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

ISSN 2307-4531
(Print & Online)

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

Data Mining of SILC Data: Turkey Case

Olgun Özdemir[a]*, İbrahim Demir[b]

[a]*PhD Student, Yıldız Technical University, İstanbul, 34220, Turkey*

[b]*Assistant professor, Yıldız Technical University, İstanbul, 34220, Turkey*

[a]*Email: olgunozdemir35@gmail.com*

[b]*Email: idemir@yildiz.edu.tr*

Abstract

Official data produced by the National Statistical Institutes (NSIs) have an essential place in the governmental economic and social decision-making process. Addressing official data with data mining methods rather than traditional statistical approaches is crucial to extract new information and hidden patterns. However, the usefulness of data mining methods for official statistics remains unexplored. In the present study, SILC (Survey of Income and Living Conditions) data for the year 2015 conducted by the Turkish Statistical Institute (TurkStat) are examined with data mining methods. Cross-sectional data of 36036 individuals were handled, and the variables affecting the individual income were determined, also the welfare status of the individuals was examined. To determine the socio-economic profiles of individuals, latent class analysis (LCA) and k-modes clustering analysis were used. The socio-economic status of individuals was classified using clustering and random forest (RF) algorithm models. In the LCA model with ten classes, it was obtained which probability of a newly selected individual would belong to which class. The latent class profile definitions of the individuals were obtained according to the variable values obtained from the latent classes with the highest probability. Ten clusters obtained as a result of k-modes were defined according to cluster modes, and cluster profile definitions of individuals were obtained, and also their results were compared with LCA results. In this study, in which categorical variables were considered, it was seen that LCA method provided more consistent results than k-modes method. In the RF model, where individual income is selected as a function of all nine input variables, the importance of the variables was determined. It is observed that education, occupation, and age variables were more important and made the most contribution to the RF model, respectively.

---

* Corresponding author.

In the SILC data, which is an extensive and detailed data, methods such as LCA and RF seem to be appropriate for the application of data mining and obtaining meaningful results from the data. Similar data mining processes can be used to obtain meaningful results for different official data.

Keywords: Data mining; SILC; cluster analysis; latent class analysis; k-modes; random forests.

1. Introduction

Data mining has been developing rapidly over the last two decades. Official statistics have been a new field for data mining in recent years. Using data mining methods to obtain new information and unexpected patterns from official statistics has not yet been adequately explored. It is possible to get undiscovered structures, models and patterns in these large amounts of data sets produced by National Statistical Institutes (NSIs) through data mining methods rather than traditional statistical methods. The advantages of data mining techniques in official statistics have been discussed in [1]. EU-SILC (European Union Statistics on Income and Living Conditions) is a comprehensive survey providing comparable cross-sectional and longitudinal multidimensional microdata on income, living conditions, poverty and social exclusion in European countries. Detailed information about EU-SILC is included in [2]. As official statistics, Turkey SILC data, which is equivalent to a national version of the EU-SILC data for Turkey provides rich and detailed information about income and demographic characteristics at individual and household levels to measure the income, poverty, social exclusion and living conditions. Empirical studies that examine the income, socio-economic and welfare status of individuals rather than households in Turkey in a comprehensive manner with data mining methods are insufficient. By conducting this study, we aimed to contribute by presenting the results related to the welfare status of individuals in a comparative way with data mining methods. Thus, by identifying the data mining methods effective in addressing such data, meaningful patterns and relationships regarding the socio-economic status of individuals will be obtained. The use of data mining methods for official statistics is very recent. Reference [3] reported that in the first five years of the new millennium, there were very few reported applications of data mining in official statistics. In addition, even with a little change in 2010, the only reported study was the study by [4], a small application of data mining techniques to official data. In recent years, data mining approaches to official data have increased. Income is an important measurement variable for socio-economic status and living standards and is also important in establishing decision-making processes. In this study, to see the effect of income on the socio-economic or welfare status of the individual, individual income instead of household income is discussed. Reference [5] analyzed the earnings of immigrants in Ireland by descriptive statistics and regression analysis using the data from the 2005 wave of the EU-SILC. Using the Czech national module of the EU-SILC of the 2007 year, Reference [6] analyzed the income distribution and living conditions of Czech farmers according to the volume of the average income per person based on disposable income. In SILC data, very few of the variables are continuous and most are categorical variables. For the analysis of large data files with categorical variables, Reference [7] examined the methods used in clustering categorical data [8], using Czech EU-SILC data for 2011, analyzed nominal indicators showing material deprivation by hierarchical cluster analysis and various k-clustering methods. Again with the same data set, Reference [9] used two-step cluster analysis and latent class analysis (LCA), which are alternative categorical data clustering methods besides recently introduced similarity measures. Reference [10] investigated the socio-economic factors affecting household

disposable income by using data mining techniques on Czech EU-SILC data for 2005, 2010, and 2014. In addition, Reference [11] used tree-based estimation methods, random forest (RF), and LCA approach to estimate inequality of opportunity by using the 2011 wave of the EU-SILC data. Although these studies contribute, the studies that have been conducted in the literature have mostly dealt with disposable income, and the results have been obtained with limited methods. In this study, SILC data was examined extensively by using advanced data mining methods, individual income and variables affecting it were obtained and, patterns and clusters were interpreted comparatively. When conducting a data mining study, the variables considered during the process and the data mining methods appropriate to the types of these variables may change. Therefore, in a data mining study, although the direction of the process may change, it is not possible to predict the results you will obtain in the light of the data and the method you use. Most of a data mining study comprises a data manipulation process and determination of appropriate methods. In this study, where we aim to evaluate high dimensional SILC data comprehensively with data mining methods, we tackled the following questions:

- What are the variables affecting the total annual income of household members and at what level do they influence?
- What is the pattern between the variables affecting the total annual income of the household members?
- What generalizations or rules can be put forward among household members in terms of variables affecting annual total income?

The main objective of this study is using data mining methods to determine the socio-economic factors affecting the income status of the individuals in Turkey and to classify the individuals according to their place in income categories and their socio-economic profile, as well as to investigate their welfare status. In our paper, we focus on the measurement of individual income and the variables affecting this income. We conduct our analysis using microdata from the 2015 wave of the SILC survey provided by Turkish Statistical Institute (TurkStat). For this purpose, first, the individual income variable was categorized and socio-economic variables affecting individual income were determined. Then, data mining methods that best fit the data set formed by the decided variables were determined. Since the data set consists of categorical variables, LCA and k-modes methods are used for clustering and RF method for classification, as the most appropriate data mining methods for the existing data set. Class and cluster definitions were compared as a result of LCA and k-modes analyses. Again, a classification model was obtained as a result of random forest (RF) analysis and it was found which variables play an important role in the model. The rest of the paper is organized as follows. Besides introducing the data set, Section 2 describes the organization of pre-analysis data (data manipulation), the selection of variables and the application of the methods used in the analysis. Section 3 provides the empirical findings on clustering and classification, and summarizes them. The study is concluded in the final Section 4.

2. Materials and methods

2.1 Data and Preparation

We use cross-sectional SILC survey microdata from the 2015 year which is recently published by TurkStat and

it was aimed to examine this data with data mining methods. For this purpose, personal, personal-register and household type of cross-sectional datasets were taken into consideration. Individual income and the variables affecting it were investigated with data mining methods and the welfare status of the individuals were examined. In the study, Microsoft Excel 2013 was used for data editing and RStudio (version 1.1.456 and R version 3.5.1) was used for data mining analysis. Microdata sets were examined in detail in the data exploration phase. For the individuals aged 15+ years, in addition to variables that affect income (basic characteristics of individuals, income and income-related variables) and those that apply to all individuals (educational status, health status, etc.), some variables have been transformed and combined according to the purpose and curiosity of the research during the analyzes. In this study, first, variables and components affecting income are discussed for the distribution of income, which is usually studied in Turkey as functional, individual, sectoral and regional. In the application where the data set for income and living conditions survey was used, with the acceptance that the research was carried out within itself, the sample selection and size, validity of survey questions and reliability of survey data were not tested. Data provided in the csv format were regulated through the Microsoft Excel program. Variables were obtained and coded from the cross-sectional data sets of the personal, personal-register and household. FG140 variable, which is the sum of the incomes obtained by the individual in the personal data set, was weighted by the personal weight variable FB030, since each individual has been sampled by TurkStat to represent a certain number of individuals throughout Turkey. Thus, instead of the income scale, 20% weighted income group ranges were obtained for individual income in order to accurately reflect the related year data and to make correct comparisons. To obtain meaningful income values, the positive and negative values of FG140 were excluded and only positive income values were taken into consideration. Weighted quantile values were calculated using FB030 and FG140 variables. The reason we use weighted quantiles is to obtain appropriate categorical ranges in such a way as to estimate the income variability in overall Turkey. Because it is intended to place an equal number of data for each income group, weighted 20% quantiles were used. Weighted 20% quantiles were calculated with the "Laeken" RStudio package and the same calculations were made in Excel environment and, cross-check was provided. The method for weighted quantiles, which is introduced by [12] was used. The weighted quantile values calculated as of 2015 for the individual income is:

