IBIMA Publishing
Journal of Eastern Europe Research in Business and Economics
http://ibimapublishing.com/articles/JEERBE/2020/359361/
Vol. 2020 (2020), Article ID 359361, 17 pages, ISSN: 2169-0367
DOI: 10.5171/2020.359361

Research Article

Data Mining Techniques in The Analysis of The Causal Factors Regarding Innovation in The Private Sector at The Level of Europe

Ionuț Costinel NICA and Simona Liliana CRĂCIUNESCU (PARAMON)

Bucharest University of Economic Studies, Bucharest, Romania

Correspondence should be addressed to: Ionuț Costinel NICA; ionut.nica@csie.ase.ro

Received date: 2 March 2020; Accepted date: 1 September 2020; Published date: 4 December 2020

Academic Editor: Georgescu Irina Alexandra

Abstract

The rapid pace of growth of data quantities and the need to make sense of the information chaos have made it possible to introduce a new concept that encompasses these two aspects. Living in a world where we can identify a wealth of data and information, the data mining technique is a useful tool for analyzing a large data set. In this article, a data set has been analyzed in order to allow the observation of how innovation has different causal factors in the private sector. The analysis was carried out in several countries in Europe. The first step of the analysis was data processing and descriptive statistics. The next step was to analyze the main components to reduce the dimensionality of the analyzed data set. The third step of the analysis is the cluster technique in order to classify the variables into classes so as to ensure a minimum variability within the class and a maximum variability between classes. The last part of the analysis proposed in this article was the use of classification trees to see what factors influence an individual's decision to become an entrepreneur.

Keywords: Data Mining, Private Sector, R Studio, Opinion Influence, Big Data.

Introduction

The present work was elaborated to apply the reduction of the dimensionality of a certain data set in order to allow the observation of the way in which the innovation in the private sector, at the level of Europe (by country), has different causal factors (direct or indirect) that allow it to grow and develop, or otherwise, to decrease. These factors can be: the share of the population with higher education, the share of foreign doctoral students, the share of companies that have high-speed internet access, public or private expenses on research, the share of companies offering

Cite this Article as: Ionuț Costinel NICA and Simona Liliana CRĂCIUNESCU (PARAMON) (2020)," Data Mining Techniques in The Analysis of The Causal Factors Regarding Innovation in The Private Sector at The Level of Europe", Journal of Eastern Europe Research in Business and Economics Vol. 2020 (2020), Article ID 359361, DOI: 10.5171/2020.359361

training or innovating, the number of publications resulting from public sector collaboration, and private sector investments in research and higher education.

The rapid pace of growth of data quantities and the need to make sense of the information chaos have made it possible to introduce a new concept that encompasses these two aspects; namely the Big Data, the concept that refers to the exponential growth of structured and unstructured data.

Methodology

Big Data is any collection of datasets that are so big and complex that can no longer be processed using traditional means. Forrester analysts described the Big Data phenomenon as "the ability of a company to store, process and analyze all the data needed to function efficiently, make informed decisions, reduce risks and serve the needs of customers."

In 2001, analyst Doug Laney gave a definition of the concept of Big Data that included the three V's:

- Volume (if in 2000, a regular computer would have 10 GB of storage space, now Facebook generates 500 terabytes of new data daily, a Boeing 737 generates 240 terabytes during a secure flight over the United States, etc.)
- Velocity (user behavior can be reflected by accessing advertisements and various sites, and this happens quickly. Online game systems host millions of competing users, each producing a lot of data every second);
- Variety (Big Data does not only refer to numbers, character strings and login data, but also includes audio and video data, unstructured data from social media sites, and 3D data. The latter making databases and traditional unprepared analytic tools. Traditional databases were designed to operate on a single server, thus being expensive and finite in terms of capacity in the Big Data context).

SAS specialists added two other dimensions, namely variability and complexity. The first term refers to the inconsistency of the data, and the second refers to the existence of multiple sources of data that make their correlation necessary. Thus, Big Data refers to any complex set of unstructured data, which can no longer be managed with the help of traditional databases, with the difficulties of storing, analyzing, manipulating data, viewing and sharing.

The term Big Data was first mentioned in August 1999, when Steve Bryson et al. published the article "Visually exploring gigabyte data sets in real time" in "Communications of the ACM". One of the chapters of the paper is called "Big Data for scientific visualization"[**Error! Reference source not found.**].

In October 2003, Peter Lyman et al. of the Berkeley University published the paper "How much information?". It tries to quantify the total of new and original information created annually in the world and kept in physical format (books, magazines, etc.), on DVDs and CDs or on magnetic disks. It is found that in 1999, around 1.5 exabytes of unique information were produced worldwide, that is to say about 250 megabytes produced by one person. [?]

R is a software that works on the basis of a programming language, providing a variety of statistical analysis and data mining techniques. The popularity of this software has increased lately due to the variety of facilities it offers in terms of analysis techniques. Thus, R encompasses linear and nonlinear models, classical statistical tests, time series analysis techniques, classification methods, clustering and so on. Through the packages that add functions to the program, the facilities are constantly increased.

Results and Discussions

The study of dimensionality reduction is based on a sample from 2017 consisting of indicators from 28 countries:

	A	B	C	D	E	F	G	H	I	J	K	L
1	cod	Country	I112	I123	I131	I211	I221	I223	I312	I322	I323	I333
2	BE	Belgium	45.7	41.80401558	23	0.74	1.73	34	45.1444	82.8389	0.07867	2.76273
3	BG	Bulgaria	32.8	6.281481481	10	0.21	0.57	8	14.7502	3.63444	0.01684	6.97138
4	CZ	Czech Rep	32.6	14.75876658	10	0.64	1.03	22	25.737	21.9825	0.02996	2.61842
5	DK	Denmark	45.3	33.40376295	31	0.97	1.89	28	39.9838	186.43	0.02666	8.02711
6	DE	Germany	30.5	9.123343527	12	0.94	2	29	49.0903	65.3235	0.11625	6.21725
7	EE	Estonia	41.2	11.96611366	12	0.61	0.66	13	15.034	15.1982	0.03587	3.80452
8	IE	Ireland	53.5	28.43691149	15	0.35	0.83	30	52.5187	41.047	0.00958	0.98145
9	ES	Spain	41	15.48320989	20	0.55	0.64	23	25.5234	23.1481	0.03362	3.0094
10	FR	France	44	40.0527643	9	0.78	1.43	20	41.6205	43.7731	0.03566	2.9104
11	HR	Croatia	32.8	3.927779538	6	0.46	0.38	22	30.8386	16.7038	0.03349	0.84326
12	IT	Italy	25.6	14.16213919	5	0.5	0.75	12	34.5953	24.5609	0.01188	6.4071
13	CY	Cyprus	56.2	14.3081761	3	0.27	0.17	22	31.1097	17.682	0.00153	3.19336
14	LV	Latvia	42.1	11.38528139	22	0.33	0.11	12	18.9673	0.50788	0.05174	1.77142
15	LT	Lithuania	54.9	4.603580563	21	0.55	0.3	10	24.0038	3.46193	0.09041	1.45858
16	LU	Luxembou	51.5	86.99472759	21	0.6	0.64	29	54.3491	39.9133	0.00729	12.6792
17	HU	Hungary	30.4	11.63457599	12	0.29	0.89	16	15.2232	33.9759	0.02917	0.92793
18	MT	Malta	34	54	12	0.23	0.39	23	30.7782	2.22017	0.00227	20.2664
19	NL	Netherlan	45.2	40.10094972	22	0.87	1.16	22	32.51	106.896	0.08277	3.64967
20	AT	Austria	39.7	28.262164	12	0.87	2.2	37	46.0554	84.9379	0.04638	7.06213
21	PL	Poland	43.5	1.957060472	11	0.32	0.63	12	11.3859	6.05786	0.01846	5.99264
22	PT	Portugal	35	25.63483736	25	0.64	0.61	23	37.8069	15.8587	0.0124	4.43033
23	RO	Romania	24.8	3.779271968	13	0.21	0.27	5	8.84043	4.85822	0.0328	0.81329
24	SI	Slovenia	43	9.652509653	16	0.49	1.51	27	33.1914	52.3208	0.04942	2.98634
25	SK	Slovakia	33.4	9.127990299	9	0.39	0.4	20	22.4371	9.58304	0.03669	1.09644
26	FI	Finland	40.7	21.14508117	26	0.91	1.81	34	37.264	92.7595	0.04686	4.43962

