainable. If Gini Index getting higher that means inequality getting higher that high income people receiving bigger shares of the whole income of the population. If in a country every occupant has the same income, then the Gini coefficient of country would be 0. However, a country's Gini coefficient would be 1, if all the income is earned by one occupant and all the others earned nothing. Similarly, comments could be done for wealth, therefore, for wealth Gini coefficient. However, measuring wealth is harder than measuring income. Wealth Gini coefficients tendency are to be much higher than income Gini coefficients. Gini coefficient is not an absolute measurement of a wealth or an income, it is a significant tool to analyze wealth or income distribution in a country or region. If each of in a low-income country and a high-income country the distribution of incomes is similar, then they could have the same Gini coefficient. For example, based on information given by OECD, in 2016 both of U.S. and Turkey's Gini coefficients about 0.39-0.40, however, GDP per capita for Turkey was smaller than half of the U.S.'s (in 2010 dollar terms) (https://www.investopedia.com/terms/g/gini-index.asp#:~:text=The%20Gini%20index%2C%20or%20Gini,wealth%20distribution%20among%20a%20population). The poverty rate is the proportion of the number of people whose income is less than the poverty line that line is taken as half the median household income of the whole population (https://data.oecd.org/inequality/poverty-rate.htm).

Economic inequalities and poverty both directly and indirectly influence people's well-being (Mishchuk et al., 2018; Bilan et al., 2020). Economic inequalities are the most obvious, as they show different positions of people in the distribution of income, wealth, goods, including wages (Stjepanović et al., 2017). Although the basis of the distribution of income is based on the principles of global economic distribution, the issue of economy, welfare at country and individual level is imperative and can be centered on the policy of equal opportunities for all members of a society (Meyer and Meyer, 2016; Belas et al., 2019; Haseeb et al., 2019). If a state government intended for reducing the level of economic inequality (Baltgailis, 2019; Prakash and Garg, 2019), it might apply a few sets of tools: (1) redistribution - from those with high incomes to those with low incomes; (2) the grid of opportunities widely available and (3) social responsibility. Redistribution involves taking from the income of those with higher incomes and supplementing the incomes of those with lower incomes (Androniceanu and Tvaronavičienė, 2019). The opportunity grid is a state policy that consisted of proposing different opportunities to the vulnerable, in difficulty and helpless. The social responsibility of the organizations (Androniceanu, 2019) is constituted by responsible social initiatives of the economic agents in response to the expectations of the customers, investors, employees and of the company as a whole with regard to the business environment focused on the provision of high quality services, admiring the rights of the employees and solving social issues (Çera et al., 2019). In case of the state governments apply these tools that would causes a reduction of economic inequalities, but simultaneously, it shall not affect the economic productivity, the labor market and the encouragements for investments respectively. The main causes of economic inequalities (Smékalová, 2018) are: wage inequalities (when some highly skilled workers earn more than those with lower or no qualifications, including gender wage inequalities), the process of globalization, monopolizing certain market segments, technological changes and computerization, fiscal policy, education, political reforms, labor market failures, wealth growth. There are different conceptual approaches to economic inequalities. Some authors (Duľová Spišáková et al., 2017; Jašková, 2019; Raišienė et al., 2019) believe that economic inequalities are necessary for economic growth. Other authors (Ključnikov and Monika, 2016; Ivancsóné and Printz-Markó, 2018) consider that an analysis of the factors (components) of inequalities is needed to understand why they exist and how their impact on the well-being of the population is reflected (Georgescu et al., 2020a).

Kinnunen et al. (2019) used Bertelsmann Foundation's qualitative data of 124 countries for the year range of 2008-2016 including characteristics

of democracy, governance and market economy. The aim is identifying the most significant political and economic characteristics estimating the level of HDI and the comparison of the classification successes of logistic regression and artificial neural network models (Georgescu et al., 2020a). Georgescu et al. (2020a) applied multinomial logistic regression on a set of 98 word stated grouped in four HDI classes according to four explanatory variables: Net income GINI index, poverty rate, Wealth GINI index and GDP per capita for 2018. The results of the research showed that GDP per capita has the least influence on predicting HDI. Georgescu et al. (2020b) applied multiple discriminant analysis on the same data used in Georgescu et al. (2020a) for predicting four classes of HDI. The discriminant analysis achieved 83.7% accuracy for the discriminant model with unequal a priori probabilities. One conclusion of the model is that GDP per capita has no discriminating power in the model, overall. Another conclusion is that from three discriminant functions retained, the first one discriminates best.

In this study, 100 world countries are classified by using eleven data mining supervised classification algorithms by using WEKA program according to their HDI level (four categories) for the year of 2018. The GDP per capita, poverty rate, Net Income Gini index and Wealth Gini index variables, which are indicators of economic inequality for a country, are used for classifying the countries to their HDI level category. The goal of this study is finding best method has the highest accuracy rate and minimum error.