Table 1: Weighted quantiles for the individual income of the year 2015

	Income	
	Min	**Max**
1. quartile	0	6615
2. quartile	6616	12000
3. quartile	12001	16200
4. quartile	16201	26400
5. quartile	26401	+

In this study, first, the clustering method was selected, then the variable selection was made for clustering. All variables discussed in the study (including missing values) are categorically encoded, to begin with, the first value of 1. For the missing values, a new variable is assigned to express these values. Since all the values in the data set are categorical, clustering methods for categorical data sets were investigated and, LCA and k-mode

clustering methods were determined to obtain cluster classes from the data in order to best suit the characteristics of the data set. Then, RF method was used for classification purposes. From the R packages, version 1.4.1 of the poLCA, version 0.6-14 of klaR and version 4.6-14 of randomForest package were used for LCA, k-mode and RF analysis, respectively. Since most of the variables in the SILC data set are categorical, data mining methods matching the categorical data were investigated, and the LCA method was preferred in this study as proposed by [9] for EU-SILC data and similar surveys. LCA is a highly robust clustering technique with a variety of applications to official statistics and a model-based approach that has become popular in recent years, offering high-quality, meaningful statistics. LCA is model-based and classifies respondents into hidden classes using prior probabilities ([13]). Categorical latent variables can be obtained from the two or more observed categorical variables with LCA, and it is seen as the categorical equivalent of factor analysis. k-modes is one of the first algorithms developed by [14] to cluster categorical data. The k-mode algorithm, which uses simple matching dissimilarity measure instead of Euclidean distance, is an extension of the standard k-means algorithm adapted to categorical data. k-modes uses modes to show cluster centers and updates modes in each iteration. Random forests are ensembles of decision trees to solve classification and regression problems ([15]). For factor type variables, RF can be used for classification. RF builds multiple decision trees and combines them to constitute a more accurate and stable prediction. The analyzes in this study were carried out on a computer with 16 GB RAM with an Intel ® Core ™ i7-8700K CPU @ 3.70GHz processor and the calculations were attempted to be accelerated by parallel processing using all the cores of the processor.

2.2 Feature Selection

From personal, personal-register and household data sets. 20 variables (some of which are encoded from existing variables) were initially determined and encoded for analysis, which were thought to affect the individual income. Among these variables, the employment sector (SEC) variable was coded as 3 different variables including 4, 7 and 10 categories, and the age (AGE) variable was also coded as 3 different variables with 3, 6 and 11 categories. The initial variables determined in the study are: Marital status (MST), Highest education level attained (EDU), General health status (HTH), Self-defined current economic status / employment status (EMP), Employment status in the main job (STA), Occupation code of the main job / occupation (OCC), Registration status to social security institutions in the main job (SSI), Gender (GEN), Geographical region where the household is located (RGN), Piped water system in the dwelling (WSY), Telephone line possession of the household (TEL), Internet connection possession of the household (INT), Household income type (Between incomes of non-operating, pension, rental of assets/lands, social transfer) (HIT), 20% income group of the household member (IG), SEC4, SEC7, SEC10, AGE3, AGE6, AGE11. To reduce the age and sector variables included in more than one category in the data set, the importance levels of the variables were determined and the variable selection (between 19 variables except IG income variable) was made using the Boruta R package. According to these results, AGE3 and AGE11 from age variables and SEC7 and SEC4 variables from the sector variable were excluded from the data set.. For this reason, by applying to the expert judgment, 14 variables needed for analysis were regulated in order to determine whether all 15 variables affected the income of the individuals (SSI, HTH, TEL variables were eliminated, and the equivalent household

size[a] (EHSIZE) variable was added). To reduce the complexity before the LCA analysis to be applied, the variable selection was performed on the remaining 15 variables using LCAvarsel R package. Genetic algorithm, swap, and backward-stepwise algorithms are used in LCAvarsel for the variable selection procedure. Using the genetic algorithm in LCAvarsel package, STA, OCC, RGN, AGE_6 ve GEN variables were found to be important. When swap and stepwise algorithms are used, the same variables (WSY, INT, HIT, EHSIZE_4) were found to be important in both algorithms. Thus, the variable selection using the LCAVarsel package did not yield the desired result by finding very few variables as important despite different attempts Thus, for LCA analysis, the most extensive set of relevant variables for variable selection was determined. LCA analysis was carried out using the data related to 36036 individuals by determining 10 variables (STA, OCC, RGN, INT, SEC, EDU, IG, AGE_6, GEN, MST) which are thought to affect the individual income most based on expert judgment. Variables which were thought to be related to each other and which did not show much change according to the results of the previous poLCA analyzes were extracted. Since the variables HIT, WSY, EHSIZE_4 do not affect the classes much in the LCA experiments with poLCA; AGE, GEN, MST variables have been added instead of these variables. Finally, based on expert judgment, the most appropriate set of variables has been determined to reduce the number of classes: STA, OCC, RGN, INT, SEC, EDU, IG, AGE, GEN, MST. In terms of comparability, k-modes and RF analyses are also based on this set of variables.

2.3 Methods

For the variables STA, OCC, RGN, INT, SEC, EDU, IG, AGE, GEN, MST, the local minima (lowest BIC value in poLCA results) was not found, although the number of classes was initially set to 30 and the number of iteration was increased. Therefore, in the LCA analysis performed without covariate, the results were evaluated based on the number of classes 10 to obtain a stable result and interpretable solution. Again, using the same variables set, it was desired to compare the results obtained with the k-modes algorithm with the LCA results. The k-mode algorithm was chosen instead of the k-means algorithm because the data set discussed was categorical. The k-mode analysis using the klaR R package was performed using 3000 as the iteration value. By using the RF classification model over the determined variables, it is aimed to create a classification model by creating rules for the future household members. Data were analyzed in RStudio environment using "randomForest" and "caret" packages. 70% of the data consisting of 36036 observations were used as train data and 30% as test data. In RF analysis, the results were first created for train data, then compared to test data. The steps applied are:

- First, the RF model was applied for train data and error matrix was obtained. Then, the prediction model was created for all train data by using the same RF model. Then, the prediction results for the test data were obtained over the same RF model and the improvement in accuracy was examined by the error matrix and compared with the results obtained for the train data. The error ratio graph was obtained for this first RF model, and the tree value which provides OOB (out-of-bag) generalized error

[a] EHSIZE variable is calculated according to the "modified OECD" equivalence scale, which gives:
• a weight of 1.0 to the reference person;
• a weight of 0.5 to any other household member aged 14+;
• a weight of 0.3 to each child aged 15-,
and the equivalent household size was found by summing the results. This variable is coded for 1-2, 3-4, 5-6, 7+ intervals as EHSIZE_4.

to reach its stable value is determined by trials. Besides, by plotting the OOB error as the function of mtry, the optimal mtry value for which the error was the lowest was determined. With the most appropriate tree value and mtry parameters, it was desired to observe the change in the accuracy of the model by applying RF model to train and test data.

- The histogram graph was used to visualize the number of nodes in the RF trees, and the importance levels of the variables for the RF model were expressed, and the frequencies of the variables in the model were shown.
- The marginal effect of variables on classes was measured with partial dependence plots.
- The output of a single tree for the RF model was obtained, and the first four trees of the model were visualized.