Figure 1: The table of the indicators from 28 countries

In the figure below, the name of each indicator considered in the paper is presented in details:

	A	B
1	**Indicator**	**Name**
2	I112	Share of population with higher education (group 25 -34 years)
3	I123	The share of foreign doctoral students
4	I131	The share of companies that have high speed internet access
5	I211	Public expenditure on Research % GDP
6	I221	Private expenditure on research development
7	I223	Share of companies offering trainig (IT Competencies)
8	I312	Share of companies that have innovated (non-technological innovation)
9	I322	Number of publications resulting from public-private collaboration
10	I323	Private sector investments in research and higher education
11	I333	The number of designs (aesthetic appearance of a product) registered

Figure 2: Descriptions of the name of each indicator

In order to perform the analysis of the chosen data set by applying the dimensionality reduction, the R Studio software was used. The first step in the analysis was to create descriptive statistics. Thus, in the figure below, the correlations between the data presented can be seen. The more the blue increases in intensity, the

stronger the correlation between the indicators gets. The empty squares in the figure represent that the probability is greater than 0.05, which means that there is a statistically insignificant coefficient between certain indicators. For example, the indicator I112 (weight of the population with higher education) has very strong correlations with all the indicators in the data set:

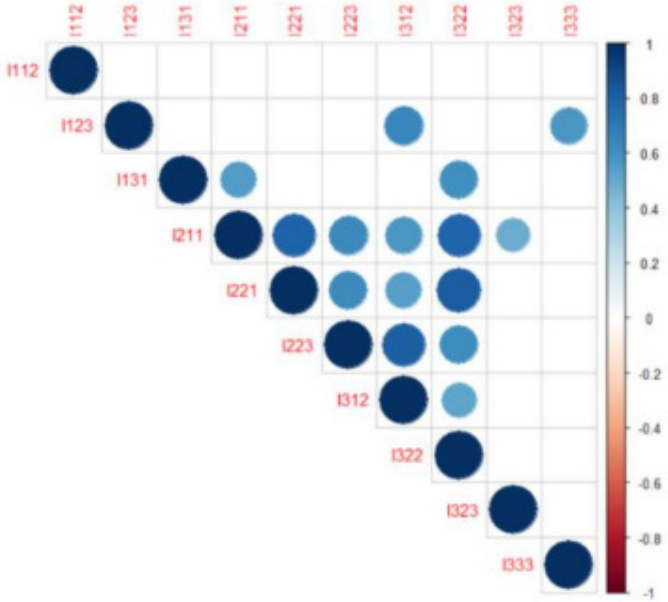

Figure 3: The correlation matrix between the indicators

In the correlation matrix of the figure below, it can be said that there is a strong direct (positive) correlation between I211 (Public expenditure on Research % GDP) and I221 (Private expenditure on research development) of 0.80, and a strong (positive) correlation directly between I223 (Share of training companies) and I312 (Share of innovating companies) of 0.81. Further in the analysis, the correlation matrix was created using the chart. The correlation () function in the Performance Analytics package that allows the correlation chart type is to be made.

The values above the main diagonal, in the figure below, represent the values of the correlation coefficients, and the stars represent the significance levels (three stars → significant correlation coefficient, with an associated probability of 0%, two stars → less significant correlation coefficient, one star → very low correlation coefficient). The figure below also presents the histograms related to the analyzed indicators, those inclined to the left (related to I123, I131, I211, I221, I322, I323, I333) have the most extreme values to the right, and the indicator I312 has a distribution inclined to the right, with several extreme values to the left. Last but not least, indicator I112 has a distribution close to the normal one.

The graphs below the main diagonal, in the figure below, represent scatter plots detailing the relationship between two quantitative variables. For example, the first scatter plot represents the relationship between indicator I112 (weight of population with higher education) and indicator I123 (weight of foreign doctoral students). Analyzing each scatter plot, a general tendency of growth was observed, more precisely, the association between variables is positive. Also, there are countries (represented by circles) that are positioned above the average or very low

below the average. These countries represent the tools of the dataset.

The Correlation Matrix Chart graph also shows the intensity of the connections between the analyzed variables, so in the 6th scatter plot, one can see a very strong connection (the dependency line is drawn at a distance of about 45 degrees) between the indicator I131 (the share of companies that have high-speed Internet access) and I211 (public spending with% GDP research).

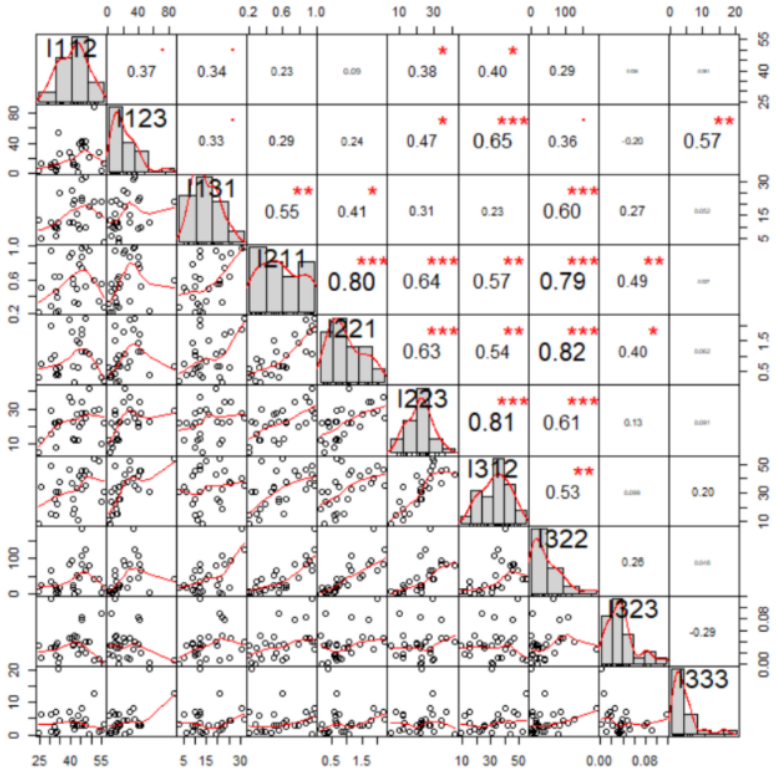

Figure 4: The correlation matrix chart

The second step in the analysis was represented by the analysis of the main components, because following the analysis of the correlation above, it was found that there is an informational overlap. The princomp () function will be used to reduce dimensionality by approaching spectral decomposition.