Organization of the rest part of the study as in following. In Section 2, the most popular eleven data mining classification algorithms are presented. In section 3, classifier performance measures for comparing the methods are presented. The results of the analysis are presented in Section 4. Lastly, conclusion and general comments are given in the last section.

2.MATERIAL METHOD

In this study eleven popular supervised classification methods: Naïve Bayes, IBK, KStar, J48, RandomForest, RandomTree, REPTree, SMO, Simple Logistic, Logistic, Multilayer Perceptron are used. Hence, in this section, these methods are briefly presented.

2.1. Naïve Bayes

The Naive Bayes (NB) classification is established on Bayes' theorem and reports what the probability of occurrence is for the value given for the target class as shown in eqn [1] (Nithi and Priya, 2018; Kazan and Karakoca, 2019):

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)} \quad [1]$$

Here, c is the class to be predicted; x is predicted class; $p(c|x)$ is the probability of event c occurrence when event x appears; $p(c)$ is the probability of event c occurrence; $p(x)$ is the probability of event x occurrence (Kazan and Karakoca, 2019).

Maximum-likelihood training could be done for Naive Bayes classifiers by assessing a closed-form representation that takes linear time, rather than by costly iterative approximation which is the case for many other kinds of classifiers (Nithi and Priya, 2018).

2.2. IBK

The k-nearest neighbor's algorithm (k-NN) is also called as IBK. k-NN is a non-parametric technique used for regression and classification. The output based on if k-NN is used for regression or classification. In case of for classification purpose k-NN is used, at first glance, all observation values are evaluated as a cluster. These clusters are combined gradually and new clusters are obtained. In this method, firstly, the distance between observations is determined. A sample point is classified by a bulk vote of its neighbours, with the sample point is classified to the class most common between its k nearest neighbours (k is a positive integer, normally small). In case of k=1, the sample point is easily classified to the class of that one nearest neighbour. In our study, for k=3 the best classification accuracy rate is obtained (Nithi and Priya, 2018).

2.3. KStar

K star algorithm functions with a generalized supported transformation which classifies as a nearest neighbor technique. It is one of the lazy learning classifications which are especially for cluster analysis approach specially meant for cluster analysis. For prediction, the algorithms use entropic distance measure. Space needed for the storage is incredibly giant as equated to different algorithms. The method of classification with K star incorporates the summation of the possibilities from the new occurrences to all the component of a class (Patil and Shinde, 2020).

2.4. J48

J48 developed by Quinlan (1993) is a C4.5 decision tree that it is developed for the classification process of nonlinear and small size data. The decision tree approach is important in solving classification problems. With

this method, a tree is created to model the classification process. After the tree is created, the classification process is performed by applying it to each data group in the database. Missing values are ignored when creating the tree. Thus, the estimation process is carried out using the remaining data. The key idea in the J48 technique is to classify using the rules generated by the decision trees (Eraldemir et al., 2017).

2.5. Random Forest

The Random Forest technique comprises of the sets of the Regression Tree or Classification Tree suitably with the goal, as much as the number of trees to be produced. Hence, one of the most generally applied algorithms between the ensembles techniques is the Random Forest. The key idea that constitutes the technique is creating ensembles with the help of a randomly chosen subset between a high numbers of estimator trees (Breiman et al., 2017; Kormaz et al., 2018). The forest is created by a number of trees, hence, as the number of trees getting higher the more robustly the forest. The more trees in the forest the more robust the forest resembles (Nithi and Priya, 2018).

Random Forest method can be used in both categorical and continuous data sets; also in small or large data sets. The disadvantage of the technique is not giving an output of a tree, different from the Classification Tree method (Akman et al., 2011; Kormaz et al., 2018).

The advantage of selecting random predictors in this style is that since less correlation is gotten between trees in the ensembles that results in higher accuracy of the model (Suchetana, 2017; Kormaz et al., 2018).

2.6. Random Tree

A random tree is a kind of supervised classification algorithm which is formed by using a stochastic process. It is a kind of community learning algorithm which generates many of discrete learners. The random tree could be used for classification and regression kind issues (Debnath, 2021). The tree produced as a consequence of the Random Tree algorithm is chosen at random from the probable tree cluster that each of tree in the tree cluster has an equal probability of being tried. The distribution of the trees shows a uniform distribution. Random trees could be generated efficiently, and models created by a lot of random trees often have high precision (Georgina et al., 2015; Yavuz et al., 2021).

