Mining data of anti-corruption institutions to identify unusual growing trends of assets of individuals

Besnik Dragusha*, Kadri Sylejmani**, Lule Ahmedi**, and Blerim Rexha**
*Anti-Corruption Agency/Office for Support, Cooperation and Information, Prishtina, Kosovo
**University of Prishtina/Department of Computer Engineering, Prishtina, Kosovo
besnikdragusha@hotmail.com, (kadri.sylejmani, lule.ahmedi, blerim.rexha)@uni-pr.edu

Abstract - **Identifying unusual grows of assets of a senior public officials can be done by analyzing their asset data such capital investments, cash, salaries, etc., which are usually declared at corresponding national institutions for anti-corruption. Usually, such agencies collect data for thousands of officials every year. Hence, manual analysis of such a relatively large of data, for a short computation time (i.e. in the range of seconds), is infeasible. Therefore, in this paper, we suggest a solution to this problem, which bases on two algorithms that arise from the field of artificial intelligence. First is the k-means algorithm, which, based on asset declaration data, groups the officials into clusters, where the two main clusters are: officials that have a normal asset grow (tagged as Best Clusters), and officials with unusual asset grow (aka. Bad Clusters). Second, by using the asset declaration data, as well as the demographic data of the existing officials, we train a Decision Tree algorithm to predict the group of the new officials, in order to place them into one of the existing groups/clusters. The experiments performed over a data set of 22300 assets declarations show that the proposed approach achieves an accuracy of 90% in comparison to the human expert analysis.**

Keywords - ***Assets Declaration; Clustering; k-means; Classifications; Decision Tree.***

I. INTRODUCTION

Assets Declaration (AD) process is a way of presenting and describing the richness of the senior public officials, where the declared data will be analyzed by anti-corruption institutions of corresponding countries [1]. When comparing the data of asset declarations in different countries, it can be noticed that the structure of these assets declarations includes components such as the domain of immovable property, movable property, cash, liabilities (credits), annual income, family annual income, etc. [2]. In each of these types of properties, it may have some declared property per person within the type (e.g., immovable property, may contain: house, land, flat, etc.). Then, each asset is described with some details (e.g.: name of the property, asset value in euro or the other currencies, year of profit, and its ownership) that outlines a given property. In general, each declaring person/declarant (i.e. senior public official) may have several statements of declarations that correspond individual years he/she has been holding a public position. Hence, the problem of analysis of assets declaration (AD) is to find and identify the officials who have unusual (abnormal) growth of their assets over the years, as well as classifying new declarants into potential declarants with unusual assets growing trends, or usual growing trends. Hereupon, this analysis aims to detect declarants who will be included in the anti-corruption investigation list to verify their assets and the way they are obtained (i.e. its origin).

The main contributions of this paper can be outlined as in the following: (1) presentation of a mathematical model for the problem of assets declaration at anti-corruption institutions, (2) application of the k-means algorithm for identifying senior public officials with unusual growing of assets, (3) application of the decision tree algorithm for prediction of assets growing trends for the new officials, and (4) an experimental study, using human experts, for evaluation of the proposed approach.

This paper is organized as in the following: Section II presents a literature review of approaches solving problems similar to AD problem; Section III presents the modelling of the problem and its mathematical definition; Section IV presents the proposed approach with representation of the solution; Section V presents the experiments and result interpretation; and Section VI concludes the paper.

II. LITERATURE REVIEW

In the literature, there is a large number of solutions to the problem of detection and finding of money laundering (ML) known as anti-money laundering solutions (AML). This type of problem is not the same as our problem of analysis of asset declaration of senior public officials, but it is somehow related. The AML problem has to do with controlling bank transfers to detect 'dirty' money, for the transfer of large amounts of money, whereas the problem tackled in this paper is about identifying senior public officials, whose assets (declared at anti-corruption institutions that include bank account information) have grown dramatically over the course of past coups of years. Klac et al. [3] solved the AML problem by using data mining (DM) methods to analyze huge datasets and build customer profiles of different groups, according to techniques: clustering (to group customers according to their properties), classification (to classify clients by predicting their behavior in the future). Gao and Ye [4] present a framework for DM based on the AML methodology, by using the concept of algorithmic classification, clustering, and searching, so the data is categorized in unusual/abnormal/anomalous, and detected

which persons are categorized as outliers. Further, the approach of Gao and Ye [4] detects and analysis the relation and interaction between members of the group, between groups and subgroups. Tang and Yin [5] present a solution for AML problem based on Statistical Learning Theory (SLT). They use methods based on Support Sector Machine (SVM) to filter transactions on database systems, by detecting the set of unusual transactions, and then classifying the data into two classes, where one of them would contain outlier data.