Next, the eigen values of the correlation matrix were determined:

```
eigen() decomposition
$values
 [1] 4.5672943 1.8969684 1.0709738 0.9001628 0.5860366 0.3303287 0.2486787 0.1628657 0.1248762 0.1118149

$vectors
             [,1]         [,2]         [,3]         [,4]         [,5]         [,6]         [,7]         [,8]         [,9]        [,10]
 [1,] -0.2083605  0.13874573  0.778534294  0.110871894  0.122077712  0.52661728 -0.100837196  0.07024968  0.04562437 -0.09753892
 [2,] -0.2700002  0.51305446  0.006282385  0.135969494  0.214314983 -0.21364192  0.568916639 -0.11853011 -0.40100780 -0.23505591
 [3,] -0.2922331 -0.10310216  0.122977702  0.722314502 -0.109434822 -0.46437530 -0.250673589 -0.12433800  0.23201329 -0.08579201
 [4,] -0.4083575 -0.22472796 -0.124222040  0.006720359  0.003151212 -0.02167898  0.121451577  0.86351886 -0.04224717 -0.06771389
 [5,] -0.3860233 -0.19361196 -0.319923234 -0.109016952 -0.185336817  0.35461335  0.052443956 -0.31436583  0.21605676 -0.62522615
 [6,] -0.3858829  0.09746159  0.074011359 -0.408580828 -0.090425658 -0.25126638 -0.603906708 -0.09077635 -0.47235295 -0.04250973
 [7,] -0.3715200  0.22535099  0.079002820 -0.418872544  0.185471392 -0.25331837  0.117956382 -0.07252354  0.66287351  0.26897894
 [8,] -0.4076737 -0.11683742 -0.099357751  0.138850994 -0.388056762  0.28570891  0.216967289 -0.21911715 -0.19833971  0.65142627
 [9,] -0.1564888 -0.51906156 -0.057742189  0.040323865  0.782844987  0.03503573 -0.002676385 -0.21741955 -0.15749291  0.12235183
[10,] -0.0699748  0.52492735 -0.485812854  0.268800126  0.294651852  0.35355943 -0.400160752  0.11082130  0.09030170  0.13747289
```

Figure 5: Output of the eigen values of the correlation matrix

Then, the authors of this paper calculated the number of the main components that will have lambda greater than 1, complying with Kaiser's Criterion, and observed the

presence of three main components in the chosen data set, following the analysis of the variances representing the square elevation of the standard deviations of each main component.

```
> comp_principale
Call:
princomp(x = c2, cor = TRUE, scores = TRUE)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8    Comp.9   Comp.10
2.1371229 1.3773048 1.0348787 0.9487691 0.7655302 0.5747423 0.4986770 0.4035662 0.3533782 0.3343873

 10  variables and  28 observations.
```

Figure 6: Output of the princomp () function

In the figure above, following the application of the princomp () function, the standard deviations of the main components were displayed, which shows the percentages of information taken by each new component created. It is also observed that the first three main components take the highest percentage of information from the data set.

Using the summary () function, the simple proportions and cumulative variance ratios of the main components resulted in R. This confirms that only the first three main components will be kept in the analysis, the standard deviation of the first component is 2.1371229, of the second 1.3773048, and of the third 1.0348787, exceeding the value 1, according to the criterion to Kaiser. It is also found that the first main component retained in the analysis accounts for 45.6% of the total information, the second holds only 18.9% of the total information, and the third holds only 10.7% of the total information.

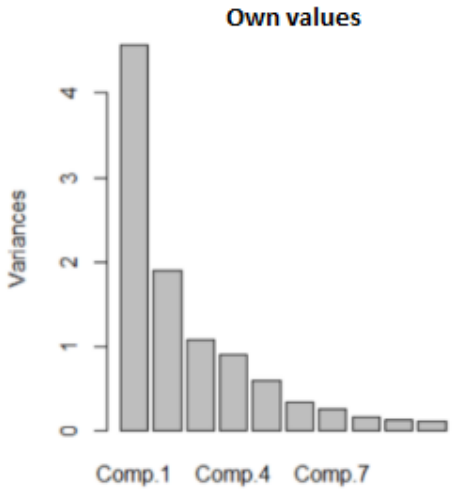

Figure 7: The variances of the main component

In the figure above, the three main components (Comp.1, Comp. 4 and Comp. 7) that will be kept in the analysis can been seen graphically. The results are based on the calculated values.

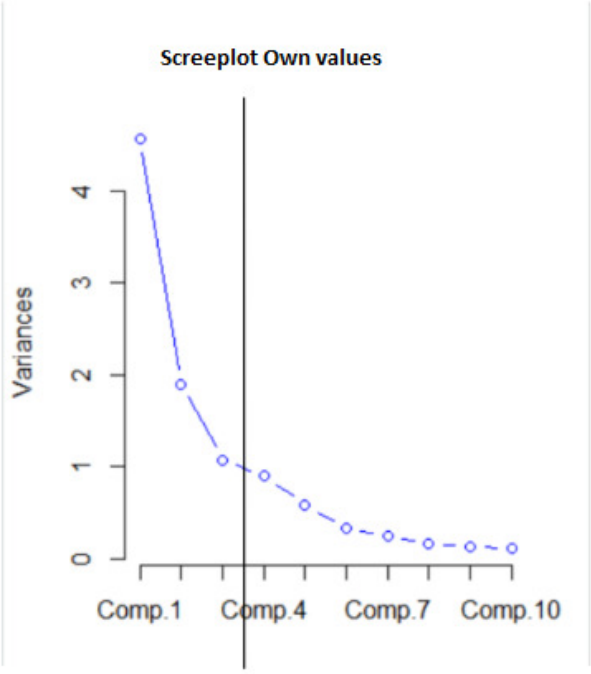

Figure 8: The screeplot of the main component

In order to determine the optimal number of the main components that should be kept in the analysis, a line-type screeplot was developed. In the figure above, it is observed how the eigen values decrease, thus, the first variance passes to the value 4 (variance = 4.5672943), the second variance decreases to the value of approximately 2 (variance = 1.8969684), and the third component decreases to about 1 (variance = 1.0709738). Then, the slope changes and the values reach below 1, becoming insignificant in the analysis. Therefore, only the first three main components in the analysis will be analyzed.

```
> coef

Loadings:
     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
I112 -0.208  0.139  0.779  0.111  0.122  0.527  0.101
I123 -0.270  0.513         0.136  0.214 -0.214 -0.569 -0.119 -0.401 -0.235
I131 -0.292 -0.103  0.123  0.722 -0.109 -0.464  0.251 -0.124  0.232
I211 -0.408 -0.225 -0.124                       -0.121  0.864
I221 -0.386 -0.194 -0.320 -0.109 -0.185  0.355         -0.314  0.216 -0.625
I223 -0.386                -0.409         -0.251  0.604         -0.472
I312 -0.372  0.225         -0.419  0.185 -0.253 -0.118         0.663  0.269
I322 -0.408 -0.117          0.139 -0.388  0.286 -0.217 -0.219 -0.198  0.651
I323 -0.156 -0.519                 0.783                -0.217 -0.157  0.122
I333         0.525 -0.486  0.269  0.295  0.354  0.400  0.111         0.137

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
SS loadings       1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0    1.0     1.0
Proportion Var    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1     0.1
Cumulative Var    0.1    0.2    0.3    0.4    0.5    0.6    0.7    0.8    0.9     1.0
```