2.7. RepTree

RepTree utilizes the regression tree mentality and produces multiple trees in dissimilar iterations. Afterwards it chooses best one from whole of created trees (Kalmegh, 2015). REPTree is one of the quick determination tree

classification algorithms. The algorithm utilizes the information gain criterion in the creation of the regression tree or decision and pruning the resulting tree based on the diminished error pruning method. In the REPTree algorithm, solely numerical variables are enumerated (Onan, 2015).

2.8. SMO

Sequential Minimal Optimization is a highly preferred algorithm due to its simplicity and it is used to solve the optimization problems that arise during training. SMO is an algorithm that uses SVM (Support Vector Machine) algorithm. SMO makes choices to solve the smallest possible optimization problems in every single step and produces results (Demirhan ve Hacioglu, 2017).

2.9. Simple Logistic

Simple logistic regression's only difference from linear regression is the dependent variable is not a measurement that it has to be nominal. In simple logistic regression, one target is for the probability of a particular value of the nominal variable being connected with the measurement variable. Another target is predicting the probability of special variables based on the measurement variable. In this method an equation providing the best prediction of the target variable value for each value of the input attribute (Debnath, 2021).

2.10. Logistic

Logistic regression is a kind of classification algorithm used in case of the dependent variable belonging to a specific kind of class. This algorithm predicts the class value of an object based on a probability model. It is a kind of linear regression, however, using the complex cost function i.e., the sigmoid function. The logistic regression hypothesis is prone to appear in the range of 0-1. The obtained equation very identical to the equation obtained in regression analysis. It estimates a logistic model's parameters (Debnath, 2021). Multinomial logistic regression is used for building a trained model for predicting with a ridge estimator (Le Cessie and Van Houwelıngen, 1992). Difference from SimpleLogistic is that Logistic uses a ridge estimator. For further information, Le Cessie and Van Houwelıngen (1992) could be examined.

2.11. Multilayer Perceptron

Artificial neural networks (ANN) are information processing systems that generally imitate the working rules of the central nervous system or human brain. Studies on this subject first started with the modeling of neurons,

the biological units that make up the brain, and its application in computer systems. Neurons are interlinked by links, and each link has a numerical weight that states the strength or significance of its input. Weights are the main tool of long-term store in ANNs. A neural network conducts learning by adjusting these weights repeatedly. ANNs are generally divided into two as singlelayer perceptron and multi multilayer perceptron (Arı and Berberler, 2017).

Multilayer perceptron (MLP) comprises the input layer where information is entered, one or more hidden (intermediate) layers and an output layer (Arı and Berberler, 2017). Each layer contains many neuron cells. These cells are linked to each other by weighted links. In the MLP, there are transitions called forward and backward propagation between layers. In the forward propagation state, the error value and the output of the network are computed. In the back propagation state, the linking weight values among the layers are updated to minimize the computed error value (Arı and Berberler, 2017; Depren et al. 2017; Kazan and Karakoca, 2019).

MLP network converts n input vector into an output vector by performing nonlinear operations on it. The output of the network is determined by the output layer, which has an activation function. The difference between the calculated output value and the target value is defined as the mean square error function. Training of the MLP network is a process expressed as the minimization of this defined error function. In this process, the weighted connections between neurons are optimized. Optimization is performed with Gradient Descent and Backpropogation algorithms. MLP type classifiers can be easily adapted to multiple classification problems (Kazan and Karakoca, 2019).

3.CLASSIFIER PERFORMANCE MEASURES

The efficiency of classification methods is found by measures for instance true positive rate, false positive rate, false negative rate and true negative rate. The classification model is derived from two classes: Predicted and actual class. Confusion matrix Contingency is defined as in Table 1 (Sujatha et al., 2017).

Confusion matrix

Table 1

Actual Class	**Predicted Class**	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

The number of accurately classified objects is the sum of diagonals in the matrix and the rest of observations are incorrectly assigned to classes.

Accuracy

The most favorite and easiest technique used for measuring model performance is the accuracy rate of the model. It is the proportion of the number of correctly classified objects (TP + TN) to the whole number of objects (TP + TN + FP + FN) (Çoşkun and Baykal, 2011; Rajagopalan, 2017; Suchetana et al., 2017; Sujatha and Bilgin, 2017)

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision/ Specificity

Precision, also called as Specificity, is the proportion of the number of TP objects estimated to be class Positive to the total number of objects estimated to be class Positive (Çoşkun and Baykal, 2011; Debnath, 2021).