To the best of the author's knowledge, there is no any other approach in the literature that automatic analysis the data declared and accumulated at anti-corruption institutions, hence, in this aspect, our proposed approach is novel to this field.

III. PROBLEM DEFINITION

The problem of AD is defined in such a way that the declarations of assets, accumulated over the years, are compared by themselves, finding their growth rate, and based on this measure, it is decided which declarants have had an unusual increase in the property. Then, the new declarants are classified into the group of usual or unusual ones, according to their match to the existing declarants, by considering their previous declarations. And then, by processing the identified data in the same training set, we make the classification of new declarants into declarants with usual or unusual growth. For more information about this problem definition, see [2].

The mathematical definition of the AD problem can be done by using following notations:

D -*Assets declaration;*

s -*Number of senior;*

m -*Number of declarations;*

n -*Types of declaration attributes;*

A_i -*Attribute, where: i = 1, 2, ..., n;*

Z_i -*Declarants (Senior Official), where: i = 1, 2, ..., s;*

Declarations_1 – existing declarants (that have two or more assets declarations);

Declarations_2 – new declarants (that have only one assets declaration);

$$diff_{AVR} = (\sum_{i=0}^{m-1}\sum_{j=0}^{n}(d_{(i+1)j} - d_{ij}))/(m-1) \quad (1)$$

where: m - no. of declarations, n – no. of attributes

$$Z_{ij} = diff_{AVR}(D_{ij}),\ where\ 1 \le i \le m, dhe\ 1 \le j \le n \quad (2)$$

Equation (1) is used to calculate the average difference of the values of the attributes (i.e. d_{ij}) for multiple declarations of a given declarant. Then, based on these values, all declarants are represented in a matrix (Z_{ij}) of m rows and n columns, as denoted by Equation (2). Consequently, as property attributes, we use six categories, and we classify them into three groups: assets (Immovable Property - IP, Movable Property - MP, and Cash - C), Liabilities (debts) - L, and revenues (Annual Incomes - AI and the annual Family Income - FI). After finishing the process of grouping the declarants into clusters, we calculate the Total Property (P_T), Total Liabilities (L_T), and Total Income (I_T) for all members of each cluster (as specified by equations (3), (4), and (5)). Next, with the aim of identifying the best and bad clusters, for each group G, we calculate the difference between P_T, I_T, and L_T, as specified by Equation (6).

$$P_T = \frac{1}{m}\sum_{i=0}^{m}\sum_{j=0}^{n}(IP_{ij} + MP_{ij} + C_{ij}) \quad (3)$$

$$L_T = \frac{1}{m}\sum_{j=0}^{n} L_j \quad (4)$$

$$I_T = \frac{1}{m}\sum_{i=0}^{m}\sum_{j=0}^{n}(AI_{ij} + FI_{ij}) \quad (5)$$

$$diff_G = \sum_{i=0}^{k}(P_T - (I_T - |L_T|)) \quad (6)$$

where: m – No. of declarants, n - No. of attributes, k – No. of members in group G.

In order to calculate the average differences of growing asset values, the declarant must have at least two assets declarations. And declarants that meet this condition shall pass to the grouping process. On the other hand, the new declarants (those who have only one declaration, and that in the current year) will only be subject to the classification process. In order to realize the grouping, the average differences in asset growth should be first calculated. When the process of the grouping is finished (i.e. The group with the usual growing and unusual assets growing are determined), then it can be continued with the process of classifying the new declarants. These can also be considered as a constraint of the property declaration problem.

IV. SOLUTION APPROACH

The solution to the problem of analysis of asset declarations is made by undergoing several steps (see Figure 1), which are: (1) data pre-processing, where initially the declarants are separated into *existing* and *new* ones, and the calculation of changes of asset values over the years is made, (2) the k-means algorithm is applied for grouping declarants into specific clusters, (3) identification of the clusters with usual and unusual asset growing trends, where the data of this step are used as *Training Set* for the classification step, and (4) classification of new declarants (*aka Testing Set*) by predicting their class (usual or unusual) of the asset growing through a Decision Tree algorithm.

