{MERAA} Mathematics in Education, Research and Applications
ISSN 2453-6881
Math Educ Res Appl, 2021(7), 2

Received: 2021-10-21
Accepted: 2021-12-09
Online published: 2021-12-31
DOI: https://doi.org/10.15414/meraa.2021.07.02.73-80

Original Paper

Classification model of poverty risk in the European Union

Janka Drábeková*

Slovak University of Agriculture in Nitra, Faculty of Economics and Management, Institute of Statistics, Operation Research and Mathematics, Slovak Republic

ABSTRACT

Analysis of the at-risk-of-poverty dataset using WEKA machine learning software tool aims for mining the relationship in selected data from database Eurostat for efficient classification. We used eight classification algorithms for analyzing dataset. We used WEKA tools to search the best classification algorithm. We evaluated accuracy of classification algorithms using various accuracy measures like Kappa statistic, TP rate, FP rate, Precision, Recall, F-measure, ROC Area and PRC Area. The accuracy of the models was monitored by the number of instances classified correctly. In this paper we describe the values of the monitored indicators of the best algorithm J48.

KEYWORDS: Data Mining, Classification, WEKA, at-risk-of-poverty, EU countries

JEL CLASSIFICATION: C38, C88, I32

INTRODUCTION

Poverty and income inequality are a highly topical issue, not least because of the covid19 pandemic we are currently experiencing. This issue is not only important in developing countries, but reducing income inequality and reducing poverty are important goals for the Member States of the European Union. Monitoring the development of poverty levels is important in determining the socio-economic progress of society [12]. Eurostat publications state that one-fifth or more of the population was at risk of poverty in up to 7 EU countries in 2018 [2].

Poverty and social exclusion are multidimensional phenomena. Just as there is no only one or correct definition of poverty, there is no single generally accepted way of measuring it [11]. The at-risk-of-poverty line is set at 60% of the median national equivalent disposable income

* Corresponding author: Janka Drábeková, Slovak University of Agriculture in Nitra, Faculty of Economics and Management, Institute of Statistics, Operation Research and Mathematics, Tr. A. Hlinku 2, 949 76 Nitra, Slovak Republic, E-mail: janka.drabekova@uniag.sk

and is expressed in PKS (purchasing power parity). The foundation for comparing living standards between countries is often gross domestic product (GDP) per capita, which in monetary terms shows the basic measure of the total size of the economy divided by the number of people living in it and is used to measure a country's wealth and prosperity. However, this headline indicator does not provide information on the distribution of income within a country, nor does it provide information on non-monetary factors that can play an important role in determining the quality of living conditions of the population [2].

Long-term observations of income inequality and poverty show that countries with higher income inequality are most likely countries with high levels of poverty and countries with low income inequality, as well as countries with low at-risk-of-poverty. Janovičová & Bartová assessed the development of income inequality, the poverty risk rate in V4 countries over the years 2005-2017 using the panel of annual data and by econometric models [5]. They found that in Poland and Hungary, the at the risk of poverty rate was significantly higher than in Czech and Slovakia in the observed period. Carlsen and Bruggemann [1] studied the inequality within the 27 European Member States by partial ordering methodology multi-indicator system. They found that Luxembourg, The Netherlands, Austria, and Finland had rather low inequality and on the other hand Bulgaria and Romania was with the highest degree of inequality in the period under review. They also found that Luxembourg and Hungary were isolated countries, i.e., incomparable to any other EU Member State. Muster [7] based on Eurostat research (the EU-SILC survey) presented the dynamics of changes in the phenomenon of in-work poverty in individual EU countries in 2006-2019 in his work. He said that a particularly significant increase in poverty in 2006-2019 was observed in Bulgaria, Germany, Hungary, Malta and the Netherlands. Between the factors that have a key impact on the problem of impoverishment of the economically active he included low level of education, flexible work, part-time work, young age, low work experience and living in multiperson households. Janovičová stated that proportion of population aged 65 years and more, unemployment rate and people aged 18-59 living in jobless household have statistically significant positive effect on income inequality and at the risk of poverty rate growth [4]. She assessed development of income inequality, poverty risk rate in the 19 EU Member States over the years 2005-2017.