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall/ Sensitivity

If TPR shows the True Positive Rate that is the fraction of objects classified as class A amongst whole objects which actually have class A (Sujatha et al., 2017). Recall is the proportion of the correctly estimated positive values to the positive estimated values. It is as well called as Sensitivity (Debnath, 2021). A higher recall value indicates the model returns most of the relevant data (Georgina et al., 2015).

$$\text{Sensitivity / Recall} = \frac{TP}{TP + FN}$$

F- Measure

The precision and recall criteria solely are not sufficient to draw a substantial comparison result. Assessing both criteria together produces more correct results. Hence, the F-criterion has been defined. F-criterion is the harmonic mean of recall and precision (Çoşkun and Baykal, 2011; Sujatha and Rajagopalan, 2017; Çavuşoğlu and Kaçar, 2019; Debnath, 2021).

$$F - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Kappa Statistics

The Kappa Statistic is the most usually used statistic for test interrater consistency. A kappa statistic value of 1 shows excellent agreement (Debnath, 2021). Kappa is a measure used to quantitatively state the correspondence between observed and predicted classifications in a data set. Kappa coefficient changes between -1 and +1. -1 shows that there is an incompatibility or a relationship in the opposite direction. 1 shows perfect correspondence. The interpretable range of the Kappa coefficient is between 0 and +1, and negative kappa values have no significance in terms of reliability (Bilgin, 2017). If kappa value is 0.4 or above, it could be treated as that there is an adequate correspondence beyond chance (Demirhan and Hacioglu, 2017). Despite the fact that there is no definitive standard explanation of the Kappa statistic, commonly, the values between 0.00-0.20 are treated as low, 0.21-0.40 as notable; 0.41-0.60 as moderate; 0.61-0.80 as significant and 0.81-1.00 as perfect (Landis and Koch, 1977; Aksu and Doğan, 2019).

$$\text{Kappa} = \frac{\left(\text{Observed Accuracy} - \text{Expected Accuracy}\right)}{\left(1 - \text{Expected Accuracy}\right)}$$

ROC Area

ROC analysis is another important analysis used in performance measurement of the classification process performed on the data set. In this analysis, the correct estimation rate of the classification process is examined using a different approach. Two concepts expressed as Sensitivity and Specificity are used in obtaining the ROC curve. The calculated Sensitivity value on the Y axis, the calculated (1- Specificity) on the X axis. ROC curve is obtained by the combination of the points obtained by the intersection of these two values (Çavuşoğlu and Kaçar, 2019).

The value obtained from the ROC analysis shows the size of the area under the ROC curve. The ROC field curve shows the predictive success of the diverse classification algorithms. Area under the ROC curve is one of the fundamental assessment criteria used for choosing the best classification algorithm. If the area under the curve approaches 1, it shows that the classification is done accurately (Çığşar and Ünal, 2019). In the ROC test, 1 indicates the best value and 0.5 indicates an unsuccessful classification. When the ROC value is 1, it is understood that no wrong estimates are made on the data set, and this ROC curve draws a line combining the points (0,0), (1,0) and (1,1) on the coordinate plane (Çavuşoğlu and Kaçar, 2019).

RMSE

The square root of the mean squared error (MSE) gives root mean squared error deviation. Generally, the RMSE is calculated as a measure of the difference between the actual values and the predicted values of an estimator or model. It also means, the RMSE indicates the standard deviation of the difference between the observed values and predicted values. The model with smallest RMSE value is preferable (Çığşar and Ünal, 2019).

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

4.RESULTS OF APPLICATION

In this paper we use data for 100 world countries from World Economic Forum from 2018. The dependent variable is HDI with its four categories: low HDI (lower than 0.550), medium HDI (0.550-0.699), high HDI (0.7-0.799) and very high HDI (0.8-1). Out of 100 countries, 47 fall into very high category (47%), 25 fall into high category (25%), 14 fall into medium category (14%) and 14 into low category (14%). The explanatory variables are Net income GINI index, GDP per capita, poverty rate and Wealth GINI index Georgescu et al. (2020a; 2020b) also used the same data set, however, in their analysis there are 98 countries. They have applied Multiple Discriminant Analysis and Multinomial Logistic Regression on the same data set. Here, 100 countries are used in analysis and the explanatory variables are obtained from World Economic Forum from 2018 and the HDI values are obtained from Human Development Data Center (http://hdr.undp.org/en/data). First of all, the mean and standard deviations of the independent variables for each of the HDI levels is given in Table 2.