Figure 9: Output of the eigenvectors of the correlation matrix

In the figure above, one can observe the eigenvectors of the correlation matrix, more precisely the coefficients that are used in the construction of the main components.

```
> n<-cbind(c2, scoruri[,1:3])
> n
   I112      I123 I131 I211 I221 I223     I312       I322        I323      I333      Comp.1      Comp.2      Comp.3
1  45.7 41.804016   23 0.74 1.73   34 45.14440  82.8388567 0.078673975 2.7627327 -2.85102525 -0.634401855  0.31555726
2  32.8  6.281481   10 0.21 0.57    8 14.75016   3.6344402 0.016844894 6.9713815  3.11553737  0.302859053 -0.84768711
3  32.6 14.758767   10 0.64 1.03   22 25.73695  21.9825139 0.029960268 2.6184228  0.91129942 -0.506187354 -0.67563558
4  45.3 33.403763   31 0.97 1.89   28 39.98384 186.4295087 0.026662337 8.0271073 -3.83711507  0.013355223 -0.61955460
5  30.5  9.123344   12 0.94 2.00   29 49.09030  65.3234599 0.116249337 6.2172544 -2.08120593 -2.114371067 -1.98770172
6  41.2 11.966114   12 0.61 0.66   13 15.03398  15.1982151 0.035869824 3.8045195  1.65225901 -0.563160922  0.07336375
7  53.5 28.436911   15 0.35 0.83   30 52.51866  41.0470293 0.009575487 0.9814522 -0.57620391  1.171237046  2.02066548
8  41.0 15.483210   20 0.55 0.64   23 25.52341  23.1480988 0.033621318 3.0093984  0.61260192 -0.301440313  0.45848727
9  44.0 40.052764    9 0.78 1.43   20 41.62046  43.7731181 0.035655392 2.9104048 -0.79788350  0.277565466  0.08517536
10 32.8  3.927780    6 0.46 0.38   22 30.83859  16.7037769 0.033493860 0.8432611  1.81874555 -0.575280387 -0.05840775
11 25.6 14.162139    5 0.50 0.75   12 34.59531  24.5608912 0.011883829 6.4070989  1.87353267  0.524473734 -1.63864697
12 56.2 14.308176    3 0.27 0.17   22 31.10969  17.6820276 0.001530849 3.1933631  1.77410970  1.295891684  2.04821900
13 42.1 11.385281   22 0.33 0.11   12 18.96729   0.5078831 0.051741130 1.7714152  2.04582124 -0.751750762  0.99735289
14 54.9  4.603581   21 0.55 0.30   10 24.00380   3.4619350 0.090407165 1.4585808  1.08038145 -1.696791031  1.91892702
15 51.5 86.994728   21 0.60 0.64   29 54.34913  39.9133014 0.007286728 12.6791691 -2.08833915  4.078023623  0.54335941
16 30.4 11.634576   12 0.29 0.89   16 15.22325  33.9759432 0.029166999 0.9279319  2.08387865 -0.788589525 -0.54258579
17 34.0 54.000000   12 0.23 0.39   23 30.77816   2.2201747 0.002272594 20.2664417  1.13896332  4.097365638 -1.93057235
18 45.2 40.100950   22 0.87 1.16   22 32.50998 106.8959993 0.082767928 3.6496701 -2.00370535 -1.004548159  0.12978037
19 39.7 28.262164   12 0.87 2.20   37 46.05537  84.9379304 0.046379458 7.0621259 -2.63850921  0.006670333 -1.14524581
20 43.5  1.957060   11 0.32 0.63   12 11.38587   6.0578590 0.018459829 5.9926396  2.57059357  0.048408344  0.19102465
21 35.0 25.634837   25 0.64 0.61   23 37.80689  15.8586952 0.012396282 4.4303347  0.09168293  0.573626089 -0.06939830
22 24.8  3.779272   13 0.21 0.27    5  8.84043   4.8582224 0.032797101 0.8132872  3.72357883 -1.091332178 -0.75141307
23 43.0  9.652510   16 0.49 1.51   27 33.19138  52.3208157 0.049420034 2.9863413 -0.39525147 -0.791944558  0.14729716
24 33.4  9.127990    9 0.39 0.40   20 22.43714   9.5830418 0.036685072 1.0964387  2.09246283 -0.590783406 -0.01809398
25 40.7 21.145081   26 0.91 1.81   34 37.26398  92.7595098 0.046861256 4.4396204 -2.57170901 -0.840813744 -0.43290548
26 47.3 34.686098   32 0.98 2.26   25 35.10149 126.1798655 0.039532064 4.7113810 -3.37370109 -0.692396449 -0.16344944
27 47.2 42.949468   10 0.52 1.13   28 45.44738  67.2197642 0.020905360 3.0511216 -0.94315492  1.119532029  0.73258699
28 49.2 21.600103   19 0.95 1.08   42 43.25129  79.7854232 0.040209756 0.3950196 -2.42764460 -0.565216553  1.21950136
```

Figure 10: Output of the values of the components corresponding to each of the 29 countries

The table above shows the values of the components corresponding to each of the 28 countries, as well as the values of the indicators. For example, the value of component 1 for Germany (6th country in the dataset) is 1.65225901.