3. Results

3.1 LCA Results

LCA analysis was performed (without co-variable / equivalent variable) using the following code depending on the variables ”STA, OCC, RGN, INT, SEC, EDU, IG, AGE, GEN, MST", and as a result, ten separate latent classes were obtained:

Table 2: LCA analysis commands

Commands: LCA analysis with ten selected variables
> set.seed(100000)> library(poLCA)
> library(foreach)
> library(doParallel)
> registerDoParallel(cores=6)
> f_Data_sel <- cbind(STA, OCC, RGN, INT, SEC, EDU, IG, AGE_6, GEN, MST)~1
> sink("output.txt")
> print(lc_sel <- foreach(i=2:30, .packages="poLCA") %dopar% poLCA(f_Data_sel, Data, nclass=i, maxiter=3000, tol=1e-5, na.rm=FALSE, nrep=10, verbose=TRUE, calc.se=TRUE))
> sink()

The outputs are as follows:

In the output, “Estimated class population shares” and “Predicted class memberships” values are close to each other, showing that the model fits well. Estimated class population shares are estimated rates, to be the number of observations fall into each latent class. For example, the share of observations in the 10-class model was estimated as 14.39% in latent class 1, 9.92% in latent class 2, and 11.42% in latent class 3, respectively. The values of “Predicted class memberships” refer to the probability that a newly selected individual will be included in the relevant class. The latent class probabilities for the variables in the LCA analysis outputs are tabulated in Appendix I. In the output, for example, the latent class "Class 1" can be interpreted as follows:For

the variable EDU, it is 51.92% probability that an individual answering "2 (primary school)" will be included in the latent class 1. Thereby, it is observed that individuals in latent class 1 are more likely to be a primary school graduate (Pr (2)).In the latent classes obtained in Appendix I, the variable values that the variables get at the highest probability are summarized as follows to characterize the latent class:

```
Estimated class population shares
 0.1439 0.0992 0.1142 0.1427 0.0862 0.1517 0.0552 0.064 0.0553 0.0877

Predicted class memberships (by modal posterior prob.)
 0.1445 0.1021 0.1136 0.1447 0.0867 0.1514 0.0529 0.0636 0.053 0.0874

=========================================================
Fit for 10 latent classes:
=========================================================
number of observations: 36036
number of estimated parameters: 439
residual degrees of freedom: 35597
maximum log-likelihood: -393680.6

AIC(10): 788239.1
BIC(10): 791967.2
G^2(10): 121462.7 (Likelihood ratio/deviance statistic)
X^2(10): 3770834 (Chi-square goodness of fit)
```

Figure 1: Output of LCA analysis with ten selected variables

Table 3: Latent classes of LCA

	Manifest Variable									
Latent Class	**STA**	**OCC**	**RGN**	**INT**	**MST**	**AGE**	**GEN**	**SEC**	**EDU**	**IG**
1	6	7	1	2	1	5	1	10	2	3
2	1	3	1	2	1	3	1	4	2	4
3	1	4	1	1	1	3	1	2	2	4
4	1	1	1	1	1	2	1	6	6	5
5	4	4	5	2	1	4	1	1	2	4
6	6	7	1	2	2	2	2	10	1	1
7	2	6	6	2	1	1	1	3	2	2
8	6	7	1	1	2	1	2	10	3	1
9	1	3	1	2	1	3	2	6	2	2
10	1	3	1	1	2	1	1	6	3	3

Again, the latent classes attained can be defined in Appendix II according to the variable values that the variables get at the highest probability.

3.2 k-modes Results

The k-mode algorithm was applied to the data set with the help of the following code:

In Table 3, the second cluster has the most significant number of members with the sample size 7840. The number of individuals in the eighth cluster is only 1434. It can be said that the data is distributed evenly to the clusters. Initially, more than three variables were chosen to decide how many clusters we needed, but at the same time, we preferred not to have more than 10 clusters to reduce complexity and increase interpretability.

For this reason, the number of clusters was determined as 10 and 10 clusters were obtained by using LCA variables, and the results were compared with LCA results.

Table 4: k-modes analysis commands

Commands: k-modes analysis with ten selected variables
> set.seed(100000)
> library(klaR)
> c1 <- kmodes(Data_kmodes, 10, iter.max = 3000, weighted = FALSE, fast =TRUE)

In the outputs of the k-mode analysis applied to the SILC data; the cluster sizes, mode values and simple matching distances were obtained for each cluster.

Table 5: Cluster sizes

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Size	3405	**7840**	2893	4723	3587	1653	1899	**1434**	5390	3212

Table 6: Cluster modes of k-modes clustering

	Variables									
Clusters	**STA**	**OCC**	**RGN**	**INT**	**MST**	**AGE**	**GEN**	**SEC**	**EDU**	**IG**
1	1	1	5	1	1	2	2	6	6	5
2	6	7	1	2	2	6	2	10	1	1
3	1	5	4	2	1	1	1	2	3	5
4	4	4	6	2	1	3	1	1	2	2
5	1	4	1	2	1	2	1	2	2	4
6	1	3	3	2	1	4	2	6	2	1
7	1	2	1	1	2	2	1	6	6	4
8	1	3	3	2	2	2	1	4	3	2
9	6	7	1	2	1	5	1	10	2	3
10	1	1	1	1	1	3	1	6	6	5

The clusters attained from the k-mode analysis can be defined in Appendix III according to the corresponding cluster modes from Table 6.

A suitable clustering method produces high-quality clusters with minimum in-cluster distances. Simple-matching distance was used to measure the distance. Table 7 shows the simple matching distance of each cluster. The total distance of the second cluster is 25458. The eighth cluster has the lowest simple matching distance. The low distance demonstrates that the intra-cluster data show high similarity to each other. Here, it depends on the lowest number of individuals in the eighth cluster. The income groups in each cluster were obtained as follows.

Table 7: Cluster simple-matching distance

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Simple-matching distance	12730	**25458**	13124	17787	13124	6791	7309	**5601**	15990	9815

Table 8: Table of income groups in each cluster

Clusters										
Income Groups	1	2	3	4	5	6	7	8	9	10
1. 20%	242	**4508**	422	746	141	**628**	218	180	410	49
2. 20%	377	2343	458	1476	253	487	227	694	1343	117
3. 20%	466	506	766	780	799	289	425	345	2120	246
4. 20%	381	386	426	957	2191	160	799	119	1162	256
5. 20%	**1939**	97	821	764	203	89	230	96	355	**2544**

Table 8 shows the number of individuals in the income group in each cluster. We see that most of the lowest income group (4508) are in the second cluster, and most of the highest income group (2544) are in the tenth cluster. Again most of the individuals in clusters 1 and 10 are in the highest income group, while most of the individuals in clusters 2 and 6 are in the lowest income group. Below; the first graph shows the distribution of data between all variables according to the clusters, and the other graphs show how the data between selected variables are distributed among the clusters:

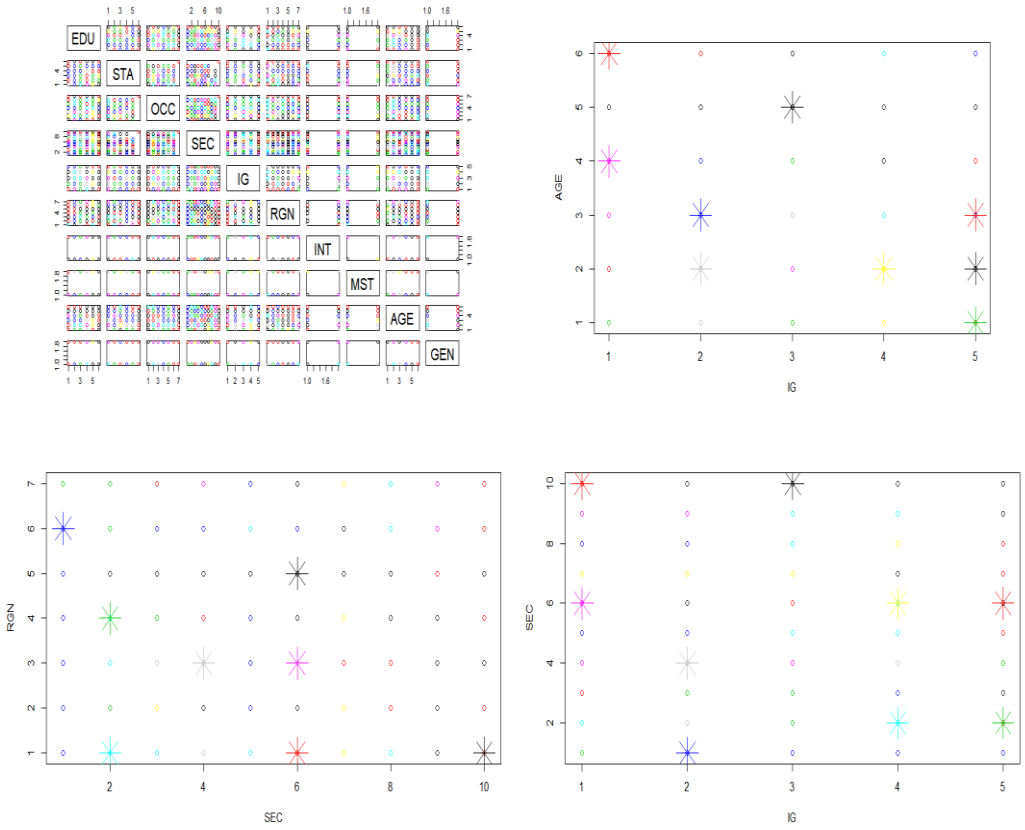

Figure 2: Cluster graphs for SILC data after k-mode clustering

Figure 2 shows cluster charts for the data. Different colors in the chart refer to different clusters. Here, binary cluster graphs are also presented for some variables. When the charts are evaluated together with Table 6, it can be seen that for example, the blue-colored cluster point corresponds to the 4th cluster in Table 6 (when IG = 2, SEC = 1, RGN = 6, and AGE = 3).