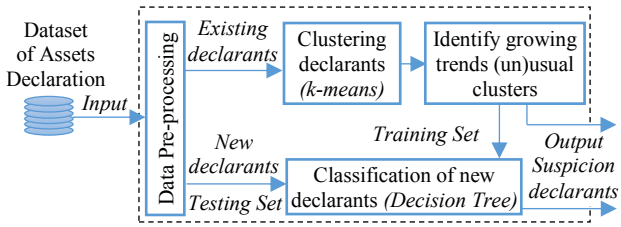


Figure 1 Block-diagram of solutions approach

A. Data pre-processing

Processing of data of assets declarations is obtained by calculating the total value of the property of an attribute, for example, declarant X, declared the declaration d_1, with the attribute A_1 = immovable property, such as: House 200,000.00 €, flat 80,000.00 €, land 30,000.00 €. Then, this attribute is represented as A_1= 310,000.00 € (200,000.00 + 80,000.00 + 30,000.00 = 310,000.00). In this way, we summarize the data of every attribute, so that it is included

in representing the total amount of assets. Thereafter, we calculate the average differences in the value of the growing asset of attributes for several years, according to the equation $diff_{AVR}$ (1), whereby all the declarants (which are part of the group tagged as Existing Declarants) are represented through the matrix Z_{mxn} (2), which contains the attribute values for all senior public officials. In the Z_{mxn} matrix, each row *i (1≤i≤m)* - represents an attribute of the declaration of a declarant, whereas each column *j* (1≤j≤n) - represents the values of asset declaration of a given attribute. The calculation of these average differences, and the representation of declarants through the matrix Z is described by the following pseudo-code (Algorithm 1).

Algorithm 1 The procedure for calculation of average difference

```
Get data of declarations (Import data from dataset);
Data = {data_ij | i← rows, j← columns} ← declarations;
Declarations_1 ← existing declarants;
Declarations_2 ← new declarants;
Z = {z_ij | i← rows, j← columns} (matrix with diff_AVR of
                                  all data Declarations_1);
rows ← 0, columns ← 0;
foreach declarations_1_ij ϵ Declarations_1 do

{
      m← max rows length;
      n← max columns length;
      while (rows <= m)
      {
         Difference = Get data declarations_1;
         while (columns <= n)
          {
           difference ← diff_AVR;
          }
         Z = difference;
       }
}
return Z;
```

B. Clustering declarants

The data of asset declaration represented through the **Z** matrix should be grouped into certain clusters according to the similarities of their attribute values, so that the declarants within the same cluster have similar values among themselves, and also have major differences with the other clusters of declarants. For this purpose, we use the *k-means* algorithm, which divides the data of dataset m into ***k*** clusters, whereby each declarant belongs to the nearby cluster, or to the nearest centroid ***c*** of cluster [6]. Simultaneously, the k-means algorithm moves the declarants from one group to another group that depends on the distance of members from the centroids of the respective groups, while the convergence of groups is achieved [7]. The determination of each declarant in which group he/she should be, is done by measuring the distance between the centroids and the declarants of each group, where the declarant will become part of the group whose centroid has the shortest distance with that declarant. The difference from a declarant and a centroid c_i is calculated by using the Euclidean distance as represented in [8]. After calculating the distance for each of the declarants of a cluster, as well as of the declarants that change the cluster, we calculate the value of new center *c'*, according to equation *C(xi)* [9], which represents a new average value over the given cluster. The following pseudo-code represents the actual method for clustering declarants into groups according to the similarities in the declared data.