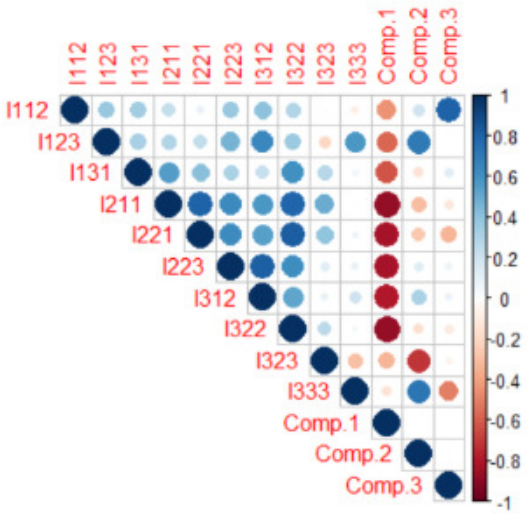

Figure 11: Correlation matrix

```
> matrice_corelatie<-rcorr(as.matrix(n))
> matrice_corelatie
        I112  I123  I131  I211  I221  I223  I312  I322  I323  I333 Comp.1 Comp.2 Comp.3
I112    1.00  0.37  0.34  0.23  0.09  0.38  0.40  0.29  0.03 -0.08  -0.45   0.19   0.81
I123    0.37  1.00  0.33  0.29  0.24  0.47  0.65  0.36 -0.20  0.57  -0.58   0.71   0.01
I131    0.34  0.33  1.00  0.55  0.41  0.31  0.23  0.60  0.27  0.05  -0.62  -0.14   0.13
I211    0.23  0.29  0.55  1.00  0.80  0.64  0.57  0.79  0.49 -0.03  -0.87  -0.31  -0.13
I221    0.09  0.24  0.41  0.80  1.00  0.63  0.54  0.82  0.40  0.06  -0.82  -0.27  -0.33
I223    0.38  0.47  0.31  0.64  0.63  1.00  0.81  0.61  0.13  0.09  -0.82   0.13   0.08
I312    0.40  0.65  0.23  0.57  0.54  0.81  1.00  0.53  0.10  0.20  -0.79   0.31   0.08
I322    0.29  0.36  0.60  0.79  0.82  0.61  0.53  1.00  0.26  0.05  -0.87  -0.16  -0.10
I323    0.03 -0.20  0.27  0.49  0.40  0.13  0.10  0.26  1.00 -0.29  -0.33  -0.71  -0.06
I333   -0.08  0.57  0.05 -0.03  0.06  0.09  0.20  0.05 -0.29  1.00  -0.15   0.72  -0.50
Comp.1 -0.45 -0.58 -0.62 -0.87 -0.82 -0.82 -0.79 -0.87 -0.33 -0.15   1.00   0.00   0.00
Comp.2  0.19  0.71 -0.14 -0.31 -0.27  0.13  0.31 -0.16 -0.71  0.72   0.00   1.00   0.00
Comp.3  0.81  0.01  0.13 -0.13 -0.33  0.08  0.08 -0.10 -0.06 -0.50   0.00   0.00   1.00

n= 28


P
        I112   I123   I131   I211   I221   I223   I312   I322   I323   I333   Comp.1 Comp.2 Comp.3
I112           0.0502 0.0730 0.2383 0.6476 0.0489 0.0334 0.1289 0.8970 0.6835 0.0176 0.3300 0.0000
I123   0.0502         0.0906 0.1328 0.2130 0.0119 0.0002 0.0568 0.2994 0.0015 0.0013 0.0000 0.9738
I131   0.0730 0.0906        0.0022 0.0290 0.1077 0.2442 0.0007 0.1610 0.7912 0.0004 0.4710 0.5187
I211   0.2383 0.1328 0.0022        0.0000 0.0003 0.0014 0.0000 0.0079 0.8915 0.0000 0.1090 0.5144
I221   0.6476 0.2130 0.0290 0.0000        0.0004 0.0030 0.0000 0.0351 0.7540 0.0000 0.1702 0.0853
I223   0.0489 0.0119 0.1077 0.0003 0.0004        0.0000 0.0005 0.5151 0.6469 0.0000 0.4959 0.6985
I312   0.0334 0.0002 0.2442 0.0014 0.0030 0.0000        0.0040 0.6168 0.3034 0.0000 0.1080 0.6792
I322   0.1289 0.0568 0.0007 0.0000 0.0000 0.0005 0.0040        0.1758 0.8090 0.0000 0.4133 0.6026
I323   0.8970 0.2994 0.1610 0.0079 0.0351 0.5151 0.6168 0.1758        0.1325 0.0820 0.0000 0.7626
I333   0.6835 0.0015 0.7912 0.8915 0.7540 0.6469 0.3034 0.8090 0.1325        0.4475 0.0000 0.0064
Comp.1 0.0176 0.0013 0.0004 0.0000 0.0000 0.0000 0.0000 0.0000 0.0820 0.4475        1.0000 1.0000
Comp.2 0.3300 0.0000 0.4710 0.1090 0.1702 0.4959 0.1080 0.4133 0.0000 0.0000 1.0000        1.0000
Comp.3 0.0000 0.9738 0.5187 0.5144 0.0853 0.6985 0.6792 0.6026 0.7626 0.0064 1.0000 1.0000
```

Figure 12: Output from R software

In the two figures above, the correlation matrix between the main components and each indicator chosen in the analysis is observed. Thus, Comp. 3 presents a strong correlation of 0.81 with the indicator I112 (weight of the population with higher education). However, in general, the informational redundancy has decreased, as a very weak correlation

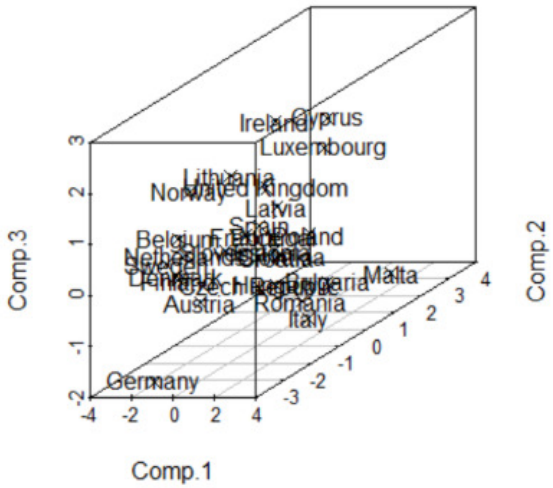

Figure 13: The three-dimensional graph

In the three-dimensional graph resulting in R Studio, using the Scatterplot 3d () command, one can see the 3 main components and the countries considered in the analysis. Thus, countries such as Finland, Belgium, Denmark and the Netherlands are grouped, more precisely; they are similar in information from the perspective of the 3 main components represented. Germany and Malta represent the outliers, and the majority of the countries are grouped above the average of the third main component, with positive values predominating, with Germany, Austria, Romania and Italy below average.

The third step of the study was to perform a cluster analysis. First, the distances between countries were calculated using the Euclidean distance as a calculation method.

```
> d<-dist(scoruri[,1:3],method="euclidean")
> d
          1         2         3         4         5         6         7         8         9        10        11        12        13
2  6.1507288
3  3.8928125 2.3543200
4  1.5054556 6.9624157 4.7770817
5  2.8439291 5.8436954 3.6418211 3.0793283
6  4.5103550 1.9337810 1.0551147 5.5628890 4.5379515
7  3.3678685 4.7550479 3.5066306 4.3525847 5.3969676 3.4301847
8  3.4825286 2.8872054 1.1905361 4.5892539 4.0653609 1.1391684 2.4540657
9  2.2583514 4.0231499 2.0284004 3.1310350 3.4154204 2.5903974 2.1433415 1.5697384
10 4.6850939 1.7537844 1.0996374 5.7140298 4.6152570 0.2126696 3.6205872 1.3405041 2.7558501
11 5.2424569 1.4890622 1.7074970 5.8233400 4.7671059 2.0403166 4.4508516 2.5826445 3.1888856 1.9260365
12 5.3028316 3.3424264 3.3780637 6.3441122 6.5407956 2.7149544 2.3537782 2.5353139 3.3920020 2.8180052 3.7680169
13 4.9454749 2.3792156 2.0362543 6.1488793 5.2725362 1.0218676 3.4088219 1.5960179 3.1588315 1.0942284 2.9337566 2.3175389
14 4.3766910 3.9742438 2.8596994 5.7922587 5.0429908 2.2401484 3.3136401 2.0733320 3.2846016 2.3901507 4.2684234 3.0747561 1.6353819
15 4.7791775 6.5777815 5.6123766 4.5751629 6.6896990 5.9794284 3.5942168 5.1460655 4.0396399 6.1057901 5.7519820 4.9923313 6.3737048
16 5.0113331 1.5325360 1.2134230 5.9755507 4.6036927 0.7851799 4.1817719 1.8450249 3.1361320 0.5917979 1.7232906 3.3396539 1.5408493
17 6.5844309 4.4133622 4.7769653 6.5695711 6.9970298 5.0989940 5.2073321 5.0333047 4.7334435 5.0794414 3.6592853 4.9073871 5.7366425
18 0.9431184 5.3732115 3.0650141 2.2268860 2.3919539 3.6829446 3.2167139 2.7290058 1.7606281 3.8510799 4.5274958 4.8212360 4.1491269
19 1.6093732 5.7693432 3.6172777 1.3088361 2.3492840 4.4967109 3.9537720 3.6382160 2.2305209 4.6246085 4.5683790 5.5974561 5.2066192
20 5.4658658 1.2002629 1.9524173 6.4588697 5.5734820 1.1095943 3.8092866 2.0069035 3.3779217 1.0082061 2.0150013 2.3748204 1.2513216
21 3.2042243 3.1341257 1.4850225 4.0064984 3.9530574 1.9359916 2.2741104 1.1470663 0.9501965 2.0743317 2.3748570 2.7993804 2.5909918
22 6.6762736 1.5240578 2.8735087 7.6421079 6.0225032 2.2911978 5.5939048 3.4301553 4.7976434 2.0916395 2.6116438 4.1637972 2.4471143
23 2.4665677 3.8097542 1.5703346 3.6170422 3.0248098 2.0615788 2.7196203 1.1632722 1.1444751 2.2340637 3.1733161 3.5607151 2.5851588
24 4.9549268 1.5916982 1.3544979 5.9905452 4.8600602 0.4504519 3.7924976 1.5814036 3.0197345 0.2771041 1.9793739 2.8161216 1.0291833
25 0.8251181 5.8159094 3.5074550 1.5380816 2.0687999 4.2632506 3.7483862 3.3504240 2.1600085 4.4143910 4.8039563 5.4412893 4.8347861
26 0.7113373 6.6006765 4.3195182 0.9596196 2.6496135 5.0331955 4.0086754 4.0534262 2.7635985 5.1948294 5.5848405 5.9451428 5.5427617
27 2.6249128 4.4313885 2.8399041 3.3803727 4.3764438 3.1626262 1.3403256 2.1247759 1.0719857 3.3355893 3.7296882 3.0241556 3.5363511
28 1.0005762 5.9794389 3.8397364 2.3881958 3.5785543 4.2378354 2.6617603 3.1451263 2.1571041 4.4345216 5.2779317 4.6696077 4.4828609
          14        15        16        17        18        19        20        21        22        23        24        25        26
```