3.3 Classification Results with RF Algorithm

3.3.1 RF Model for Train Data

In RF analysis, IG (response variable) was selected as the function of all 9 other variables (input variables). Since IG is a factor variable, this RF has resulted in classification. The number of trees (ntree value) is 500 by default. The number of variables used in each division (No. of variables tried at each split, ie the mtry value) is by default the square root of the total number of variables (square root of 9 variables). The error matrix (also known as confusion matrix) obtained when the RF algorithm is applied based on the IG income group variable for the train data is as follows:

Table 9: Error matrix for train data

Income groups	1. quantile	2. quantile	3. quantile	4. quantile	5. quantile	Classification Error
1. quantile	**3332**	1207	401	230	133	37.2%
2. quantile	1225	**2283**	1057	653	288	58.5%
3. quantile	499	1150	**1593**	1165	409	66.9%
4. quantile	238	647	986	**1951**	965	59.2%
5. quantile	90	174	340	963	**3406**	31.5%

When the random forest model was applied for train data, the "OOB estimate of error rate" ratio was found to be 50.5%, and the accuracy rate of the model was obtained in this vicinity. As we examine the error matrix, we see that the estimates are satisfactory in estimating the fifth income group. The highest error is 66.9% in predicting the third income group.

3.3.2 Error Matrix and Prediction for Train Data

When the estimation is made by using all the train data, the error matrix for the income variable is obtained as follows:

Table 10: Error matrix for prediction of train data

Income groups	1. quantile	2. quantile	3. quantile	4. quantile	5. quantile	Classification Error
1. quantile	**3973**	879	342	161	66	26.71%
2. quantile	834	**3365**	746	395	107	38.22%
3. quantile	250	688	**2824**	584	238	38.39%
4. quantile	150	397	662	**3140**	578	36.27%
5. quantile	96	177	242	507	**3984**	20.42%

It is seen that most of the data are classified correctly in the diagonal values in Table 10. For example, 3973 of the data is classified as the first income level and 3365 for the second income level. Those who are not included in diagonal are misclassified. When the diagonal values are summed and divided by the number of train data (25385), it gives the accuracy value (68.1%). It is observed that there is an accuracy value of over 50%. The reason for the high accuracy is that we used all train data in the model (i.e., it was higher than the previous 1-OOB (1-50.5%)). It can also be said that the confidence interval obtained for accuracy (0.6752, 0.6867) is also a narrow confidence interval, so the results seem good. Also, the Kappa value has resulted in a significant amount of 0.60.

3.3.3 Error Matrix and Forecast for Test Data

Table 11: Error matrix for prediction of test data

Income groups	1. quantile	2. quantile	3. quantile	4. quantile	5. quantile	Classification Error
1. quantile	**1407**	480	208	107	38	37.19%
2. quantile	527	**985**	424	271	73	56.80%
3. quantile	141	433	**676**	454	132	63.18%
4. quantile	101	250	458	**832**	399	59.22%
5. quantile	65	121	160	386	**1523**	32.46%

Table 12: Sensitivity values of classes for test data

Class	1. quantile	2. quantile	3. quantile	4. quantile	5. quantile
Sensitivity	0.6278	0.43411	0.35099	0.40585	0.7035

In the estimation with test data, the accuracy (50.92%) was lower than the train data. Since the test data is not trained by the random forest model, it can be said to be a more accurate assessment for model accuracy. The confidence interval was obtained as (0.4996, 0.5187) and it is still narrow and good. In addition, it is seen that there is less accurate classification in the error matrix than the previous one. The sensitivity values are formed by dividing the corresponding class value in the error matrix by the identical column sum. When the sensitivity values in Table 12 are considered, it is seen that the best sensitivity value is 70.35% in estimating Class 5, and it is higher than in other classes.

3.3.4 RF Error Ratio

The error ratio in the RF model can be learned graphically by using the random forest model (via train data):

As the number of trees in the graph increases, the Out-of-Bag (OOB) generalized error appears to have declined and stabilized. This error stabilized after about 300 trees (by testing different values, OOB was the lowest in 310 trees), and after this value, the model has not improved further.

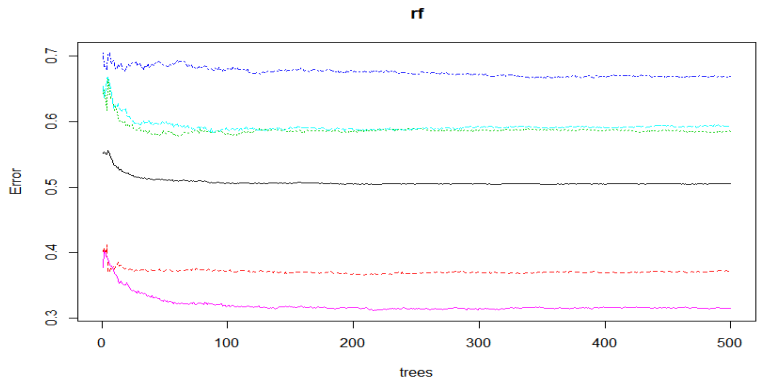

Figure 3: Error ratio of random forest

3.3.5 Improvement of RF Model (Tune Operation)

To improve the mtry value, the OOB error is plotted as a function of the mtry value:

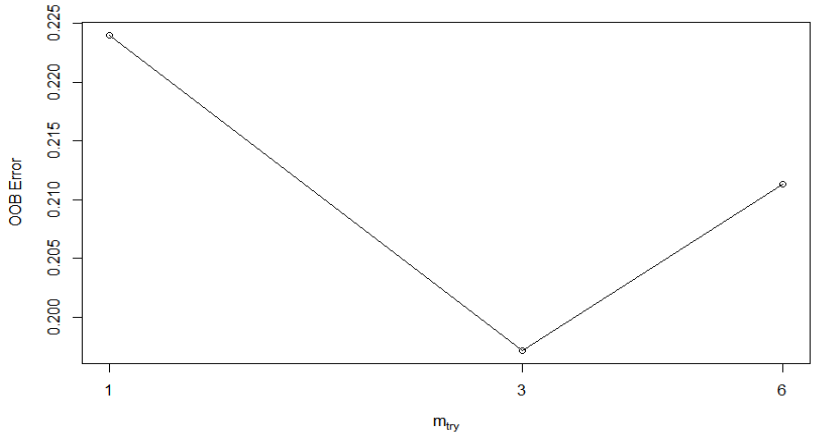

Figure 4: Determination of the most appropriate mtry value

In Figure 4, the OOB error is relatively high in mtry=1, while in mtry=3 it is the lowest value, and in mtry=6 it increases again. Therefore, this graph shows that the mtry value should be selected as 3. Returning to the initial random forest model and adding the ntree=310 and mtry=3 parameters, the following error matrix was obtained when the RF model was applied in the train data:

Table 13: Error matrix of the improved model for train data

Income groups	1. quantile	2. quantile	3. quantile	4. quantile	5. quantile	Classification Error
1. quantile	**3348**	1184	402	235	134	36.87%
2. quantile	1234	**2276**	1072	645	279	58.66%
3. quantile	508	1137	**1592**	1178	401	66.94%
4. quantile	239	645	993	**1950**	960	59.26%
5. quantile	93	174	341	952	**3413**	31.37%

The OOB error in the first model was 50.5%, and now it is 50.45%. In the estimation of the 1st and 5th classes, the classification error has improved (decreased according to the first model).

The following results are attained for the training error:

Table 14: Error matrix of the improved model for prediction of train data

Income groups	1. quantile	2. quantile	3. quantile	4. quantile	5. quantile	Classification Error
1. quantile	**3973**	879	342	161	66	26.71%
2. quantile	835	**3365**	746	393	106	38.20%
3. quantile	250	688	**2826**	584	239	38.39%
4. quantile	149	398	661	**3140**	577	36.24%
5. quantile	96	176	241	509	**3985**	20.41%

The accuracy value (0.6811) for the 310 trees was higher (previous 0.681). The actual test will be based on the test data, and accordingly, the error matrix is:

Table 15: Error matrix of the improved model for test data

Income groups	1. quantile	2. quantile	3. quantile	4. quantile	5. quantile	Classification Error
1. quantile	**1410**	479	210	104	38	37.08%
2. quantile	517	**980**	430	270	76	56.89%
3. quantile	146	431	**664**	449	131	63.54%
4. quantile	105	258	462	**845**	401	59.20%
5. quantile	63	121	160	382	**1519**	32.34%

Table 16: Sensitivity values of classes of the improved model for test data

Class	1. quantile	2. quantile	3. quantile	4. quantile	5. quantile
Sensitivity	0.6292	0.43191	0.34476	0.41220	0.7016

According to the previous results, there is an increase in the sensitivity values in Class 1 and 4.