Algorithm 2 Clustering method – based in *k-means*

```
Input:  Z = {z_ij | i ← rows, j ← columns} (set with data from
                                            Algorithm 1)
Get MaxIterations ← (maximal number of iterations)
Get k (k ← no. of clusters – input value, and k>=2)

Output: C = {c_ij | i = 1, 2, …, k; j = 1, 2, …, k} (set with data for
                                            saving value of centroids)
        L = {l(c) | c = 1, 2, ..., n} (set for labeling clusters)
        M = {m_ij | i ← rows, j ← columns} (gets data from Z and L )
        S_C = Create initial solution;
        S_B = S_C (value of centroids);
        Iterations = 0;
        While (Iteration <= MaxIterations)
        {
            Sc = Get K-means Clustering
            d_ij ← (Calculate a Euclidian distance);
            d^0 = d_ij;
            foreach z_i ϵ Z do
                        l(z_i) ← argminDistance (c_ij, d_ij), j ϵ
                                {1,2, …, k}(Eucledian distance)
            end;
            change false;
            g^0 ← (verifing true clusterings for every subject);
            S_C = d^0 (Current solution S_C = solution of d^0);
            Update clustering(S_C);
            Iterations ++;
            S_B=S_C (best solution S_B = current solution S_C);
            If (no declarants changet cluster)
                        {
                        g^0 = g^1;
                        S_C= g^1;
                        S_B=S_C;
                        }
                        return S_B;
            }
            return (C, M);
```

The presented approach can be used to group declarants into *k* clusters. After this, it is necessary to find the group (or groups) that have the highest discrepancy between incomes and capital investment's value over the years, which is a process described in the next subsection.

C. Identifying unusual asset growing trends

In this phase, we identify the group with the most usual trends of assets growing - *Best cluster*, and the group with the most unusual trend of assets growing - *Bad cluster*. The *"Bad"* cluster (group) contains the members that have the highest discrepancy of incomes and benefits, in terms of assets and capital investments growing, while the *"Best"* cluster is the one with the lowest such a discrepancy. This procedure is presented below by Algorithm 3.

The Bad Cluster list includes members with declarants that have unusual (abnormal) growing of assets. Hence, the members in this list should be part of the process for verification and investigation of their property by anti-corruption institutions. In cases, when we create more than two clusters, then, as part of the *Best* and *Bad Clusters*, we can have a range of groups, which could be selected between a predefined internal. From the practical point of view, this makes sense because there are occasions that a given group might contain only a single member (i.e. "Outlier"), who has a great difference from the other declarants (be it for bad or good context), but this does not mean that he/she is the only one who should be investigated, because it might happen that the next closer mean that he/she is the only one who should be

investigated, because it might happen that the next closer group might need to be part of investigation too. This process is explained in more details in [2] (p.48 and p.49).

Algorithm 3 The method for identifying unusual growing trends

```
Get Z, K;
Z = {z_ij | i ← rows, j ← columns} (set of data entries with
                                    clustering ← Algorithm 2);
K = {k |label value for each cluster} (Import from Algorithm 2);
m ← max Length rows;
n ← max length columns;
foreach k_j ϵ K do
{
    while (rows<= m)
        {
        while (columns<= n)
            {
                P_T = SUM(IP_ij, MP_ij, C_ij)/rows (length
                                                    of k);
                L_T = SUM(L_j)/rows(length of k);
                I_T = SUM(AI_ij, FI_ij)/rows(length of k);
            }
                Diff(G) = SUM (P_T – (I_T – abs(L_T)));
        }
        MIN(Diff(G)) ← Best Cluster
        MAX(Diff(G)) ← Bad Cluster
}
return (k, Best, Bad);
```

D. Classification of new declarants

For the case of the "new declarants", who declare assets for the first time, both, the calculation method - $diff_{AVR}$ and clustering method cannot be applied, because there are no existing data to compare with, so that their growing trend of assets over the years could be calculated. Hence, here, we apply the Decision Tree algorithm, which creates a tree model according to the data of the Training Set, which, in our case, is the data set created when the *k-means* algorithm is applied. The data of the Training Set enlists the existing declarants, who belong to either Bad or Best cluster, where each of them is annotated with a True or False mark. In order to predict the group of the new declarants, besides the three properties used for the existing declarants (immovable property-IP, movable property-MP, and cash-C), we consider properties such as birthplace, year of birth, name of the institution and job position, while liabilities are not taken into account [2]. In Algorithm 4, we present the pseudo-code of the actual decision tree algorithm that is implemented in this project, where details for calculation of metrics such as Entropy *H (S)* [10] and Information Gain *Gain(S, F)* [11] are outlined.