Figure 14: The Euclidian methods

It can be seen from the graph of Euclidean distances above that the distance between Germany (6) and Belgium (2) is 4.51. The Euclidean distance was calculated by taking into account the main scores.

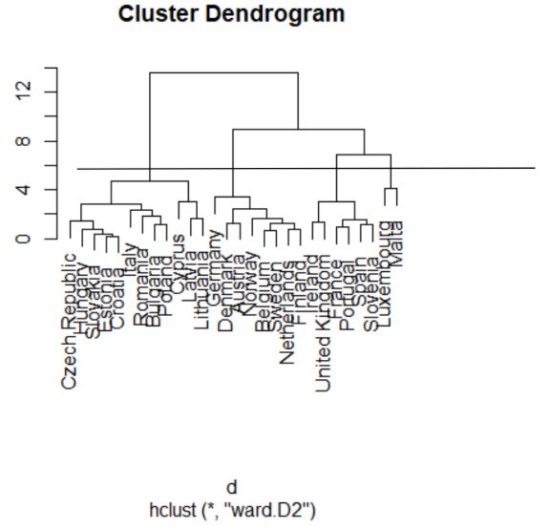

Figure 15: Ward method

In the figure above, the WARD distance was calculated, which is based on minimizing the increase in the sum of the squares of errors, after the groups are merged into one. By making a cut in the figure above, at the level of about 6, 3 clusters are obtained. The first cluster is made up of countries such as the Czech Republic, Hungary, Slovakia, Estonia, Croatia, Romania, Bulgaria, Poland, Cyprus and Lithuania.

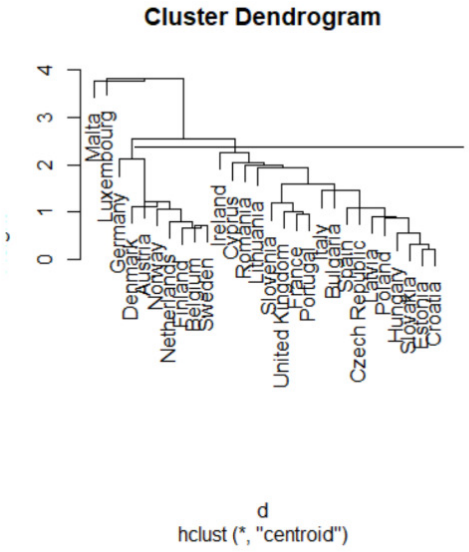

Figure 16: Centroid method

Using the centroid method, the Euclidean distances between the arithmetic averages of the components of the elements in the three groups were calculated, using the centroid of the clusters. After plotting a cut in the dendrogram, the distribution of the analyzed countries is observed in three clusters. Luxembourg and Malta are part of the first cluster, while the second cluster is made up of countries such as Germany, Denmark, Austria, Norway, the Netherlands, Finland, Belgium and Switzerland.

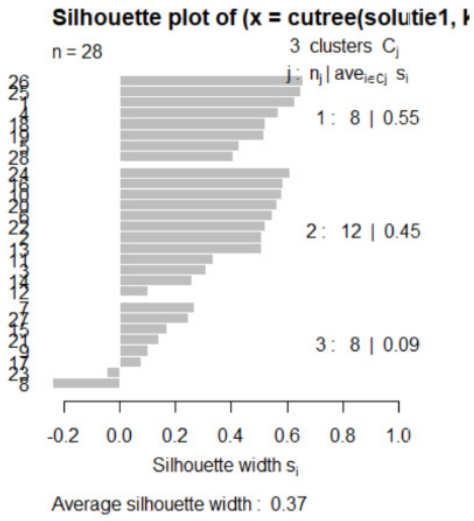

Figure 17: Silhouette plot for the Ward method

Analyzing the figure above, realized by the function silhouette () for the Ward method, it can be observed that the highest average is in the case of the division of the countries into 3 classes. Most countries which are correctly distributed (the value of the silhouette coefficient approaching more than 1) are also in the case of the division into 3 classes, but two countries will remain correctly distributed, with the coefficient of silhouette less than 0.

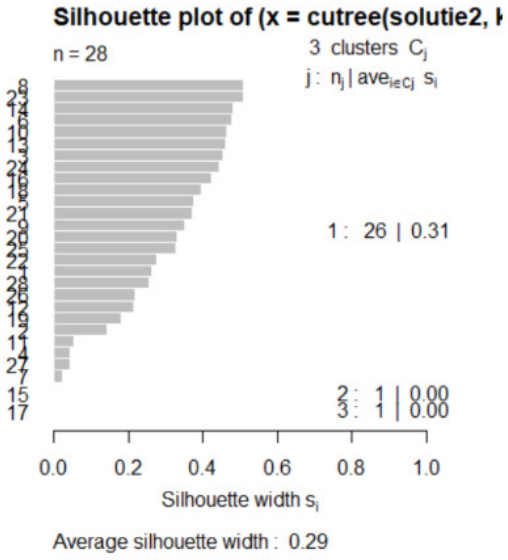

Figure 18: Silhouette plot for the Centroid method

Analyzing the figure above, realized by the function silhouette () for the Centroid method, it can be observed that with the increase of the number of clusters, the

coefficient of silhouette decreases, but it still remains positive, therefore, the countries are distributed correctly. Therefore, the division of countries into 3 clusters is recommended.