Accurate data on poverty prevalence are needed by policymakers in anti-poverty policies [12]. Žilinský et al. [12] in their study argue that subjective poverty indicators provide essential information and should be taken into account as a supplementary dimension for assessments of the poverty level in a society. They found that with the exception of a few countries, all three subjective poverty indices (headcount ratio, the poverty gap index, and the severity of poverty index) show consistent decreasing trends in subjective poverty of EU Member States. Their results suggest that objective poverty measures should be considering housing costs because social subjective poverty lines are considerably higher for households paying mortgages and tenants paying rent than for outright homeowners.

Ivanová and Grmanová [3] studied the sustainability of EU labor immigration in terms of poverty inequalities and employment. They argue in their study that immigrants coming out of the EU are significantly at higher risk of poverty because in most EU countries, the employment rate in the group “nationals” is lower than in the group “foreign” from the EU. Tkachova et al. [10] in their study determined that the policy of integration of immigrants does not ensure the achievement of the goal of inclusive and equitable social-economic welfare. Next a particularly vulnerable group in terms of the risk of poverty are the

unemployed. With almost half (48.6%) of all unemployed in the EU27 being at risk of poverty in 2018, with the undisputed highest rate recorded in Germany (69.4%). Another 11 EU Member States (Lithuania, Malta, Latvia, Sweden, Bulgaria, Hungary, the Czech Republic, Estonia, Slovakia, Spain and Belgium) reported that at least half of the unemployed were at risk of poverty in 2018 [2].

The Waikato Environment for Knowledge Analysis (WEKA) allow easy access to state-of the-art techniques in machine learning for researchers [9]. This software for analyzing data contains many machine leaning algorithms [8]. It is providing large number of different classifiers that are used in data mining task and analyze the output produced by these classifiers [6]. In this article we focused on classification as one of the data mining technique appropriate to extract patterns from data.

MATERIAL AND METHODS

We obtained the data in a secondary way from the international database Eurostat. We compiled a dataset consisting of 28 EU countries: Austria, Austria, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, the Republic of Poland, Portugal, Romania, the Slovak Republic, Slovenia, Spain, Sweden and United Kingdom (Table 1).

Table 1 Values of indicators forming the research dataset before its modification