3.3.6 Number of Nodes for Trees

The numerical size of the trees was obtained using a histogram:

Figure 5 shows the number of nodes in each of the 310 trees. The frequency of the most massive histogram bar is close to 60. There are approximately 60 trees with over 2900 nodes. There are also very few trees that contain less than 2200 or over 3200 nodes.

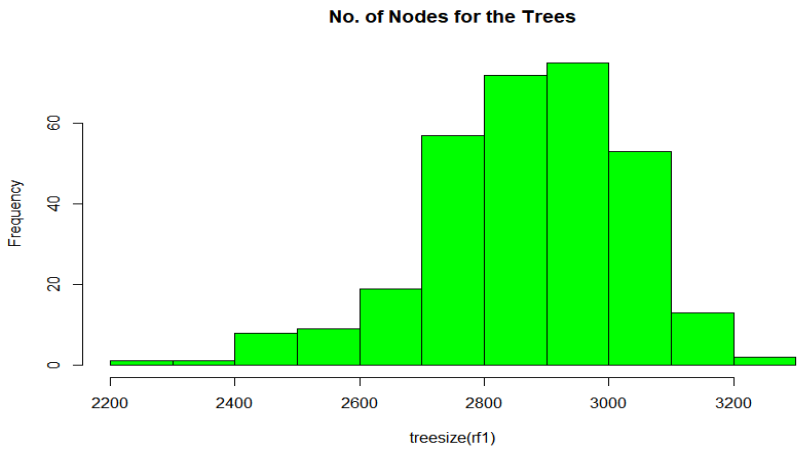

Figure 5: Number of nodes for trees

3.3.7 Variable İmportance

To see which variables play an essential role in the model, variable importance plot was generated:

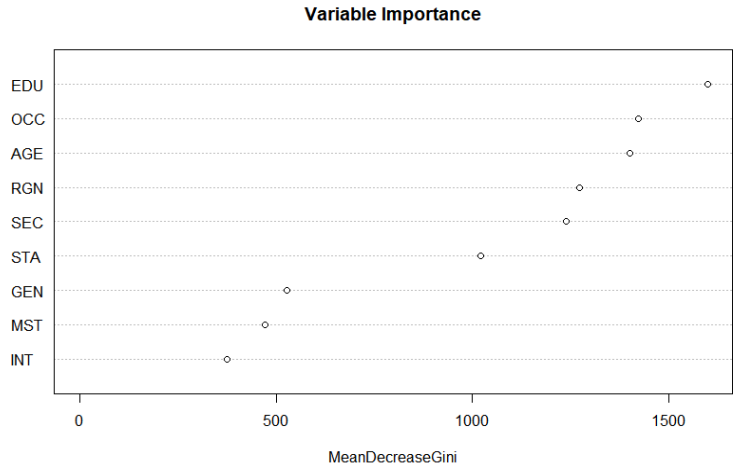

Figure 6: Variable importance

The graph in Figure 6 shows how pure the nodes are at the end of the tree when each variable is omitted. For example, when the EDU variable is removed, the average Gini value is reduced by 1500. It is observed that EDU, OCC, and AGE have the most contribution to the Gini parameter compared to other variables. In Table 17, the actual values representing the data points that make up the graph for nine input variables are shown in the "Mean decrease in Gini index" row. "Frequency of use" shows how many times each variable is found in the random forest model / how often it is seen (frequencies in the trees). For example, EDU variable is observed in the model 140887 times.

Table 17: Mean decrease in Gini index and frequency of use

Variable	EDU	STA	OCC	SEC	RGN	INT	MST	AGE	GEN
Mean decrease in Gini index	**1598.37**	1020.40	**1421.57**	1237.46	1273.45	374.19	471.47	**1401.21**	526.98
Frequency of use	140887	61423	104829	142353	198165	52228	33863	129448	26709

3.3.8 Partial Dependence Plots

Partial dependency graphs give a graphical representation of the marginal effect of a variable in the classification probability (classification). For example, when EDU and OCC variables are examined based on Class 1 and Class 5, Figure 7 is obtained:

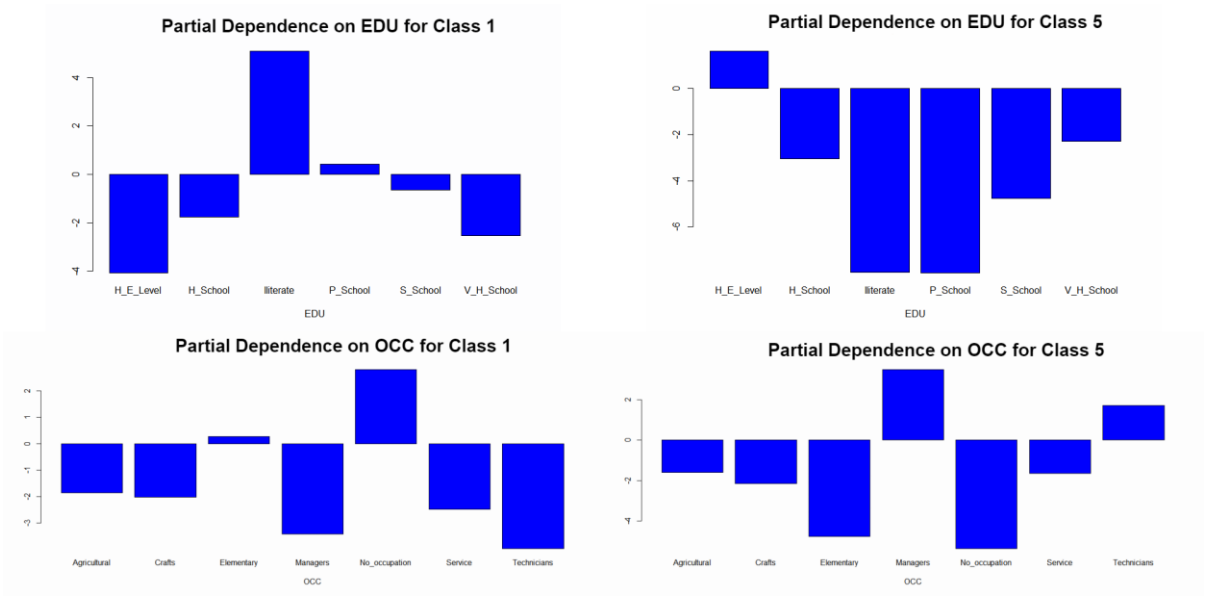

Figure 7: Partial dependence plots

- When the EDU variable receives the "Illiterate" value, it tends to predict Class 1 (lowest income level) more strongly than the other values.
- When the EDU variable receives the "H_E_Level" value, it tends to predict Class 5 (highest income level) more strongly than the other values.
- When the OCC variable receives the "No occupation" value, it tends to predict Class 1 (lowest income level) more strongly than the other values.
- When the OCC variable receives the "Managers" value, it tends to predict Class 5 (highest income level) stronger than the other values.

3.3.9 Obtaining a single tree from the forest

	left daughter	right daughter	split var	split point	status	prediction
1	2	3	OCC	55	1	<NA>
2	4	5	SEC	40	1	<NA>
3	6	7	AGE	3	1	<NA>
4	8	9	AGE	7	1	<NA>
5	10	11	SEC	64	1	<NA>
6	12	13	EDU	1	1	<NA>
7	14	15	AGE	24	1	<NA>
8	16	17	STA	5	1	<NA>
9	18	19	EDU	4	1	<NA>
10	20	21	MST	1	1	<NA>
11	22	23	STA	26	1	<NA>
12	24	25	OCC	8	1	<NA>
13	26	27	SEC	68	1	<NA>
14	28	29	STA	10	1	<NA>
15	30	31	RGN	62	1	<NA>
16	32	33	RGN	30	1	<NA>
17	34	35	EDU	12	1	<NA>
18	36	37	AGE	16	1	<NA>
19	38	39	EDU	26	1	<NA>
20	40	41	EDU	3	1	<NA>
21	42	43	OCC	5	1	<NA>
22	44	45	STA	8	1	<NA>
23	46	47	AGE	1	1	<NA>
24	48	49	MST	1	1	<NA>
25	50	51	RGN	47	1	<NA>
26	52	53	RGN	23	1	<NA>
27	54	55	MST	1	1	<NA>
28	56	57	EDU	1	1	<NA>
29	58	59	STA	16	1	<NA>
30	60	61	SEC	402	1	<NA>
31	62	63	EDU	17	1	<NA>
32	64	65	EDU	20	1	<NA>
33	66	67	EDU	11	1	<NA>
34	68	69	AGE	2	1	<NA>
35	70	71	RGN	47	1	<NA>
36	72	73	SEC	8	1	<NA>
37	74	75	SEC	8	1	<NA>
38	76	77	GEN	1	1	<NA>
39	78	79	INT	1	1	<NA>
40	80	81	OCC	5	1	<NA>
41	82	83	RGN	55	1	<NA>
42	84	85	EDU	2	1	<NA>
43	86	87	INT	1	1	<NA>
44	88	89	MST	1	1	<NA>
45	90	91	GEN	1	1	<NA>
46	92	93	SEC	385	1	<NA>
47	94	95	GEN	1	1	<NA>
48	96	97	SEC	402	1	<NA>
49	0	0	<NA>	0	-1	5_Quantile
50	98	99	AGE	1	1	<NA>