Algorithm 4 Classification method based on Decision Tree

```
Input:  Declarations_2 = {d_ij | i ← rows, j ← columns}
            (data to be "Testing Set" from Algorithm 1);
        Matrix_F = {f_ij | i ← rows, j ← columns}
            (data to be "Training Set" from Algorithm 1);
Ouput:  Tree;
        T;
Matrix_F = { f_ij | i ← rows, j ← columns} (a round
  attribute values "year of birth", "Property", "Incomes");
A ={a_i | i = (0, m), } (m = max number of attributes);
A ← Training Set;
T ← Testing Set;
Create a Tree Root (Select A_i, which is a best attribute)
Root ← A_k (Best attribute ← higher Gain(S, A), k ϵ i);
A' = {a'_i | (i= 0, m); m = number of current attributes}
        foreach a' ϵ A' do
            While (i ≤ m)
            {
                Calculate Entropy E(A') of attribute;
                Calculate Gain(S, A') – (S – simples
                            collection, A' - Attribute);
                }
                N' ← (Best attribute ← attribute with
                                high Gain(S, A'));
                N' ← be a Node of tree
                return (N');
        }
        return (Tree);
        T ← (Tree - this model applies for data of
                                        Testing Set);
        return T;
end;
```

This method measures the similarity (using the above mentioned attributes) between the new declarants and the existing ones, and, based on that, marks the new declarants either as usual or unusual declarants. Based on this, we predict whether there are new declarants that have prerequisite to become abusive, in terms of unusual growth of their assets.

V. EXPERIMENTAL RESULTS

In this section, we show the experimental results of the analysis of the data of assets declarations by using a real data set. In the following, we present the dataset structure that includes details about assets declarations, as well as the evaluation results about the proposed solution, which include comparative results, in terms of quality of analysis, computation time and user evaluation.

The algorithms are developed by using C# programming language that is part of Microsoft Visual Studio 2013 development environment. The experiments are done by means of a machine with CPU of type Intel Core i7-3770 3.4GHz with 8GB of RAM memory, while running on a Windows 10, 64-bit Operating System.

A. Dataset

The utilized dataset contains data about declarants, that are described by means of 15 attributes, such as: ordinal number (No), name and surname of the declarant, ID (this is an anonymized ID number that is assigned for the purpose of this research work), birthplace, year of birth, name of institution, name of the sub-institution, job position, date of declaration, immovable property, movable property, cash, liabilities, incomes, and family income. This dataset is composed of real declarants, whose data are available for public view, and it is obtained from the website of Kosovo Anti-Corruption Agency (ACA) [12]. This dataset contains about 22300 asset declarations, and it includes declarations for the period from 2011 to 2017.

B. Quality of prediction

In order to analyze the quality of the obtained solutions that are returned by the proposed approach, we have selected the data about 10 different declarants, who have a total of 20 property declarations. The experts from the Anti-Corruption Agency in Kosovo (ACAK) have been asked to manually analyze their declared data about the last two years. Their findings are that five out of ten of the declarants could be marked as suspicious. In Table 1, we present the complete details of this 10 declarants, where, for privacy reasons, their names and IDs are anonymized.

According to the findings of experts from ACAK, the declarants from 1 to 5 can be considered as public officials that belong the group with the most unusual growing of assets, hence they should be subject to the process of verification and investigation about the origin of their properties. On the other hand, when our approach is applied to analyze the data of these 10 declarants, the results are identical to the manual results produced by ACAK experts, when the *k-means* algorithm is applied to group declarants into five groups.