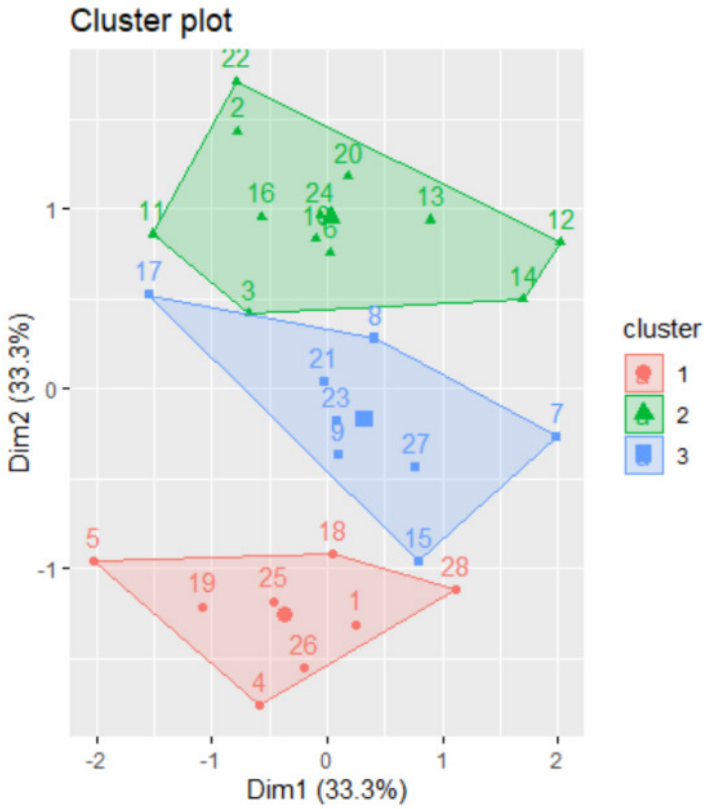

Figure 19: Cluster plot

From the figure above, realized in R Studio, the 3 clusters and the countries that form them are more clearly observed, represented by numbers. Also, the border countries, such as: 12 (Italy), 7 (Estonia) and 5 (Denmark), have small silhouette coefficients.

The use of classification trees in the analysis of the decision to become an entrepreneur

The purpose of the analysis is to see what factors influence a person's decision to become an entrepreneur. The data set used has been downloaded from www.gemconsoltum.com and is presented as follows:

	setid	country	ctryalp	cat_gcr1	cat_gcr2	yrsurv	id	weight	weight_l	weight_a	gemwork	gemwork3
1	1150900027	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900027	0.7576617	NA	0.7576617	Full: full or part time (includes self-employment)	Work:F-T, I
2	1150900066	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900066	0.6778038	0.6740203	0.6778038	Part time only	Work:F-T, I
3	1150900086	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900086	0.6778038	0.6740203	0.6778038	Part time only	Work:F-T, I
4	1150900090	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900090	0.7486202	NA	0.7486202	Retired, disabled	Retired st
5	1150900096	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900096	0.8189806	0.8144091	0.8189806	Full: full or part time (includes self-employment)	Work:F-T, I
6	1150900115	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900115	1.2968244	1.2895856	1.2968244	Full: full or part time (includes self-employment)	Work:F-T, I
7	1150900145	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900145	0.9689959	0.9635870	0.9689959	Retired, disabled	Retired st
8	1150900163	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900163	1.0234998	1.0177866	1.0234998	Full: full or part time (includes self-employment)	Work:F-T, I
9	1150900194	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900194	0.9464582	0.9411751	0.9464582	Retired, disabled	Retired st
10	1150900224	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900224	1.0330074	1.0272411	1.0330074	Full: full or part time (includes self-employment)	Work:F-T, I
11	1150900306	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900306	1.2004057	1.1937050	1.2004057	Retired, disabled	Retired st
12	1150900312	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900312	1.0234998	1.0177866	1.0234998	Full: full or part time (includes self-employment)	Work:F-T, I
13	1150900314	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900314	1.2845408	1.2773705	1.2845408	Full: full or part time (includes self-employment)	Work:F-T, I
14	1150900356	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900356	1.2845408	1.2773705	1.2845408	Full: full or part time (includes self-employment)	Work:F-T, I
15	1150900360	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900360	1.0502063	NA	1.0502063	Retired, disabled	Retired st
16	1150900362	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900362	0.6778038	0.6740203	0.6778038	Full: full or part time (includes self-employment)	Work:F-T, I
17	1150900367	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900367	1.2845408	1.2773705	1.2845408	Retired, disabled	Retired st
18	1150900395	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900395	1.2845408	1.2773705	1.2845408	Retired, disabled	Retired st
19	1150900413	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900413	1.2971960	NA	1.2971960	Retired, disabled	Retired st
20	1150900470	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900470	0.8025263	0.7980466	0.8025263	Full: full or part time (includes self-employment)	Work:F-T, I
21	1150900498	United States	United States	Stage 3: innovation driven	Stage 3: innovation driven	2015	900498	1.1307355	1.1244237	1.1307355	Full: full or part time (includes self-employment)	Work:F-T, I

Figure 20: The table of the data set

From the total of 230 countries, Germany was chosen for the analysis because it has a higher number of records (3842), which will help in obtaining greater accuracy of the results. For the analysis the following variables will be considered:

• bstart - represents the variable by which an individual wants or does not want to open a business, depending on the following features;

• suskill - represents a variable by which the individual thinks he or she has competencies as entrepreneurs to open the business;

• fearfail - represents the variable that registers the fear of failure, which prevents individuals from opening a business;

• gender - represents the gender variable;

• gemwork3 - represents the occupational status, and already has three levels;

• gemhhinc - represents the income category;

• knowent - binary variable - which indicates that the individual has acquaintances or friends who have opened a business;

• gemeduc - a variable that represents the level of education of the individual;

From the total of 3842, following the data cleaning process, the following division was obtained:

```
> germany<-na.omit(germany)
> table(germany$bstart)

  No  Yes
2647  192
```

Figure 21: Output of R

Thus, at present, the authors have a total of 2839 registrations, of which, 2647 consider that they do not want to open a business, while 192 individuals want to open a business.