Relation: udajeCSV-weka.filters.unsupervised.attribute.Remove-R4-6,16-17

No.	1: country	2: RiskRP	3: JoblessHch	4: UnemplT	5: TGGexp	6: GDPpc	7: PopTerEdu	8: PopSecEdu	9: PopPrimEdu	10: Pop65viac	11: S80/S20	12: **Gini**
	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	Belgium	14.8	10.9	5.4	52.1	35940.0	36.0	38.1	25.9	18.9	3.61	25.1
2	Bulgaria	22.6	9.3	4.2	36.3	6840.0	24.7	53.5	21.9	21.3	8.1	40.8
3	Czechia	10.1	5.5	2.0	41.3	18330.0	21.6	66.1	12.3	19.6	3.34	24.0
4	Denma...	12.5	7.5	5.0	49.2	49720.0	33.1	40.8	26.1	19.6	4.9	27.5
5	Germa...	14.8	8.4	3.1	45.2	35840.0	26.0	54.5	19.5	21.5	4.89	29.7
6	Estonia	21.7	7.3	4.4	38.9	15760.0	36.5	47.7	15.8	19.8	5.8	30.5
7	Ireland	13.1	10.9	5.0	24.5	60170.0	40.7	38.3	21.0	14.1	4.3	28.3
8	Greece	17.9	8.3	17.3	47.5	17740.0	27.8	46.3	25.9	22.0	5.11	31.0
9	Spain	20.7	8.3	14.1	42.1	25200.0	35.1	25.3	39.6	19.4	5.94	33.0
10	France	13.6	11.5	8.5	55.6	33270.0	33.7	42.9	23.4	20.1	4.27	29.2
11	Croatia	18.3	5.7	6.6	47.0	12450.0	22.0	59.9	18.1	20.6	4.76	29.2
12	Italy	20.1	9.4	10.0	48.6	26920.0	17.4	42.8	39.8	22.8	6.1	32.8
13	Cyprus	14.7	6.0	7.1	40.1	24570.0	40.0	38.5	21.5	16.1	4.58	31.1
14	Latvia	22.9	7.6	6.3	38.4	12510.0	31.4	53.8	14.9	20.3	6.54	35.2
15	Lithuania	20.6	8.5	6.3	34.6	14010.0	37.9	51.0	11.1	19.8	6.44	35.4
16	Luxem...	17.5	6.2	5.6	42.2	83640.0	41.0	32.3	26.7	14.4	5.34	32.3
17	Hungary	12.3	5.6	3.4	45.6	13260.0	22.5	57.6	20.0	19.3	4.23	28.0
18	Malta	17.1	7.0	3.6	37.2	21800.0	26.7	32.1	41.3	18.7	4.18	28.0
19	Netherl...	13.2	5.3	3.4	42.0	41870.0	34.8	39.7	25.5	19.2	3.94	26.8
20	Austria	13.3	6.2	4.5	48.4	38170.0	31.1	50.2	18.7	18.8	4.17	27.5
21	Poland	15.4	8.3	3.3	41.8	13000.0	28.2	58.5	13.3	17.7	4.37	28.5
22	Portugal	17.2	4.5	6.5	42.7	18590.0	23.8	28.7	47.6	21.8	5.16	31.9
23	Romania	23.8	7.3	3.9	36.1	9120.0	16.0	58.9	25.1	18.5	7.8	34.8
24	Slovenia	12.0	2.6	4.5	43.3	20700.0	29.3	54.9	15.8	19.8	3.39	23.9
25	Slovakia	11.9	7.5	5.8	42.7	15860.0	23.1	62.3	14.5	16.0	3.34	22.8
26	Finland	11.6	4.8	6.7	53.3	37170.0	38.5	44.6	16.9	21.8	3.69	26.2
27	Sweden	17.1	10.5	6.8	49.4	43900.0	37.8	41.5	20.8	19.9	4.33	27.6
28	United...	19.0	10.5	3.8	41.1	32910.0	40.6	40.2	19.1	18.4	5.63	33.5

Source: data Eurostat, author processing, output from WEKA

We used 11 numeric attributes for the analyzes (Table 1): Unemployment Rate (UnemplT), Children aged 0-17 years living in jobless households (JoblessHch), Total general government expenditure (TGGexp), Gross domestic product per capita (GDPpc), Population by educational attainment level, tertiary, levels 5-8 (PopTerEdu), Population by educational

attainment level, Upper secondary, post-secondary non-tertiary, levels 3-4 (PodSecEdu), Population by educational attainment level, Less than primary, primary and lower secondary education, levels 0-2 (PopPrimEdu), Proportion of population aged 65 years and more (Pop65viac), Income quintile share ratio (S80/S20), Gini coefficient of equivalised disposable income (Gini).

There was no missing data in the dataset, it was adjusted by discretizing the variables needed for classification methods (Table 2). As a classification attribute, we set the indicator - At the risk of poverty rate (RiskRP). We discretized the classification attribute to three nominal categories (Table 2).