The mean and standard deviations for independent variables for each of HDI Levels

Table 2

		GDP per capita US$	Net Income Gini index	Wealth inequality Gini	Poverty Rate
Very High HDI	Avg.	32632.53	32.69	68.66	8.05
	Std. dev.	23651.14	5.96	11.47	5.59
High HDI	Avg.	4908.79	41.78	73.24	11.37
	Std. dev.	2334.50	7.67	12.06	10.16
Medium HDI	Avg.	1840.71	41.96	70.35	41.71
	Std. dev.	1424.59	5.62	11.74	19.23
Low HDI	Avg.	870.78	37.57	67.06	75.62
	Std. dev.	566.83	5.31	6.98	17.12

47% of the countries fall into "Very High HDI" class with the highest average GDP per capita of USD 32633. Very high class has the lowest income Gini and lowest poverty rate of four HDI levels and it has the lowest wealth Gini, on average, following the HDI low class. Moreover, for GDP per capita variable's variance is getting larger for higher levels of HDI. 25% of the countries fall into "High HDI" class with the second highest average GDP per capita of USD 4909, the income is considerably lower than the top class. Moreover, the highest wealth Gini belongs to this class and it has the second highest income Gini close to medium class, however, poverty rate is almost as low as in the very high class. 14% of the countries fall into "Medium HDI" class with GDP per capita of USD of 1841. This category has an income less than half of the high level HDI class. Income Gini is nearly same level with high level and it is the second category close to high category in terms of highest wealth Gini, however, in comparison with high level, the poverty rate averagely is much higher with a large variance. 14% of the countries fall into "Low HDI" class and this category has the lowest income of 871 USD. The income in this category, averagely, is less than half of the medium class countries. Wealth Gini is the lowest and income Gini is the second lowest of the HDI levels (classes), the poverty rate, averagely, is very high of 76% in HDI low level countries.

We have used 11 data mining algorithms. The analyses are implemented in Waikato Environment for Knowledge Analysis (WEKA). University of New Zealand was developed WEKA that it is a data-mining application software. It is an open source software program created in Java under General Public License. It includes various supervised and unsupervised techniques like classification, data visualization, clustering and association. In this data application, the WEKA 3.8.5 version is used for classifying the countries with various supervised classification algorithms (Çığşar and Ünal, 2019).

In experimental studies, usually the original data set is randomly partitioned into 10 equal pieces by the 10-fold cross validation technique. Then, one of these pieces is kept as validation data for testing the model, while the rest of nine pieces are used as training data. The cross validation process is conducted 10 times, allowing each of the 10 pieces to be used as validation data once (Onan, 2015). Witten and Frank (2005) mentions that, stratified 10-fold cross-validation, that is the common evaluation method in cases where merely limited data is available and it is considered as the strictest one (Ganganagowder and Kamath, 2017). The following methods are also applied with 10-fold cross validation. The classification performance results for all eleven classification algorithms are given in Table 3.

The classification performance results for eleven supervised classification algorithms for HDI data

Table 3

Classification	Accuracy	Precision	Recall	RMSE	ROC Area	F-Measure	Kappa Statistic
Naïve Bayes	81.00%	81.80%	81.00%	0.2626	0.954	0.811	0.7213
IBK	***72.00%***	***72.70%***	***72.00%***	***0.3072***	***0.882***	***0.719***	***0.5805***
KStar	76.00%	76.50%	76.00%	0. 3252	0.937	0.762	0.6469
J48	83.00%	83.60%	83.00%	0.2836	0.898	0.833	0.7511
RandomForest	85.00%	85.50%	85.00%	0.2442	0.951	0.851	0.7786
RandomTree	79.00%	79.70%	79.00%	0.324	0.851	0.788	0.6895
REPTree	78.00%	78.80%	78.00%	0.2903	0.902	0.778	0.6773
SMO	74.00%	76.00%	74.00%	0.352	0.834	0.733	0.6040
Simple Logistic	85.00%	85.50%	85.00%	0.2472	0.962	0.849	0.7778
Logistic	86.00%	86.10%	86.00%	0.2495	0.939	0.860	0.7937
Multilayer Perceptron	***88.00%***	***87.60%***	***88.00%***	***0.2151***	***0.970***	***0.877***	***0.8211***

For the HDI dataset, looking the accuracy proportions in the Table 3, it is obvious that the Multilayer Perceptron accomplished the highest accuracy fraction of 88.00%. The precision and recall rates of the Multilayer Perceptron method are 87.60% and 88.00%, respectively. In terms of all classifier performance measures Multilayer Perceptron is the leading one such as Kappa statistic is 0.8211 is an as excellent level, ROC-Area is 0.970 is closest to 1 and F-measure is the highest one. It has also the smallest RMSE value of 0.2151. The Logistic method is the second best technique after the Multilayer Perceptron. The Logistic technique succeeded 86.00% accuracy rate, 86.10% precision rate and 86.00% recall rate. However, IBK has the lowest accuracy rate of 72.00%, precision rate of 72.70%, recall rate of 72.00% and Kappa Statistic value of 0.5805.