TABLE I THE DATA OF TWO LAST DECLARATIONS OF TEN DECLARANTS

No. of declarant	*No. of declaration*	*Name& Sur-name*	*ID*	*Date of declaration*	*IP (€)*	*MP (€)*	*C (€)*	*L (€)*	*AI (€)*	*FI (€)*
1	1	A	12345	3/30/2016	3,280,000.00	9,000.00	891.89	139,520.00	10,660.00	160,950.00
	2	A	12345	3/27/2017	3,908,000.00	9,000.00	2,400.00	626,962.00	11,152.00	160,800.00
2	1	B	23456	5/12/2016	120,000.00	0.00	0.00	0.00	3,500.00	5,860.00
	2	B	23456	3/30/2017	277,000.00	4,500.00	2,628.00	6,000.00	1,666.00	1,700.00
3	1	C	56789	3/22/2016	232,000.00	42,500.00	187,050.00	37,069.50	64,205.33	46,600.00
	2	C	56789	3/14/2017	2,482,000.00	37,500.00	234,000.00	24,607.50	61,334.00	46,600.00
4	1	D	67890	3/14/2016	685,000.00	5,000.00	800.00	0.00	9,678.00	10,800.00
	2	D	67890	3/29/2017	985,000.00	5,000.00	800.00	0.00	9,678.00	28,036.68
5	1	E	98765	3/17/2016	860,000.00	13,100.00	2,345.00	15,000.00	8,363.00	8,400.00
	2	E	98765	3/30/2017	902,610.00	3,700.00	0.00	42,700.00	5,241.00	8,500.00
6	1	F	09876	3/22/2016	350,000.00	4,000.00	40,000.00	53,000.00	20,908.00	7,200.00
	2	F	09876	3/29/2017	370,000.00	4,000.00	40,000.00	26,000.00	35,322.97	8,400.00
7	1	G	54321	3/31/2016	563,300.00	16,300.00	1,430.00	0.00	22,538.94	4,764.00
	2	G	54321	3/31/2017	484,500.00	7,000.00	9,932.14	0.00	21,813.28	5,916.00
8	1	H	65432	3/18/2016	122,000.00	4,000.00	42,000.00	0.00	3,076.23	620.34
	2	H	65432	3/29/2017	122,000.00	11,500.00	42,060.00	0.00	26,751.12	7,444.08
9	1	I	88880	3/31/2016	140,000.00	23,000.00	0.00	0.00	23,112.00	4,740.00
	2	I	88880	3/28/2017	140,000.00	23,000.00	0.00	0.00	29,426.64	4,939.80
10	1	J	99881	3/15/2016	204,000.00	4,000.00	0.00	0.00	682.00	75.00
	2	J	99881	3/30/2017	204,000.00	0.00	0.00	0.00	682.00	0.00

Nevertheless, the results of our approach are slightly different from the expert analysis, when the k-means algorithm groups the declarants into three or ten groups, where (in Table 2) it can be observed that person E is marked as a "suspected" declarant by ACAK experts, whereas our approach, in both cases, marks her/him as "unsuspected". In overall, based on these experiments with real users, it can be concluded that our approach can achieve at least 90% of accuracy in predicting whether a give declarant should be subject of further investigation of the origin of their properties. In addition, it should be noted down that the manual analysis done by ACAK experts includes the data of only last two years of declarations, whereas, our approach considers the data of all declarations of the officials that have been done in the past.

TABLE II. EVALUATION RESULTS FOR THE K-MEANS APPROACH

Name & Surname of Declarant	**Suspected Declarants**			
	ACAK Results	*Algorithm results*		
		K=3	*K=5*	*K=10*
A	Y	Y	Y	Y
B	Y	Y	Y	Y
C	Y	Y	Y	Y
D	Y	Y	Y	Y
E	Y	N	Y	N
F	N	N	N	N
G	N	N	N	N
H	N	N	N	N
I	N	N	N	N
J	N	N	N	N

TABLE III THE DATA OF DECLARANTS USED FOR CLASSIFICATION

	Declarants	*Birthplace*	*Year of birth*	*Institution*	*Job title*	*Total Property (€)*	*Total Income (€)*	*Output (Suspected)*
Training Set	A	Prishtina	1960	Kosovo Police	Director	200,000.00	20,000.00	FALSE
	B	Prishtina	1960	Kosovo Police	Manager	200,000.00	10,000.00	FALSE
	C	Ferizaj	1970	Kosovo Police	Secretary	200,000.00	10,000.00	TRUE
	D	Gjilan	1980	Government	Secretary	200,000.00	10,000.00	TRUE
	E	Gjilan	1980	Presidency	Secretary	100,000.00	20,000.00	TRUE
	F	Gjilan	1980	Presidency	Manager	100,000.00	20,000.00	FALSE
	G	Ferizaj	1970	Presidency	Director	100,000.00	20,000.00	TRUE
	H	Prishtina	1960	Government	Manager	200,000.00	30,000.00	FALSE
	I	Prishtina	1960	Presidency	Director	100,000.00	20,000.00	TRUE
	J	Gjilan	1980	Government	Secretary	100,000.00	32,000.00	TRUE
	K	Prishtina	1960	Government	Director	100,000.00	18,000.00	TRUE
	L	Ferizaj	1970	Government	Secretary	200,000.00	20,000.00	TRUE
	M	Ferizaj	1970	Kosovo Police	Manager	100,000.00	20,000.00	TRUE
	N	Gjilan	1980	Government	Director	200,000.00	20,000.00	FALSE
Testing Set	**X**	**Prishtina**	**1960**	**Government**	**Director**	**200,000.00**	**10,000.00**	**?**

In this regard, our approach in comparison to the manual one, besides being very quick in producing the results of the analysis, it can be more comprehensive, in terms of producing results that cover a wider period of declarations.