We selected the relevant data on the basis of selection using the values of correlation coefficients expressing the relationship between individual attributes and the classification attribute. Based on the obtained values of correlation coefficients $(r > 0.2)$, we can conclude that there is a relationship between the at-risk-of-poverty rate and 6 attributes: Income quintile share ratio S80/S20 $(r = 0.586)$, Gini coefficient of equivalised disposable income $(r = 0.388)$, Gross domestic product per capita $(r = 0.24)$, Population by educational attainment level - upper secondary, post-secondary non-tertiary, levels 3-4 $(r = 0.21)$, Population by educational attainment level - less than primary, primary and lower secondary education, levels 0-2 $(r = 0.209)$, Total general government expenditure $(r = 0.203)$. In the following analyzes, we will use only 6 of the listed attributes.

Table 2 Discretization of selected attributes

Attribute	*Label*	*Variation*	*Count*
Gini coefficient of equivalised disposable income	$(-\infty, 28.8\rangle$	below average EU	13
	$(28.8, 34.8\rangle$	average EU	12
	$(34.8, \infty)$	above average EU	3
Income quintile share ratio S80/S20	$(-\infty, 5.72\rangle$	risk - free	21
	$(5.72, \infty)$	at risk	7
At the risk of poverty rate	$(-\infty, 14.67\rangle$	low	10
	$(14.67, 19.23\rangle$	medium	11
	$(19.23, \infty)$	high	7
Total general government expenditure	$(-\infty, 41.2\rangle$	below average EU	9
	$(41.2, 46.3\rangle$	average EU	10
	$(46.3, \infty)$	above average EU	9
Gross domestic product per capita	$(-\infty, 16800\rangle$	below average	9
	$(16800, 34555\rangle$	average	10
	$(34555, \infty)$	above average	9
Population by educational attainment level - upper secondary, post-secondary non-tertiary, levels 3-4	$(-\infty, 38.9\rangle$	low	7
	$(38.9, 52.5\rangle$	medium	11
	$(52.5, \infty)$	high	10
Population by educational attainment level - less than primary, primary and lower secondary education, levels 0-2	$(-\infty, 18.4\rangle$	low	9
	$(18.4, 25.3\rangle$	medium	10
	$(25.3, \infty)$	high	9

Source: data Eurostat, author processing by WEKA

We used data mining methods to extract the models describing the investigated data. We used and tested several methods of the classification: methods using information theory (algorithm J48), based on decision trees (Random Forest, Random Tree), methods based on conditional probability (Bayes Net, Naive Bayes), rules PART, classifiers on the principle of k-nearest neighbors (classifier lazy IBK, Instance Bases Learning with parameter K), meta-algorithm Bagging. We used Weka toolkit to analyze the performance of the classifiers. We used several sampling methods to test and build the model (Evaluation of Test Set): Cross Validation Fold, Use Training Set, 66% Percentage Split. We chose the best model based on the values of the following indicators: correctly classified instances, incorrectly classified instances, kappa statistics, area under receiver operating characteristic curve (ROC), area under precision-recall curve (PRC).

RESULTS AND DISCUSSION

Given the values of the monitored indicators, we chose the model created by the J48 algorithm as the best model (Figure 1). Although this model achieved the values of correctly and incorrectly classified instances the same as the model created on the basis of the rules PART classification and also the Bagging meta and the Bayes Net or lazy IBK classifier, J48 achieved a lower error rate and slightly higher area values under the ROC curve.

```
=== Summary ===

Correctly Classified Instances          25               89.2857 %
Incorrectly Classified Instances         3               10.7143 %
Kappa statistic                          0.8369
Mean absolute error                      0.1196
Root mean squared error                  0.2445
Relative absolute error                 27.3166 %
Root relative squared error             52.3038 %
Total Number of Instances               28

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,900     0,111     0,818       0,900   0,857      0,774    0,919     0,789     low
               0,818     0,059     0,900       0,818   0,857      0,774    0,922     0,850     medium
               1,000     0,000     1,000       1,000   1,000      1,000    1,000     1,000     high
Weighted Avg.  0,893     0,063     0,896       0,893   0,893      0,830    0,941     0,866

=== Confusion Matrix ===

 a b c   <-- classified as
 9 1 0 | a = low
 2 9 0 | b = medium
 0 0 7 | c = high
```

