Journal of Theoretical and Applied Information Technology
31st May 2015. Vol.75. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

THE DEVELOPMENT OF A TURNOVER PREDICTIVE MODEL APPLICATED TO THE ESTATE SECTOR IN MOROCCO

[1,*]RACHID ELMANSOURI, [1]OMAR ELBEQQALI, [2]HAMID BENNIS
[1]LIIAN/GRMS2I, FSDM, Sidi Mohamed Ben Abdellah University, Fes, Morocco
[2]LITEN, FST, Hassan 1st University, Settat, Morocco
E-mail : [1,*]elmansouri.rachid@gmail.com, [1]omarelbeqqali@gmail.com, [2]hamid.bennis@gmail.com

ABSTRACT

We present in this paper a descriptive statistical analysis of a set of data using the method of Principal Component Analysis PCA. Then, we discuss the construction of a prediction model based on decision trees by CHAID method. Then, we move to the evaluation of our model and we discuss the trees optimization problem. Comparative experiments between fifteen trees showed that our model gives interesting predictive performance. Finally, we present concluding remarks about decision trees and we propose development prospects.

Keywords: *Prediction, Decision Trees, Classification, Machine Learning, Principal Components Analysis*

1. INTRODUCTION

The construction of decision trees from data is a technique that has begun to be treated in 1963 by building the first regression tree in the process of prediction and explanation. On the basis of Morgan and Sonquist work [1], a whole literature has emerged, it was followed by a whole family of methods, extended to the problems of discrimination and classification, which were based on the same paradigm of representation by THAID trees [2] and CHAID (CHi-squared Automatic Interaction Detector) [3]. This approach has culminated with the CART method of Breiman (Classification And Regression Tree) [4].

In machine learning, most studies are based on information theory. It is customary to quote the method ID3 (Induction of Decision Tree) of Quinlan [5], who himself, relates his work to Hunt's [6]. Since 1993, Quinlan approach has taken an important turn when he presented the C4.5 methods [7] and C5.0 which are essential references in this field.

In this paper, we are interested in a classification problem where we want to explain and predict the value taken by a Boolean variable from a variables set, called predictive or explanatory variables. In the machine learning terminology, we are in the supervised learning framework. We will not discuss other types of use that are regression trees, when it is still a prediction problem but the variable to predict is continuous [8], neither classification trees, where the objective is to build homogeneous groups in the descriptors space [9].

This paper is organized as follows. In the second section, from an array of 20 variables and 4285 observations, we carry a descriptive statistical analysis of different variables. Section 3 is devoted to the analysis of data using the Principal Components Analysis (PCA) technique. The construction of a prediction model based on decision trees and CHAID method is presented in the Section 4. Section 5 is dedicated to the evaluation of our model. Optimization trees are treated at section 6. In Section 7, we review the advantages and disadvantages of decision trees. Section 8 presents the "Grid Computing" technology and calls the Moroccan scientific community to intensify the exploitation of the national grid "MaGrid". The last section comes as a conclusion.

2. STATISTICS AND DESCRIPTIVE STUDY

2.1 Description of variables

Before starting our predictive study, it is advised to conduct a descriptive study that will allow us to a better understand of data and the variables behavior that will be all hidden for confidential reasons. The following Table summarizes some descriptive statistical properties of numerical variables:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1 : Descriptive statistics.

	Effective	Minimum	Maximum	Average	Standard Deviation
V_5	4 285	8	5 220	134,41	117,018
V_6	2 113	8	5 218	135,79	142,319
V_7	4 285	6 000	5 649 600	279 856,81	284 685,817
V_8	2 749	6 000	5 649 600	252 966,95	274 016,428
V_9	4 285	59	16 000	2 235,72	1 619,437
V_10	4 285	6 000	5 649 600	290 506,77	293 127,720
V_12	4 285	20	113	44,62	12,216
V_13	4 285	0	190 000	6 360,54	8 910,420
V_15	4 205	0	12	2,03	1,918
V_17	4 285	-2 000	5 649 600	134 195,68	198 143,400
V_19	4 285	0	1	,34	,472
V_20	4 285	Feb.- 06-1984	Sept.-04-2013	Jan.- 15 - 2011	850

2.2 Computation and Analysis of the Inter-Correlations Matrix

The inter-correlation matrix concerns 19 variables, it should be noted that a variable which its type is date has been eliminated in this computation.