For finding the subset of components producing the best prediction and classification performance, a process is done by ordering components according to their discriminative power. WEKA contains various attitude-selection techniques. Attribute selection is a technique used to extract the ranking of attributes. ClassifierAttributeEval technique that assesses the worth of an attribute using a user-specified classifier. Here, since for our data set Multilayer Perceptron has the best performance for this method the importance of attributes (means explanatory variables) in classification is determined. The results for this attribute selection method is given in Table 4. From Table 4 it is clear that all of 4 explanatory variables are important for classifying the countries in terms of HDI Levels. Among them, GDP per capita US$ is found as the most effective variable on determining the HDI levels of countries.

WEKA output of ranking of attributes with respect to ClassifierAttributeEval evaluation method

Table 4

Ranked attributes:
0.286 1 GDP per capita US$
0.216 3 Poverty Rate
0.074 2 Net Income Gini index
-0.008 4 Wealth inequality Gini
Selected attributes: 1,3,2,4 : 4

4. RESULTS

The predictive power of economic inequalities (measured by the wealth and income gini) with respect to the United Nations' human development index, HDI, has been studied together with the gross domestic product (GDP per capita) in 100 world countries. As a result of analysis we conclude that the Multilayer Perceptron is the best one with an accuracy rate of 88.00% compared with the other data mining classifiers. Georgescu et al. (2020a; 2020b) have studied on 98 countries for the year 2018 and they have applied classical classification methods like multinomial logistic regression and multiple discriminant analysis on training data set. For discriminant analysis they have found that GDP per capita is not an important variable in classification. However, in this study 100 countries are evaluated for the same year of 2018 and eleven data mining classifier methods are applied on the data set by using cross-validation method. Since there is not a test set, the performances of methods are compared by using cross-validation technique. As a result Multilayer Perceptron is found to be the best one in terms of accuracy rate and other classification measures. Moreover, there is an attribute selection technique in WEKA that we obtained as a result, all of explanatory variables: Net income GINI index, poverty rate, Wealth GINI index and GDP per capita are important in classification. It is concluded that these 4 variables showing the economic inequalities are effective in determining the HDI levels of countries. Moreover, unlike Georgescu et al. (2020a; 2020b) studies, GDP per capita is the most significant variable in classifying the countries according to their HDI levels.

REFERENCES

1. **Akman, M., Genç, Y**., **Ankaralı, H.**, 2011, *"Random Forests yöntemi ve sağlık alanında bir uygulama"*, Türkiye Klinikleri Journal of Biostatistics, 3 (1), 36-48.
2. **Aksu, G.**, **Doğan, N.**, 2019, *"An Analysis Program Used in Data Mining: WEKA"*, Journal of Measurement and Evaluation in Education and Psychology, 10(1), 80-95.
3. **Andrei, A.M.**, **Georgescu, I.**, 2018, *"Socio-economic inequality and economic growth: measurements for central and eastern Europe"*, Chinese Business Review, 17(11), 547-570.
4. **Androniceanu, A.**, **Tvaronavičienė, M.**, 2019, *"Developing a holistic system for social assistance services based on effective and sustainable partnerships"*, Administratie si Management Public, 33, 103-118.
5. **Androniceanu, A.**, 2019, *"Social responsibility, an essential strategic option for a sustainable development in the field of bio-economy"*, Amfiteatru Economic, 21 (52), 347- 364.
6. **Arı, A.**, **Berberler, M. E.**, 2017, *"Yapay Sinir Ağları ile tahmin ve sınıflandırma problemlerinin çözümü için arayüz tasarımı"*, Acta Infologıca, 1(2), 55-73.
7. **Baltgailis, J.**, 2019, *"The issues of increasing the effectiveness of teaching comparative economics"*, Insights Into Regional Development, 1(3), 190-199.
8. **Belas, J.**, **Belas, L.**, **Cepel, M.**, **Rozsa, Z.**, 2019, *"The Impact of the public sector on the quality of the business environment in the SME segment"*, Administratie si Management Public, 32, 18-31.
9. **Bilan, Y.**, **Mishchuk, H.**, **Samoliuk, N.**, **Yurchyk, H.**, 2020, *"Impact of income distribution on social and economic well-being of the state"*, Sustainability, 12(1), 429.
10. **Bilgin, M.**, 2017, "Gerçek veri setlerinde klasik makine öğrenmesi yöntemlerinin performans analizi". 19. Akademik Bilişim Konferansı-AB 2017, Aksaray University, Aksaray, Turkey, 8-10 February, 1-6.
11. **Breiman, L**., **Friedman, J. H.**, **Olshen, R. A.**, **Stone, C. J.**, 2017, *"Classification and Regression Trees"*, Taylor Francis, Berkeley, California, 368.
12. **Çavuşoğlu, Ü.**, **Kaçar, S.**, 2019, *"Anormal trafik tespiti için veri madenciliği algoritmalarının performans analizi"*, Academic Platform Journal of Engineering and Science, 7-2, 205-216.
13. **Çera, G.**, **Meço, M.**, **Çera, E**., **Maloku, S.**, 2019, *"The effect of institutional constraints and business network on trust in government: an institutional perspective"*, Administratie si Management Public, 33, 6-19.
14. **Çığşar, B.**, **Ünal, D.**, 2019, *"Comparison of Data Mining Classification Algorithms Determining the Default Risk"*, Scientific Programming Volume 2019, Article ID 8706505, 1-8.
15. **Çoşkun, C.**, **Baykal, A.**, 2011, *"Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması"*, Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri, 2 – 4 February 2011, İnönü University, Malatya, Turkey.
16. **Debnath, P. et al.**, 2021, *"Analysis of Earthquake Forecasting in India Using Supervised Machine Learning Classifiers"*, Sustainability, 13(2), 971.
17. **Demirhan, T.**, **Hacioglu, I.**, 2017, *"Performance and achievement analysis of a data set of distance education samples with Weka"*, The Eurasia Proceedings of Educational & Social Sciences (EPESS), 8, 9-18.
18. **Depren**, **S.K.**, **Aşkın, Ö.E.**, **Öz, E.**, 2017, *"Identifying the Classification Performances of Educational Data Mining Methods: A Case Study for TIMSS"*, Uram ve Uygulamada Eğitim Bilimleri Educational Sciences: Theory & Practic, 17(5), 1605–1623.