Figure 1 Output algorithm J48 (Use training set)
Source: data Eurostat, author processing by WEKA

The correctly classified instances were 89.29% and the incorrectly classified instances were 10.71%. Reached value of Kappa statics (0.84) is considered as very good. It is outstanding degree of agreement between two sets of categorized data, observed and predicted values. Mean absolute error is measure set of predicted value to actual value i.e. how close a predicted model to actual model [6]. The mean absolute difference between the predicted and

observed values reached the value of 0.12. Root mean square error (RMSE) is measuring the differences between values predicted by a model and the values actually observed, so small value of RMSE means better accuracy of model [6]. Root mean square error reached the value of 0.25. The relative absolute difference between the predicted and actual values was 27%. The ratio of the number of observations predicted as a low at-risk-of-poverty rate to the total number of observations representing a given low-at-risk-of-poverty category was $\frac{9}{10} = 0{,}9$. The ratio of the number of observations predicted as the average at-risk-of-poverty rate to the total number of observations representing the given category of the average at-risk-of-poverty rate was $\frac{9}{11} = 0.\overline{81}$. All of observations predicted as a high at-risk-of-poverty rate belonged to representing a given high-at-risk-of-poverty category $\left(\frac{7}{7} = 1\right)$. A false negative rate in the low at-risk-of-poverty category obtained value 0.11 and in the medium-on-poverty-weight category obtained value 0.59. Detailed accuracy (False positive rate, Precision, Recall, Matthews Correlation Coefficient, ROC Area, PRC Area) by class is shown in Figure 1. We have observed a high correlation between observed and predicted values (83%). The area under the ROC curve is graphically shown for two categories in Figure 2 and Figure 3. The area under the feedback and accuracy curve took on values of 79%, 85%, 100%, which means high accuracy and feedback for all categories.

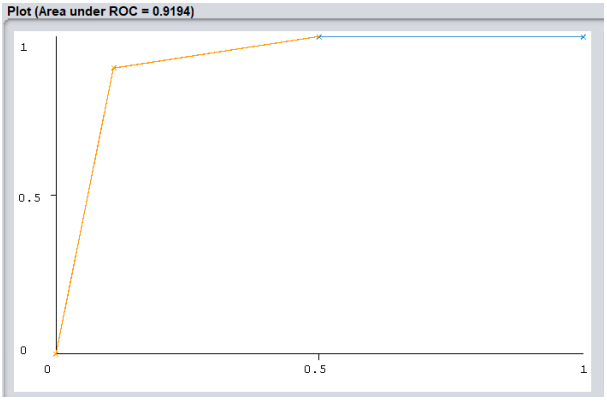

Figure 2 ROC curve of low variation at the risk of poverty rate (J48)
Source: data Eurostat, author processing by WEKA

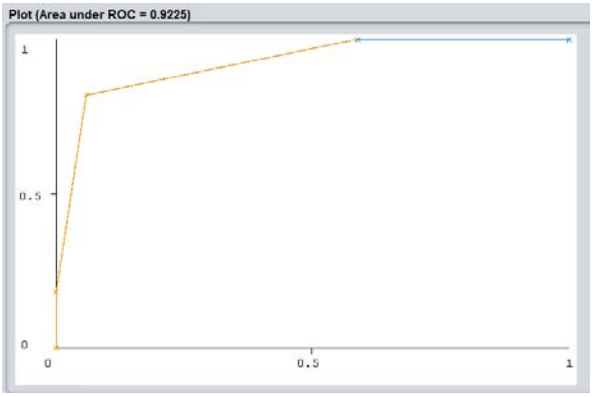

Figure 3 ROC curve of medium variation at the risk of poverty rate (J48)
Source: data Eurostat, author processing by WEKA

According to confusion matrix (Figure 1) we can say that one country had a low at-risk-of-poverty rate but was predicted as a medium at-risk-of-poverty rate and in two cases, countries achieved a medium at-risk-of-poverty rate but have been predicted to have a low poverty rate.