Figure 8: RF tree output from a single tree

In Figure 8, there is a section of the tree output obtained from a single tree. Here, the value "-1" in the "Status" column refers to the terminal node. The first four trees of the RF are visualized in Figure 9 and Figure 10:

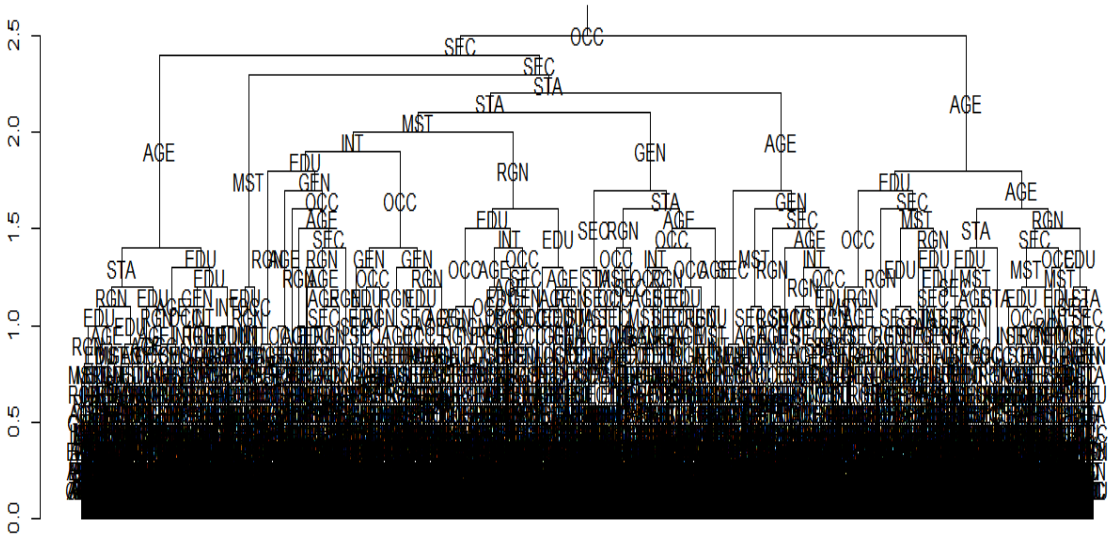

Figure 9: First tree of RF

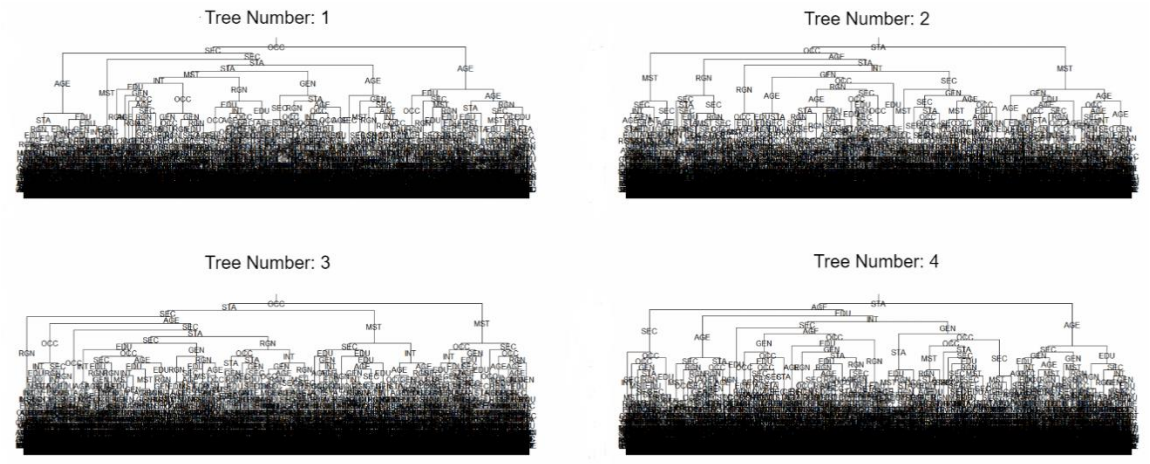

Figure 10: The first four trees of RF

4. Conclusion

With today's growing and expanding volume of data, it is important for economic and social decision-makers to reach new information and hidden patterns from official statistics and make predictions for the future. Currently, traditional statistical methods are insufficient to extract meaningful information from official statistics. Analyzing SILC data, which is a large and comprehensive official statistical data with data mining methods will facilitate the development of new data mining methods on official statistics and increase the usefulness of the data mining methods addressed for the current official statistics. For this purpose, SILC cross-sectional data provided from TurkStat for the year 2015 were analyzed using appropriate data mining methods. The welfare status of the individuals was investigated by determining the variables affecting individual income. For this purpose, the socio-economic profiles of individuals were determined by using LCA and k-modes clustering methods. In addition, by using the RF method, the importance level of the variables affecting the individual income was determined and a classification model was obtained for the welfare status of the individuals in the model, where individual income was handled as a dependent variable. Looking at the results of the cluster definitions in Appendix II and Appendix III; 6-2, 4-1, 1-9, 3-5, 9-6 from LCA and k-mod class/cluster pairs have close similarities respectively, while 2-8, 5-4 pairs have low similarities. The k-mode cluster definitions 3, 7 and 10 do not show similarity with any of the LCA class definitions and are not consistent. From these results, it was seen that in the categorical SILC data LCA method gave more consistent results than k-modes method and it can be applied successfully in such official data and produce meaningful results. We conclude that the RF method is a feasible data mining method that produces meaningful results for SILC data. In the study of [10], similar to our study, the income levels of the poor and the rich were predominantly determined by variables such as age, education, economic activity and the sector in which the head of household worked. The study of [9] found that LCA best predict the rates of rich and poor households. This study is consistent with our finding that the LCA method provided more consistent economic interpretation. While we restricted ourselves to the income and welfare status of individuals in SILC data, for future research, the methods used in this study can be used to develop other methods in studies related to official data. In this study, where LCA, k-modes and RF methods are used, the data used to be a nominal cross-sectional data for a given year, the number of variables considered,

and the algorithm parameters constitute a limitation. Despite these limitations, the results are consistently clustered around variables in connection with income, and consistent socio-economic profiles are obtained throughout the country.

5. Constraints/limitation of the Study

There are some limitations in this study. While we restricted ourselves to the income and welfare status of individuals in SILC data, for future research, the methods used in this study can be used to develop other methods in studies related to official data. In this study, where LCA, k-modes and RF methods are used, the data used to be a nominal cross-sectional data for a given year, the variables considered, and the algorithm parameters constitute a limitation. The number of classes preferred in LCA analysis and the number of clusters in k-modes analysis constitute a constraint. It is a limitation that the data partition process is 70/30 ratio for train and test data before RF analysis. Despite these limitations, the results are consistently clustered around variables in connection with income, and consistent socio-economic profiles are obtained throughout the country.

6. Recommendations

Future researches may be performed using panel data of official data or categorical/continuous cross-sectional data of different years and using different algorithms appropriate to the data. In addition, SES (socio-economic status) index of individuals can be determined and their results can be compared with existing socio-economic profiles. With this study, how data mining methods can be applied to a comprehensive official data such as SILC is shown in detail as a data mining process. These results will guide decision-makers through the socio-economic profiles of individuals.

Acknowledgements

The views expressed in this article are only those of the authors and do not necessarily reflect those of the TurkStat or any other person or organization.

References

[1] H. Hassani, G. Saporta and E.S. Silva. “Data Mining and Official Statistics: The Past, The Present & The Future.” Big Data, vol. 2(1), pp. 34BD-BD43, 2014.