In Table 3, with the aim of didactic explanation, we present the details of a sample group of declarants (along with the corresponding attributes) that could be used as part of the training set. The aim is that based on attributes such as birthplace, year of birth, institution, job title, total property and total income, to predict whether a given new declarant can be associated with an existing declarant (i.e. one of the declarants from A to N in the training set) so that her/his class (suspected or not suspected) could be predicted as the class of the new declarant (i.e. declarant X). Based on the decision tree algorithm that was described earlier, the declarant X is classified to the class "FALSE", which means that she/he is predicated to be an unsuspected declarant, as presented in Figure 2.

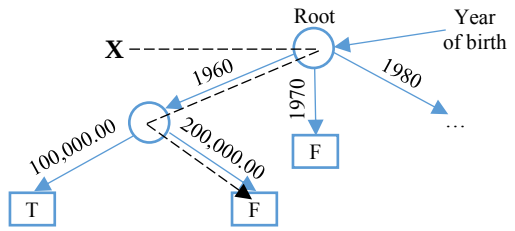
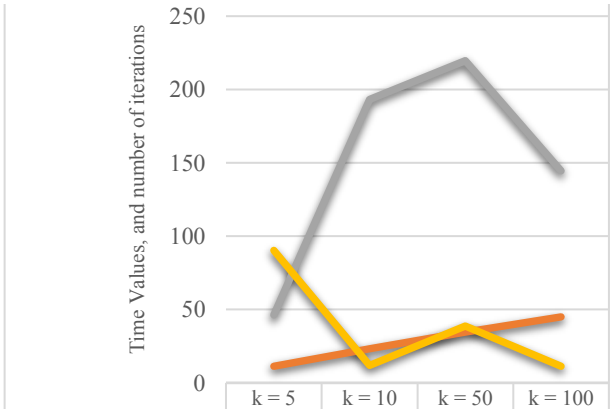


Figure 2 Decision Tree for classification of declarant X

C. Computation Time

By using a dataset of declarations with over 22,000 records, we extensively experimented with our proposed approach by analyzing the computation time for different algorithm settings, in regard to the number of clusters and the length of the identification interval. The computation results are presented in Figure 3, by outlining the number of iterations and clustering/classification time.



	k = 5	k = 10	k = 50	k = 100
Number of Iterations	11.34	23.34	34.67	45
Clustering Time /ms	46.33	193.33	219.67	144.67
Classification Time /ms	90.33	12	38.67	11.33

Figure 3 Computation time for different approach settings

The time for grouping declarants into clusters depends on the number of clusters utilized. The longest computation time observed from the experimentation is when the number of clusters is k = 50, which is 219.67ms, whereas the best case scenario is when the number of clusters is only 5, where the computation time is only 46.33ms.

D. User Evaluation

In order to measure the usability of the proposed approach, a prototype implementation is made and offered by the experts of ACAK to evaluate its features. Consequently, an only survey is made with them, where they were asking a number of questions regarding specific functionalities of the prototype system. In overall, the impression of the ACAK experts is that the prototype would facilitate to a great extent their daily duties, both, in terms of effectively and efficiently, towards a fast identification of the declarants that could be considered as "suspects". In Figure 4, we present the results of the five of the main questions asked in the survey, whereas the complete results of the survey can be found in [2].