The decision tree is shown in Figure 4. The J48 algorithm decided that the root decision node would be the variable Income quintile share ratio (S80/S20). It builds the decision tree from labeled training data set using information gain and to make the decision the attribute with highest normalized information gain is used. The splitting procedure stops if all instances in a subset belong to the same class [9]. The tree contains two intermediate nodes (branches) formed by the variables Gini coefficient of equivalised disposable income (Gini) and the population with secondary education (PopSecEdu). The tree is terminated by 5 leaf nodes (leaves), which also contain the numbers of correctly and incorrectly classified variables.

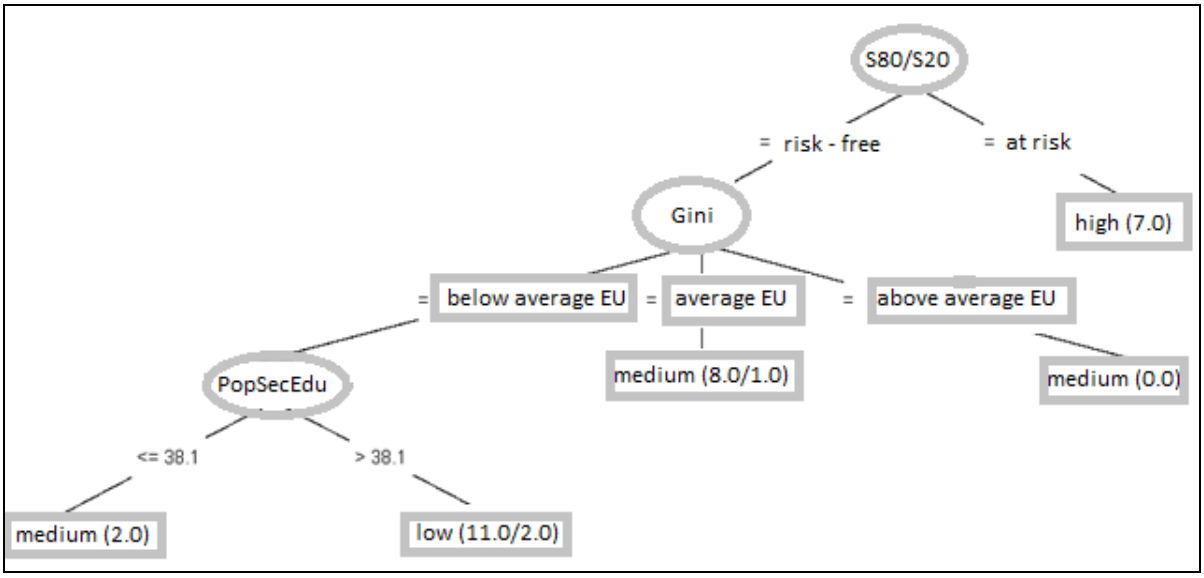

Figure 4 Decision tree of algorithm J48 (Use training set)
Source: data Eurostat, author processing by WEKA

CONCLUSIONS

In this article, we dealt with one technique of Data Mining. We applied the classification methods to the dataset of data obtained from the international Eurostat database. As a classification attribute, we determined the at-risk-of-poverty rate in the population. We focused on EU28 countries and 11 attributes. We found that the classification attribute was significantly positively affected by Total general government expenditure, GDP per capita, Income quintile share ratio (S80/S20), Gini coefficient of equivalised disposable income and also Proportion of population according to the level of primary and secondary education. We used Weka tools to search the best classification algorithm. The models created by classification techniques were building based on training data. To evaluate the performance of classifiers Weka data mining tool was used and the accuracy measures like Kappa statistic, TP rate, FP rate, Precision, Recall, F-measure, ROC Area and PRC Area. Overall observation was that the best algorithm suitable for predicting the at-risk-of-poverty rate in the monitored countries is J48.