[2] V.S. Arora, M. Karanikolos, A. Clair, A. Reeves, D. Stuckler, M. McKee. “Data Resource Profile: The European Union Statistics on Income and Living Conditions (EU-SILC).” International Journal of Epidemiology, vol. 44(2), pp. 451-61, 2015.

[3] S. Sumathi and S.N. Sivanandam. Introduction to Data Mining and Its Applications. New York: Springer, 2006.

[4] H. Hassani, S. Gheitanchi and M.R. Yeganegi. “On the Application of Data Mining to Official Data.”

Journal of Data Science, vol. 8, pp. 75-89, 2010.

[5] A. Barrett and Y. McCarthy. (2007, Aug.). "The Earnings of Immigrants in Ireland: Results from the 2005 EU Survey of Income and Living Conditions." IZA Discussion Paper No. 2990. Available at SSRN: https://ssrn.com/abstract=1012371 [Oct. 22, 2019].

[6] L. Stejskal and J. Stávková. "Living Conditions of Czech Farmers According to the EU Statistics on Income." Agricultural Economics: Zemědělská ekonomika, vol. 56 (7), pp. 310-316, 2010.

[7] H. Řezanková. "Cluster Analysis and Categorical Data." Statistika, vol. 3, pp. 216-232, 2009.

[8] H. Řezanková. "Cluster Analysis of Economic Data." Statistika: Statistics and Economy Journal, vol. 94(1), pp. 73-86, 2014.

[9] Z. Šulc and H. Řezanková. "Evaluation of Recent Similarity Measures for Categorical Data", in Proc. 17th AMSE , 2014. pp. 249-258.

[10] N. Birčiaková, I. Antošová, F. Balák. "Determinants of Czech Disposable Household Income and Related Housing Quality." Acta Universitatis, vol. 65(2), pp. 601-610, 2017.

[11] P. Brunori, P. Hufe, D.G. Mahler. (2019, Feb.). "The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees and Forests." Working Paper http://www.unicaldine.it/research/BHM_2017.pdf, [Oct. 22, 2019].

[12] A. Alfons and M. Templ. "Estimation of Social Exclusion Indicators from Complex Surveys: The R Package laeken." Journal of Statistical Software, vol. 54(15), pp. 1-25, 2013.

[13] P.F. Lazarfeld and N.W. Henry. Latent Structure Analysis. Boston: Houghton Mifflin, 1968.

[14] Z. Huang and M. Ng. "A Fuzzy k-Modes Algorithm for Clustering Categorical Data." IEEE Transactions on Fuzzy Systems, vol. 7(4), pp. 446-452, 1999.

[15] L Breiman. "Random Forests." Machine Learning, vol. 45(1), pp. 5-32, 2001.

Appendix I: Latent Class Probabilities

Table 18: Prior class probabilities based on LCA model

Manifest Variable	Category	Conditional item response	Class 1 (14.39%)	Class 2 (9.92%)	Class 3 (11.42%)	Class 4 (14.27%)	Class 5 (8.62%)	Class 6 (15.17%)	Class 7 (5.52%)	Class 8 (6.4%)	Class 9 (5.53%)	Class 10 (8.77%)

		(column) probabilities									
STA	Regular employee	Pr(1)	0.0000	**0.5231**	**0.8430**	**0.9151**	0.0270	0.0000	0.3500	0.0000	**0.6904**	**0.8994**
	Casual employee	Pr(2)	0.0000	0.0198	0.0400	0.0006	0.0200	0.0000	**0.5971**	0.0000	0.1408	0.0522
	Employer	Pr(3)	0.0000	0.1434	0.0376	0.0558	0.0282	0.0000	0.0014	0.0000	0.0003	0.0057
	Self-employed	Pr(4)	0.0000	0.3008	0.0793	0.0283	**0.8255**	0.0000	0.0388	0.0000	0.1465	0.0346
	Unpaid family worker	Pr(5)	0.0000	0.0128	0.0000	0.0002	0.0992	0.0000	0.0127	0.0000	0.0220	0.0082
	Inactive	Pr(6)	**1.0000**	0.0000	0.0000	0.0000	0.0000	**1.0000**	0.0000	**1.0000**	0.0000	0.0000
OCC	Managers, Professionals	Pr(1)	0.0000	0.1378	0.0095	**0.6248**	0.0000	0.0000	0.0000	0.0000	0.0049	0.0640
	Technicians and associate professionals	Pr(2)	0.0000	0.0679	0.0996	0.2762	0.0000	0.0000	0.0000	0.0000	0.0139	0.1910
	Service and sales workers	Pr(3)	0.0000	**0.5962**	0.0060	0.0935	0.0000	0.0000	0.0234	0.0000	**0.4701**	**0.3451**
	Skilled agricultural and fishery workers, Plant and machine operators and assemblers	Pr(4)	0.0000	0.0000	**0.4118**	0.0035	**0.9842**	0.0000	0.1198	0.0000	0.0217	0.1056
	Crafts and related trades workers	Pr(5)	0.0000	0.0875	0.3595	0.0020	0.0000	0.0000	0.3971	0.0000	0.1080	0.1694
	Elementary	Pr(6)	0.000	0.11	0.113	0.000	0.01	0.000	**0.45**	0.00	0.38	0.12

	occupations		0	05	6	0	58	0	**97**	00	13	49
	No occupation	Pr(7)	**1.000 0**	0.00 00	0.000 0	0.000 0	0.00 00	**1.000 0**	0.00 00	**1.00 00**	0.00 00	0.00 00
RGN	Marmara	Pr(1)	**0.300 1**	**0.24 35**	**0.378 3**	**0.278 7**	0.15 53	**0.200 0**	0.10 64	**0.24 93**	**0.28 16**	**0.31 93**
	Aegean	Pr(2)	0.156 5	0.12 32	0.155 4	0.130 6	0.15 94	0.129 0	0.07 89	0.12 43	0.19 19	0.13 57
	Central Anatolia	Pr(3)	0.186 9	0.15 68	0.159 5	0.194 1	0.11 69	0.121 5	0.10 18	0.16 47	0.12 74	0.16 94
	Mediterranean	Pr(4)	0.109 1	0.13 30	0.092 9	0.113 2	0.08 74	0.119 0	0.15 38	0.13 50	0.13 54	0.10 00
	Black Sea Region	Pr(5)	0.138 0	0.10 45	0.109 9	0.116 5	**0.21 39**	0.157 3	0.08 05	0.08 04	0.12 76	0.10 08
	Eastern Anatolia	Pr(6)	0.065 3	0.12 84	0.045 2	0.108 8	0.20 07	0.154 8	**0.26 17**	0.12 29	0.06 28	0.08 53
	Southeastern Anatolia	Pr(7)	0.044 0	0.11 07	0.058 8	0.058 0	0.06 63	0.118 4	0.21 68	0.12 34	0.07 33	0.08 95
INT	Yes	Pr(1)	0.456 3	0.46 01	**0.540 8**	**0.853 9**	0.19 60	0.178 1	0.09 98	**0.53 26**	0.42 37	**0.53 27**
	No	Pr(2)	**0.543 7**	**0.53 99**	0.459 2	0.146 1	**0.80 40**	**0.821 9**	**0.90 02**	0.46 74	**0.57 63**	0.46 73
MST	Married	Pr(1)	**0.865 2**	**0.94 51**	**0.935 7**	**0.808 3**	**0.89 85**	0.381 0	**0.74 65**	0.20 67	**0.80 21**	0.20 82
	Other	Pr(2)	0.134 8	0.05 49	0.064 3	0.191 7	0.10 15	**0.619 0**	0.25 35	**0.79 33**	0.19 79	**0.79 18**
AGE	15-24	Pr(1)	0.000 0	0.00 50	0.004 6	0.026 6	0.01 77	0.010 0	0.22 43	**0.67 63**	0.00 36	**0.58 28**
	25-34	Pr(2)	0.036 0	0.23 78	0.304 8	**0.397 8**	0.08 44	0.056 5	**0.27 99**	0.22 24	0.18 78	0.37 59
	35-44	Pr(3)	0.054 6	**0.36 39**	**0.384 9**	0.341 2	0.20 27	0.100 0	0.25 72	0.06 72	**0.42 26**	0.04 13
	45-54	Pr(4)	0.230 0	0.27 08	0.252 3	0.178 1	**0.29 05**	0.102 1	0.16 88	0.03 17	0.26 53	0.00 00
	55-64	Pr(5)	**0.403 0**	0.09 73	0.050 5	0.049 4	0.25 37	0.133 1	0.05 64	0.00 24	0.07 97	0.00 00
	65+	Pr(6)	0.276 3	0.02 52	0.002 9	0.006 9	0.15 09	**0.598 3**	0.01 34	0.00 00	0.04 09	0.00 00
GEN	Male	Pr(1)	**0.757 2**	**0.95 82**	**0.940 5**	**0.641 0**	**0.82 63**	0.310 0	**0.95 52**	0.49 05	0.20 29	**0.68 50**
	Female	Pr(2)	0.242	0.04	0.059	0.359	0.17	**0.690**	0.04	**0.50**	**0.79**	0.31