19. **Duľová Spišáková, E.**, **Mura, L.**, **Gontkovičová, B.**, **Hajduová, Z.**, 2017, *"R&D in the context of Europe 2020 in selected countries"*, Economic Computation and Economic Cybernetics Studies and Research, 51(4), 243-261.
20. **Eraldemir, S. G.**, **Arslan, M. T.**, **Yıldırım, E.**, 2017, *"Comparison of Random Forest and J48 Decision Tree classifiers using HHT based features in EEG"*, International Advanced Researches & Engineering Congress-2017, 16-18 November 2017, 1250-1256. *http://iarec.osmaniye.edu.tr/ Osmaniye/TURKEY*.
21. **Gallardo, G.**, 2009, *"The Human Development Index as an effort to measure well-being in Honduras"*, The 3rd OECD World Forum on "Statistics, Knowledge and Policy" Charting Progress, Building Visions, Improving Life Busan, Korea - 27-30 October, 1-13.
22. **Ganganagowder, N. V.**, **Kamath, P.**, 2017, *"Intelligent classification models for food products basis on morphological"*, Colour and Texture Features. Acta Agron., 66 (4), 486-494.
23. **Georgescu, I.**, **Kinnunen, J.**, 2019, *"Well-being and economic freedoms in OECD"*, Proceedings of the 12th LUMEN International Scientific Conference Rethinking Social Action: Core Values in Practice, 15-17 May 2019, Iasi, Romania, 108-125.
24. **Georgescu, I.**, **Kinnunen, J.**, **Androniceanu, A.**, **Androniceanu, A. M.**, 2020a, *"Global well-being and economic inequality"*, Paper presented at the 19th International Conference on Information in Economy (IE 2020), Timisoara, Romania, 21-24 May 2020, 266-273.
25. **Georgescu, I.**, **Androniceanu, A.**, **Kinnunen, J.**, 2020b, *"A discriminant analysis to the quantification of Human Development Index under economic inequality"*, Proceedings of the 14th International Management Conference "Management Sustainable Organizations" 5th–6th November, Bucharest, Romania, 1053-1062.
26. **Georgina N., O.**, **Isah, A.**, **Alhasan, J.**, 2015, *"Analytical Study of Some Selected Classification algorithms in WEKA using real crime data"*, (IJARAI) International Journal of Advanced Research in Artificial Intelligence, 4(12), 44-48.
27. **Haseeb, M.**, **Kot, S.**, **Hussain, H.I.**, **Jermsittiparsert, K.**, 2019, *"Impact of economic growth, environmental pollution, and energy consumption on health expenditure and R&D expenditure of ASEAN countries"*, Energies, 12 (3598), 1-20.
28. **Ivancsóné, H. Z.**, **Printz-Markó, E.**, 2018, *"Territorial differences between countries with regard to the wellness lifestyle of their youth"*, Forum Scientiae Oeconomia, 6(3), 101-117.
29. **Jašková, D.**, 2019, *"Assessment of social development in Slovakia in the context of human resources"*, Central European Journal of Labour Law and Personnel Management, 2(2), 21-32.
30. **Kalmegh, S.**, 2015, *"Analysis of WEKA data mining algorithm REPTree, Simple Cart and RandomTree for classification of Indian news"*, IJISET - International Journal of Innovative Science, Engineering & Technology, 2(2), 438-446.
31. **Kazan, S.**, **Karakoca, H.**, 2019, *"Makine öğrenmesi ile ürün kategorisi sınıflandırma"*, Sakarya University Journal of Computer and Information Sciences, 2(1), 18-27.
32. **Kinnunen, J.**, **Androniceanu, A.**, **Georgescu, I.**, 2019, *"The role of economic and political features in classification of countries-in-transition by Human Development Index"*, Informatica Economica, 23(4), 26-40.