REFERENCES

[1] Carlsen, L. & Bruggemann, R. (2021). Inequalities in the European Union—A Partial Order Analysis of the Main Indicators. *Sustainability*, 13(11), 6278. doi: https://doi.org/10.3390/su13116278

[2] Eurostat (2020). *Statistics of income poverty* (in Slovak). Retrieved 2021-10-20 from http://ec.europa.eu/eurostat/statistics-explained/index.php?title=Income_poverty_statistics/sk

[3] Ivanová, E. & Grmanová, E. (2021). The sustainability of EU labor immigration in terms of poverty inequalities and employment. *Sustainability (Switzerland)*, 13(4), 2265. doi: https://doi.org/10.3390/su13042265

[4] Janovičová, L. (2018). *Income Inequality in Selected EU Countries*. Zborník abstraktov a článkov z konferencie Prehliadka prác mladých štatistikov a demografov 2018, p. 22, Bratislava. Retrieved 2021-10-10 from http://www.ssds.sk/publikacie/VS2018_Zbornik.pdf

[5] Janovičová, L. & Bartová, L. (2018). *Income Inequality and Poverty Risk in V4 Countries*. Zborník abstraktov z 27. Medzinárodného vedeckého seminára Výpočtová štatistika 2018, s.11, Bratislava. Retrieved 2021-10-10 from http://www.ssds.sk/publikacie/VS2018_Zbornik.pdf

[6] Kumar, Y. & Sahoo, G. (2012). Analysis of Parametric & Non Parametric Classifiers for Classification Techniqur using WEKA. *International Journal of Information Technology and Computer Science* (IJITCS), 4(7), 43-49. Retrieved 2021-10-31 from https://www.mecs-press.org/ijitcs/ijitcs-v4-n7/IJITCS-V4-N7-6.pdf

[7] Muster, R. (2021). Employees' poverty: Poland in comparison to other EU countries. *Problemy Polityki Spolecznej*, 53, 26-53. doi: https://doi.org/10.31971/pps/142005

[8] Parmar, M. (2018). Comparative Analysis of Classification Techniques using WEKA on Different Datasets. *International Journal of Latest Engineering and Management Research* (IJLEMR), 3(6), 1-5. Retrieved 2021-10-31 from http://www.ijlemr.com/papers/volume3-issue6/1-IJLEMR-33190.pdf

[9] Sharma, T.Ch. & Jain, M. (2013). WEKA Approach for Comparative Study of Classification Algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(4), 1925-1931. Retrieved 2021-10-31 from https://www.ijarcce.com/upload/2013/april/60-trilok-WEKA%20approach%20for%20comparative.pdf

[10] Tkachova, N., Krushelnytska, T., Marchenko, O. & Kuznetsova, N. (2021). Migration policy in the context of sustainable development. *WSEAS Transactions on Business and Economics*, 18, 619-627. doi: https://doi.org/10.37394/23207.2021.18.61

[11] Vlačuha, R. & Kováčová, Y. (2018). *EU SILC 2017 Indicators of poverty and social exclusion.* Statistical Office of the Slovak Republic, Demography and social statistics (in Slovak). Retrieved 2021-10-01 from https://slovak.statistics.sk/wps/wcm/connect/c685cd00-9241-4785-9695-b8b571595de3/EU_SILC_2017_ Indikatory_chudoby_a_socialneho_vylucenia.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-c685cd00-9241-4785-9695-b8b571595de3-ml2MZIg

[12] Želinský, T., Mysíková, M. & Garner, T.I. (2021). Trends in Subjective Income Poverty Rates in the European Union. *European Journal of Development Research* (2021). doi: https://doi.org/10.1057/s41287-021-00457-2