	V_1	V_2	V_3	V_4	V_5
V_1	1.000	-0.018	0.257	0.318	0.031
V_2	-0.018	1.000	0.116	-0.123	-0.113
V_3	0.257	0.116	1.000	0.116	0.021
V_4	0.318	-0.123	0.116	1.000	0.028
V_5	0.031	-0.113	0.021	0.028	1.000
V_6	0.026	-0.115	0.015	0.027	**0.998**
V_7	0.248	0.014	0.119	0.153	0.699
V_8	0.246	0.013	0.118	0.154	0.698
V_9	0.375	0.306	0.222	0.244	-0.109
V_10	0.246	0.005	0.059	0.136	0.654
V_11	-0.086	-0.058	-0.046	-0.003	0.038
V_12	-0.006	-0.074	-0.019	0.021	0.037
V_13	0.077	-0.025	0.084	0.102	0.215
V_14	-0.082	-0.014	-0.046	-0.069	-0.027
V_15	-0.083	-0.063	-0.046	-0.019	0.002
V_16	-0.031	-0.046	-0.018	0.013	0.057
V_17	0.237	-0.009	0.114	0.229	0.303
V_18	0.104	0.099	0.089	0.107	-0.081
V_19	0.324	0.008	0.020	0.295	-0.088

	V_6	V_7	V_8	V_9	V_10
V_1	0.026	0.248	0.246	0.375	0.246
V_2	-0.115	0.014	0.013	0.306	0.005
V_3	0.015	0.119	0.118	0.222	0.059
V_4	0.027	0.153	0.154	0.244	0.136
V_5	**0.998**	0.699	0.698	-0.109	0.654
V_6	1.000	0.695	0.697	-0.113	0.652
V_7	0.695	1.000	**0.999**	0.386	**0.894**
V_8	0.697	**0.999**	1.000	0.385	**0.895**
V_9	-0.113	0.386	0.385	1.000	0.336
V_10	0.652	**0.894**	0.895	0.336	1.000
V_11	0.038	0.008	0.008	-0.033	0.033
V_12	0.039	0.030	0.031	-0.005	0.034
V_13	0.213	0.342	0.342	0.129	0.338
V_14	-0.024	-0.056	-0.055	-0.048	-0.073
V_15	0.006	-0.039	-0.037	-0.032	-0.035
V_16	0.059	0.057	0.057	-0.020	0.085
V_17	0.298	0.493	0.492	0.180	0.515
V_18	-0.091	-0.092	-0.097	0.046	-0.109
V_19	- 0.093	-0.149	-0.151	-0.030	-0.181

	V_11	V_12	V_13	V_14	V_15
V_1	-0.086	-0.006	0.077	-0.082	-0.083
V_2	-0.058	-0.074	-0.025	-0.014	-0.063
V_3	-0.046	-0.019	0.084	-0.046	-0.046
V_4	-0.00	0.021	0.102	-0.069	-0.019
V_5	0.038	0.037	0.215	-0.027	0.002
V_6	0.038	0.039	0.213	-0.024	0.006
V_7	0.008	0.030	0.342	-0.056	-0.039
V_8	0.008	0.031	0.342	-0.055	-0.037
V_9	-0.033	-0.005	0.129	-0.048	-0.032
V_10	0.033	0.034	0.338	-0.073	-0.035
V_11	1.000	-0.012	0.072	0.085	0.095
V_12	-0.012	1.000	0.019	**0.301**	**0.554**
V_13	0.072	0.019	1.000	0.055	0.030
V_14	0.085	**0.301**	0.055	1.000	**0.635**
V_15	0.095	**0.554**	0.030	**0.635**	1.000
V_16	0.039	-0.001	-0.001	-0.018	-0.028
V_17	-0.015	0.033	0.286	-0.016	-0.033
V_18	-0.059	-0.012	0.019	-0.017	-0.039
V_19	-0.053	0.043	0.011	-0.019	-0.036

	V_16	V_17	V_18	V_19
V_1	-0.031	0.237	0.104	0.324
V_2	-0.046	-0.009	0.099	0.008
V_3	-0.018	0.114	0.089	0.020
V_4	0.013	0.229	0.107	0.295
V_5	0.057	0.303	-0.081	-0.088
V_6	0.059	0.298	-0.091	-0.093
V_7	0.057	0.493	-0.092	-0.149
V_8	0.057	0.492	-0.097	-0.151
V_9	-0.020	0.180	0.046	-0.030
V_10	0.085	0.515	-0.109	-0.181
V_11	0.039	-0.015	-0.059	-0.053
V_12	-0.001	0.033	-0.012	0.043
V_13	-0.001	0.286	0.019	0.011
V_14	-0.018	-0.016	-0.017	-0.019
V_15	-0.028	-0.033	-0.039	-0.036
V_16	1.000	0.006	-0.031	-0.074
V_17	0.006	1.000	0.441	0.294
V_18	-0.031	0.441	1.000	0.464
V_19	-0.074	0.294	0.464	1.000

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The analysis of the inter-correlations matrix allowed us to make the following decisions:

The two variables "V_5 and V_6" are strongly correlated linearly and positively, they have a correlation coefficient is equal to 0.998. Later in the study, we keep only the variable V_6.