			8	18	5	0	37	**0**	48	**95**	**71**	50
SEC	Agriculture	Pr(1)	0.0000	0.0022	0.0097	0.0017	**0.9392**	0.0000	0.1867	0.0000	0.0607	0.0028
	Industry	Pr(2)	0.0000	0.0657	**0.5415**	0.0914	0.0016	0.0000	0.0890	0.0000	0.1941	0.2913
	Construction	Pr(3)	0.0000	0.0351	0.1399	0.0216	0.0000	0.0000	**0.5527**	0.0000	0.0003	0.0408
	Trade	Pr(4)	0.0000	**0.4497**	0.0459	0.0506	0.0000	0.0000	0.0437	0.0000	0.0856	0.2383
	Transportation	Pr(5)	0.0000	0.0104	0.1493	0.0238	0.0592	0.0000	0.0477	0.0000	0.0032	0.0348
	Service	Pr(6)	0.0000	0.3503	0.0596	**0.4307**	0.0000	0.0000	0.0716	0.0000	**0.4344**	**0.3169**
	Financial	Pr(7)	0.0000	0.0048	0.0008	0.0433	0.0000	0.0000	0.0000	0.0000	0.0010	0.0123
	Public/Administration	Pr(8)	0.0000	0.0750	0.0419	0.2242	0.0000	0.0000	0.0087	0.0000	0.0014	0.0153
	Health	Pr(9)	0.0000	0.0068	0.0114	0.1127	0.0000	0.0000	0.0000	0.0000	0.2192	0.0476
	No sector	Pr(10)	**1.0000**	0.0000	0.0000	0.0000	0.0000	**1.0000**	0.0000	**1.0000**	0.0000	0.0000
EDU	lliterate / not a graduate	Pr(1)	0.0251	0.0365	0.0121	0.0000	0.1790	**0.6672**	0.2259	0.0646	0.1892	0.0337
	Primary school	Pr(2)	**0.5147**	**0.4393**	**0.4919**	0.0010	**0.6499**	0.3245	**0.4041**	0.0073	**0.6313**	0.0012
	Secondary school	Pr(3)	0.1195	0.1899	0.1930	0.0070	0.0877	0.0048	0.2900	**0.4192**	0.0782	**0.4096**
	High school	Pr(4)	0.0968	0.1690	0.1029	0.0900	0.0391	0.0000	0.0471	0.2066	0.0595	0.1810
	Vocational high school	Pr(5)	0.0931	0.1190	0.1675	0.0832	0.0270	0.0034	0.0257	0.1514	0.0327	0.1897
	Higher education level	Pr(6)	0.1508	0.0463	0.0326	**0.8188**	0.0173	0.0000	0.0072	0.1509	0.0091	0.1848
IG	1. weighted 20% quantile	Pr(1)	0.0322	0.0200	0.0102	0.0167	0.2159	**0.5026**	0.2900	**0.7695**	0.3270	0.2397
	2. weighted 20% quantile	Pr(2)	0.2326	0.1369	0.0799	0.0320	0.1957	0.3914	**0.4015**	0.1640	**0.4115**	0.2663
	3. weighted 20% quantile	Pr(3)	**0.3687**	0.2120	0.2508	0.0465	0.1550	0.0805	0.1811	0.0492	0.1844	**0.3288**

4. weighted 20% quantile	Pr(4)	0.2812	**0.3353**	**0.4320**	0.1366	**0.2228**	0.0229	0.1209	0.0095	0.0699	0.1525
5. weighted 20% quantile	Pr(5)	0.0852	0.2957	0.2271	**0.7683**	0.2105	0.0026	0.0065	0.0077	0.0073	0.0127

Appendix II: Definitions of Latent Classes from 10 Class Model Solution

Table 19: Latent class definitions

Class	Size	Definition
1: "Middle-income and low-educated older men, second largest slice"	14% (n = 5185)	Inactive employment status No occupation In Marmara Region No internet connection Married Aged 55-64 Male No working sector Primary school educated In the third 20% income group
2: "Regular employee in the trade sector"	10% (n = 3574)	Regular employee Service/sales worker In Marmara Region No internet connection Married Aged 35-44 Male Employed in trade sector Primary school educated In the fourth 20% income group
3: "Regular employee in the industry sector"	11% (n = 4115)	Regular employee Plant and machine operator/assembler In Marmara Region Having internet connection Married Aged 35-44 Male Employed in industry sector Primary school educated In the fourth 20% income group

4: "Highest income and most educated"	14% (n = 5142)	Regular employee Manager/professional In Marmara region Having internet connection Married Aged 25-34 Male Employed in service sector Having higher education level In the highest 20% income group
5: "Self-employed in the agricultural sector in the Black Sea region"	9% (n = 3106)	Self-employed Agricultural occupation In Black Sea region No internet connection Married Aged 45-54 Male Employed in agriculture sector Primary school educated In the fourth 20% income group
6: "Lowest income and most uneducated, largest slice"	15% (n = 5466)	Inactive employment status No occupation In Marmara region No internet connection Not married Aged 65 years and over Female No working sector llliterate/not a graduate In the lowest 20% income group
7: "Low-income, casual employee in construction sector in Eastern Anatolia region"	6% (n = 1988)	Casual employee Elementary occupation Eastern Anatolia region No internet connection Married Aged 25-34 Male

		Employed in construction sector Primary school educated In the second 20% income group
8: "Inactive young women with the lowest income"	6% (n = 2306)	Inactive employment status No occupation In Marmara region Having internet connection Not married Aged 15-24 Female No working sector Secondary school educated In the lowest 20% income group
9: "Low-income, middle-aged married women"	6% (n = 1992)	Regular employee Service/sales worker In Marmara region No internet connection Married Aged 35-44 Female Employed in service sector Primary school educated In the second 20% income group
10: "Middle-income, not married young men"	9% (n = 3160)	Regular employee Service/sales worker In Marmara region Having internet connection Not married Aged 15-24 Male Employed in service sector Secondary school educated In the third 20% income group

Appendix III: Definitions of Clusters from 10 Cluster k-modes Model Solution

Table 20: k-modes cluster definitions

Cluster	Size	Definition
1: "Highest income and most educated"	9% (n = 3405)	Regular employee Manager/professional In Black Sea region Having internet connection Married Aged 25-34 Female Service and sales worker Having higher education level In the highest 20% income group
2: "Lowest income and most uneducated, largest slice"	22% (n = 7840)	Inactive employment status No occupation In Marmara Region No internet connection Not married Aged 65 years and over Female No working sector llliterate/not a graduate In the lowest 20% income group
3: "Highest income, young regular employee in the industry sector in the Mediterranean region"	8% (n = 2893)	Regular employee Crafts and related trades worker In Mediterranean Region No internet connection Married Aged 15-24 Male Employed in industry sector Secondary school educated In the highest 20% income group
4: "Self-employed in the agricultural sector in the Eastern Anatolia region"	13% (n = 4723)	Self-employed Agricultural occupation In Eastern Anatolia Region

		No internet connection Married Aged 35-44 Male Employed in agriculture sector Primary school educated In the second 20% income group
5: "Regular employee in the industry sector"	10% (n = 3587)	Regular employee Agricultural occupation In Marmara region No internet connection Married Aged 25-34 Male Employed in industry sector Primary school educated In the fourth 20% income group
6: "Low-income, middle-aged married women"	5% (n = 1653)	Regular employee Service/sales worker In Central Anatolia Region No internet connection Married Aged 45-54 Female Employed in service sector Primary school educated In the lowest 20% income group
7: "Most educated young men with high-income"	5% (n = 1899)	Regular employee Technician/associate professional In Marmara region Having internet connection Not married Aged 25-34 Male Employed in service sector Having higher education level In the fourth 20% income group
8: "Low-income, not married regular employee young men in the trade sector"	4% (n = 1434)	Regular employee Service/sales worker

		In Central Anatolia Region No internet connection Not married Aged 25-34 Male Employed in trade sector Secondary school educated In the second 20% income group
9: "Middle-income and low-educated older men, second largest slice"	15% (n = 5390)	Inactive employment status No occupation In Marmara region No internet connection Married Aged 55-64 Male No working sector Primary school educated In the third 20% income group
10: "Most educated middle-aged married men with the highest income"	9% (n = 3212)	Regular employee Manager/professional In Marmara region Having internet connection Married Aged 35-44 Male Employed in service sector Having higher education level In the highest 20% income group