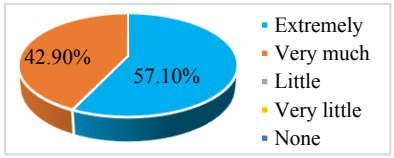
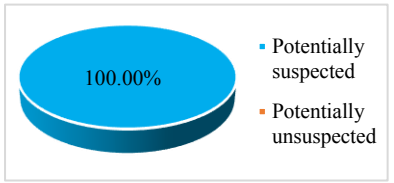
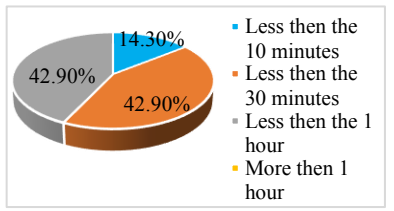
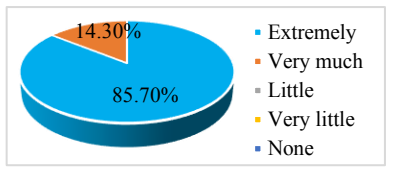
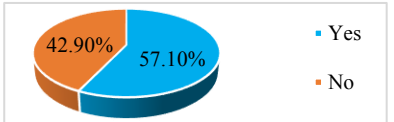
1. Is it helpful for your work to group declarants into clusters based on their declared assets over a range of years?	
2. Do you consider a declarant as a "suspect" when he/she has a substantial increase of property in comparison from the previous year, while his/her income or loans have not increased?	
3. How much time does it take for you to compare the declarations of the last two years for a given declarant?	
4. How much would it facilitate your day-to-day work a system like this if it would be able to generate a list of "suspected" declarants?	
5. Should a system like this compare the new declarants against the existing ones, in order to predict their potential behavior?	

Figure 4 User survey resulsts

VI. Conclusion

In this paper, we presented an approach for identifying unusual trends of growing of the assets of the public officials, who commonly declare their assets at anti-corruption agencies. Further, we presented a mathematical model for the problem of asset analysis, which is based on the practicalities of the anti-corruption institutions. Based

on the experimental study with real users, our *k-means* based approach achieves a 90% of accuracy, whereas our *decision tree algorithm* implementation can help in figuring out whether any new public official could have affinities for becoming a "suspected" declarant in the future.

As part of future work, we plan to extend our proposed approach by hybridizing (or partial substitution) of the *k-means* and *decision tree algorithm* with other grouping and prediction algorithms.

ACKNOWLEDGMENT

We thank a group of seven experts from the Anti-Corruption Agency of Kosovo for manual evaluation of the dossiers of the 10 declarants, whose results are used for evaluation of the proposed approach in this paper.

References

[1] T. Hoppe, H. Papa, et al., "Income and Asset Declarations in Practice", Danilovgrad, Montenegro: Regional School of Publik Administration - ReSPA, 2013.

[2] B. Dragusha, "Zhvillimi i algoritmeve për krahasimin e të dhënave të deklarimit të pasurive," University of Prishtina - Faculty of Electrical and Computer Engineering, Prishtina, 2017.

[3] Nhien An Le Khac, Sammer Markos, M. O'Neill, A. Brabazon, and M-Tahar Kechadi, "An investigation into Data Mining approaches for anto money laundering," in *2009 International Conference on Computer Engineering and Applications*, Sindapore, 2009.

[4] Zengan Gao, Mao Ye, "A framework for data mining-based ant-money laundering research," *JML,* vol. 10, 2007.

[5] Jun Tang, Jiam Yin, "Developing an intelligent data discriminating system of anti-money laundering based on SVM," in *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, 2005.

[6] Z. Anna, "Acceleration of K-means Clustering by K-dijkstra method for Graph Partitioning," Nara Institute of Science and Technology, NAIST-IS-MT1351213, 2015.

[7] K. Sylejmani, J. Dorn and N. Musliu, "Tourist trip planning: solo versus group traveling," in *30th workshop of the UK Planning andScheduling Special Interest Group*, 2012.

[8] D. Cieslakiewicz, "Unsupervised asset cluster analysis implemented with parallel genetic algorithms on the nvidia cuda platform," University of the Witwatersrand, Johannesburg, 2014.

[9] Z. Zhang, "K-means Algorithm," The University of Lowa, Lowa USA, 2012.

[10] "Decision trees, entropy, information gain, ID3," 2009. [Online]. Available: http://euclid.nmu.edu/~mkowalcz/cs495f09/slides/lesson015.pdf. [Accessed 10 06 2017].

[11] E. Alpaydim, Introduction to Machine Learning, 2nd ed. ed., London, London: Messachusetts Institute of Technology, 2010.

[12] K. A.-C. Agency, "Anti-Corruption Agency," [Online]. Available: http://www.akk-ks.org/sq/deklarimet#indexmain. [Accessed 7 January 2018].