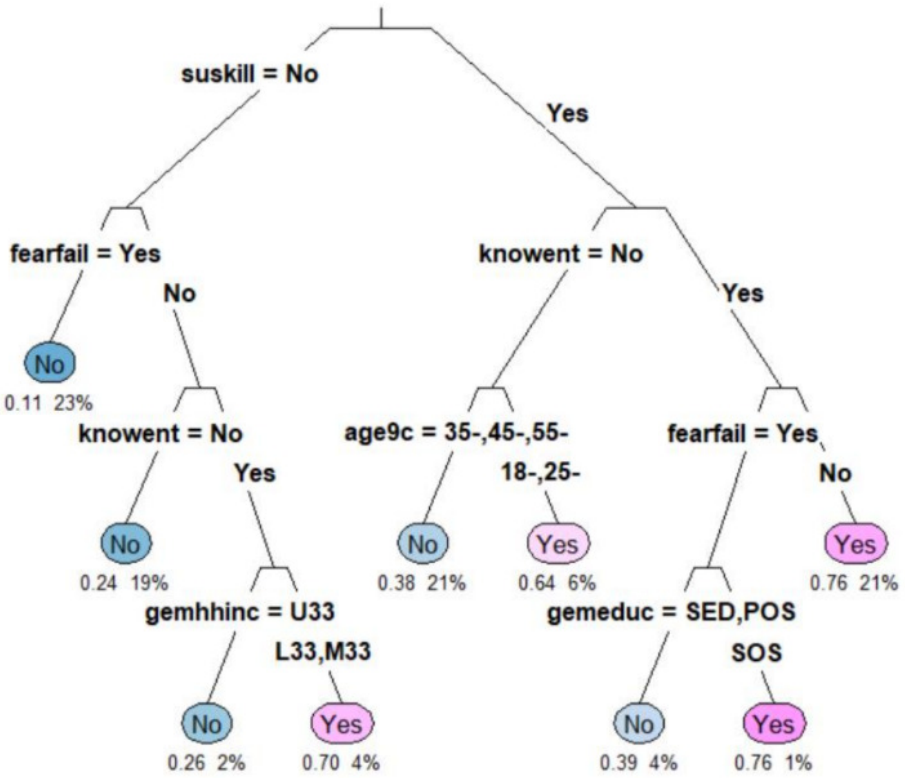

Figure 22: The classification tree

On each branch, the variable used and the level it takes can be seen ("Yes"/"No"). Only the leaf nodes are labeled, and for "type" = 3 on each branch of the tree, the variable used and the level it takes are seen. For the numerical information, it is observed that "extra" = 106 was used as well as the probability of class 2; the class "yes", that is the weight of the observations from the total sample from the initial node to the root node. It started from 0.11 as a probability of the "no" class, meaning that a proportion of 23% of the respondents said that they were not afraid of failure. The first variable used was "suskill" whereby the individual believes in having entrepreneurial skills to open the business. It is noticeable that on the left branch, there is "No"; only those who have the answer “no” have been selected, and the no tag is associated. The tree in the previous figure shows a pale pink leaf which shows that there is a fairly low probability of 0.64; 6% as a risk of misclassification. For respondents who do not know close persons who have a business, it is found that those between 35 and 45 years old do not consider that they can open a business, whereas a percentage of 6% people consider that they can be entrepreneurs in the near future. The "plum" function was applied on the initial tree, using the minimum accepted value of the complexity parameter and obtained a new tree with the variable "suskill" in the first place of importance. Even if the "root" variable is not found on the tree, it appears in the tree with a probability of 60.4%.

The tree below was made using the training set. It is noticeable that there is a probability of 0.98 for individuals who consider that they do not have entrepreneurial skills to open a business.

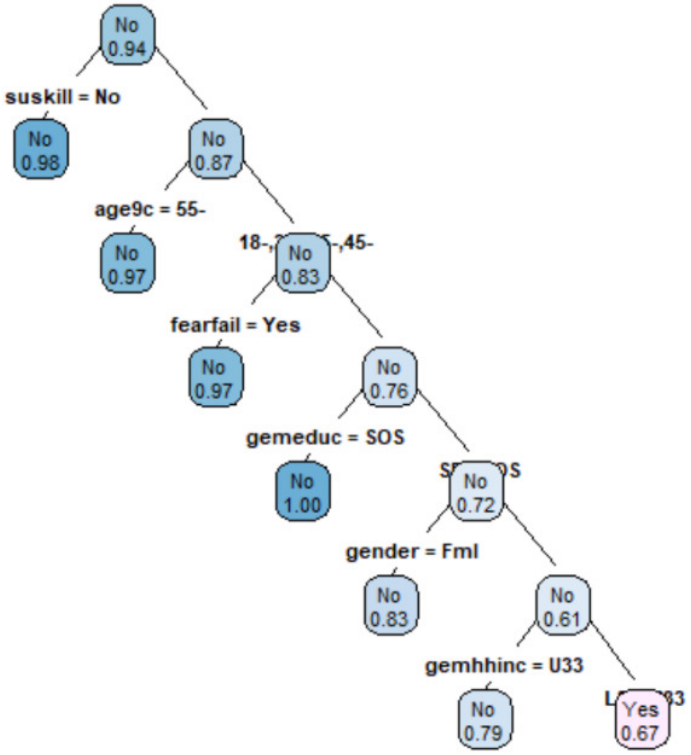

Figure 23: Training set method of the tree

Conclusions

This article demonstrates the usefulness of the analysis and analysis tools applied to the large datasets generated in the process of observing innovation in the private sector in Europe.

In the first part of the analysis, a strong correlation was observed between public spending on research and private spending on research and development. A strong positive correlation was also observed between the share of companies offering training and the share of companies that innovated.

From this first analysis, it was observed that the more training offered by companies, the more performance of a company increases. An important correlation was also observed between the share of companies that have access to high-speed Internet and public spending on research. For a clearer analysis, the principal components analysis technique and the clustered technique were used to correctly group the countries considered in the analysis from the variable point of view taken into account.

The last part of the analysis was the evaluation of the factors that influence a person's decision to become an entrepreneur. The tree classification technique was used. Germany, having the most records in the dataset, was the country analyzed for the proposed objective.

Future analyses of this article are aimed at implementing data mining techniques by applying an online questionnaire to several countries in Europe to determine information about the behavior of

individuals who want to be entrepreneurs and how to change this behavior at certain circumstances. For this purpose, a correspondent analysis and a conjoint analysis will be used.

References

- Behera, H.S., Nayak, J., Naik, B., Abraham, A. (2019). *Computational Intelligence in Data Mining.* Proceedings of the International Conference on CIDM 2017, *Springer;*
- Chiriță, N., Nica, I. (2020). An approach of the use of cryptocurrencies in Romania using data mining technique. *Theoretical and Applied Economics, Volumes XXVIII, no.1(622)*
- Davenport T. H., Dyche J. (2013) Big Data in Big Companies, *International Institute for Analytics;*
- Kerstholt, F., Van Wezel, J. (1976). Optimizing Social Participation over the life Cycle: Towards an integrated socio-economic theory and policy;
- Kokate, U.; Deshpande, A; Mahalle, P.; Patil, P. (2018) Data Stream Clustering Techniques, Application, and Models: Comparative Analysis and Discussion. Big Data Cogn, Comput;
- Lyman, P.; Hal R Varian;Swearingen, K.;Charles, P. (2003) *How much information?*School of Information Management and Systems, University of California at Berkeley;
- Nica, I., Chiriță, N., Ciobanu, F.A. (2018). Analysis of Financial Contagion in Banking Networks. *32nd IBIMA Conference, Vision 2020: SUSTAINABLE ECONOMIC DEVELOPMENT AND APPLICATION OF INNOVATION MANAGEMENT, pages 8391-8409;*
- Nica, I., Chiriță, N. (2020). Conceptual dimensions regarding the financial contagion and the correlation with the stock market in Romania. *Theoretical and Applied Economics, Volumes XXVIII, no.1(622);*
- Perner, P. (2017). *Advances in Data Mining. Applications and Theoretical Aspects*. 17th Industrial Conference, ICDM 2017, New York, USA, 2017 Proceedings, *Springer*;
- Shaw, J., (2014) Why Big Data is a Big Deal,*Harvard Magazine;*
- www.gemconsoltum.com
- www.bigdata.ro/big-data/
- www.forrester.com/Big-Data