33. **Ključnikov, A.**, **Monika, S.M.**, 2016, *"Impact of gender in the perception of administrative burdens among young entrepreneurs - evidence from Slovakia"*, Journal of Competitiveness, 8(2), 17-30.
34. **Kormaz, D.**, **Çelik, H., E.**, **Kapar, M.**, 2018, *"Botnet detection by using classification and Regression Trees with Random Forest algorithms: example of Van Yüzüncü Yıl University"*, Journal of the Institute of Natural & Applied Sciences, 23(3), 297-307.
35. **Landis, J.R.**, **Koch, G.G.**, 1977, *"The Measurement of Observer Agreement for Categorical Data"*, Biometrics, 33, 159-174.
36. **Le Cessie, S.**, **Van Houwelıngen, J.C.**, 1992, *"Ridge estimators in logistic regression"*, Applied Statistics, 41(1), 191-201.
37. **Meyer, N., Meyer, D.F.,** 2016, *"The relationship between the creation of an enabling environment and economic development: A comparative analysis of management at local government sphere"*, Polish Journal of Management Studies, 14(2), 150-160.
38. **Mishchuk, H., Samoliuk, N., Bilan, Y., Streimikiene, D.**, 2018, *"Income inequality and its consequences within the framework of social justice"*, Problemy Ekorozwoju, 13(2), 131-138.
39. **Nithi, M., Priya, J.**, 2018, *"Performance analysis of different classification methods in data mining"*, International Conference on Advancements in Computing Technologies - ICACT 2018, ISSN: 2454-4248, Chennai, India, February 2-3, 55-58.
40. **Onan, A.,** 2015, *"Şirket iflaslarının tahmin edilmesinde Karar Ağacı algoritmalarının karşılaştırmalı başarım analizi"*, Bilişim Teknolojileri Dergisi, 8(1), 9-19.
41. **Patil, P., Shinde, Dr. Prof. S.,** 2020, *"Performance analysis of different classification algorithms: Naïve Bayes, Decision Tree and K-Star"*, Journal of Critical Reviews, 7(19), 1160-1164.
42. **Prakash, R., Garg, P.**, 2019, *"Comparative Assessment of HDI with Composite Development Index (CDI)"*, Insights into Regional Development, 1(1), 58-76.
43. **Raišienė, A.G.**, **Bilan, S., Smalskys, V., Gečienė, J.**, 2019, *"Emerging changes in attitudes to inter-institutional collaboration: the case of organizations providing social services in communities"*, Administratie si Management Public, 33, 34-56.
44. **Smékalová, L.**, 2018, *"Evaluating the cohesion policy: targeting of disadvantaged municipalities"*, Administratie si Management Public, 31:143-154.
45. **Stjepanović, S., Tomić, D., Škare, M.**, 2017, *"A new approach to measuring green GDP: A cross-country analysis"*, Entrepreneurship and Sustainability Issues, 4(4), 574- 590.
46. **Suchetana, B., Rajagopalan, B., Silverstein, J.**, 2017, *"Assessment of wastewater treatment facility compliance with decreasing ammonia discharge limits using a Regression Tree model"*, Science of the Total Environment, 598, 249-257.
47. **Sujatha, J., Rajagopalan, Dr. S.P.**, 2017, *"Performance evaluation of machine learning algorithms in the classification of parkinson disease using voice attributes"*, International Journal of Applied Engineering Research, 12(21), 10669-10675.
48. **Witten, I.H., Frank, E.**, 2005, *"Data Mining Practical Machine Learning Tools and Techniques"*, editor Morgan Kaufmann, San Francisco, 560.
49. **Yavuz, A.A., Ergül, B., Aşık, E.G.,** 2021, *"Trafik Kazalarının Makine Öğrenmesi Yöntemleri Kullanılarak Değerlendirilmesi"*, Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi (International Journal of Engineering Research and Development, 13(1), 66-73.
50. *http://hdr.undp.org/en/data*. Access date: 01.04.2021.

51. *https://www.investopedia.com/terms/g/gini-index.asp#:~:text=The%20Gini%20index%2C%20or%20Gini,wealth%20distribution%20among%20a%20population*. Access date: 01.04.2021.
52. *https://data.oecd.org/inequality/poverty-rate.htm*. Access date: 01.04.2021.