The same for variables "V_7, V_8 and V_10", about their inter-correlation coefficients are the order of 0.9. For further study, we keep only the variable V_10 as a variable that influences mostly a customer / citizen.

It is worth mentioning that there is a certain correlation between the variables "V_12, V_14 and V_15" that is explained by the natural evolution of family situation and children number of an individual depending on his age.

3. DATA ANALYSIS USING PCA

3.1 Brief Description of the PCA

The Principal Components Analysis (PCA) is a part of the descriptive multidimensional methods called factorial methods. The methods that have emerged in the early of 30s were mainly developed in France in the 60s, especially by Jean-Paul Benzécri who operated many geometric and graphical representations. The PCA is a technique that focuses on the correlation matrices analysis. It is a set of methods to perform linear transformations of a large number of intercorrelated variables to obtain a number of components uncorrelated wich is relatively lower [10].

These methods exploit many geometric and graphical representations in so far as they are descriptive methods. They are not based on a probabilistic model, but they depend on a geometrical model [11].

The PCA offers, from a rectangular array of data (a table in the data warehouse environment) with the values of p quantitative variables (also called attributes or columns), for n units (also called individuals, rows or records), geometric representations of these units and these variables [12].

This approach facilitates analysis by grouping the data into smaller sets and eliminates problems of multicollinearity between variables.

The basic idea of the PCA is to explain the observed variance in the mass of initial data with a limited number of components, defined as a pure and simple mathematical transformations of the original variables. The algorithm used for the determination of these components verifies two important constraints: First, the extracted components must capture a proportion of variance which becomes less important. Afterward, the components must have zero linear correlation (orthogonality condition).

Ideally, we hope that the first component C1 captures a very significant proportion of the variance contained in the original data; so 70% or 60% of variance explained by the first component will be an interesting result. However, the reality is often less rewarding and the first extracted component C1 explains 40% or even less. Variance remaining unexplained by the first component C1, is not left out in the PCA. In contrast, it is subjected in turn to the same extraction process components.

We can discern various implications of this analysis. Firstly, the cumulative proportion of total variance across the different components may eventually get 100% if the number of extracted components equals the number of initial variables. But knowing that the main objective of the PCA is to reduce the amount of data. It is therefore inappropriate to extract the same number of components as original variables. In other words, we should make a wise decision about the number of principal components to extract.

In previous work [12], we studied the classical PCA, for which all individuals have the same weight in the analysis and all variables are treated symmetrically. This can be problematic. The first criticism made by practitioners is that: if the old variables are heterogeneous, it is difficult to make sense to the principal components which are linear combinations of heterogeneous variables. The second criticism is that: if we change units on these variables, we can completely change the outcome of the PCA. The last criticism is that: a variable contributes even more to the making of the first axis if its variance is high [10].

For all these reasons, we decided to go in this work for a PCA on centered and reduced variables, we speak about normed PCA.

3.2 Computation and Interpretation of Eigenvalues

By examining the correlation matrix, we notice that the presence of a strong correlation is between some variables.

Table 2 shows the evolution of eigenvalues and inertia percentages of the data table. We note that the eigenvalue associated to the first component C1 is equal to 4.75, which corresponds to 25.04 % of the total variance that is equal to 19. The second component C2 explains 2.35 units of variance, which corresponds to 12.38 % of the total variance. Therefore, we can say that after extracting two principal components, we can reduce the data from

Journal of Theoretical and Applied Information Technology
31st May 2015. Vol.75. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

19 to 2 dimensions keeping successfully 37.43 % of the initial variance.

The third component C3 explains 1.99 units of variance, which corresponds to 10.47 % of the total variance, and thus a total of 47.91 % of variance explained by the first three components.

It should be remembered that we have worked on centered and reduced variables, so each variable has 1 unit of variance.

Table 2: Evolution Of Eigenvalues And Percentages Of Cumulative Inertia.

Components	Eigenvalues	Inertia (%)	Cumulative inertia (%)
C1	4.75	25.04	25.04
C2	2.35	12.38	37.43
C3	1.99	10.47	47.91
C4	1.56	8.25	56.16
C5	1.17	6.17	62.34
C6	1.05	5.57	67.91
C7	0.98	5.16	73.08
C8	0.90	4.75	77.84
C9	0.84	4.46	82.30
C10	0.68	3.61	85.92
C11	0.66	3.52	89.44
C12	0.64	3.38	92.83
C13	0.40	2.13	94.96
C14	0.34	1.84	96.80
C15	0.28	1.49	98.29
C16	0.19	1.05	99.34
C17	0.12	0.64	99.98
C18	0.002	0.01	99.99
C19	0.0002	0.000012	100
Total :	19	100	

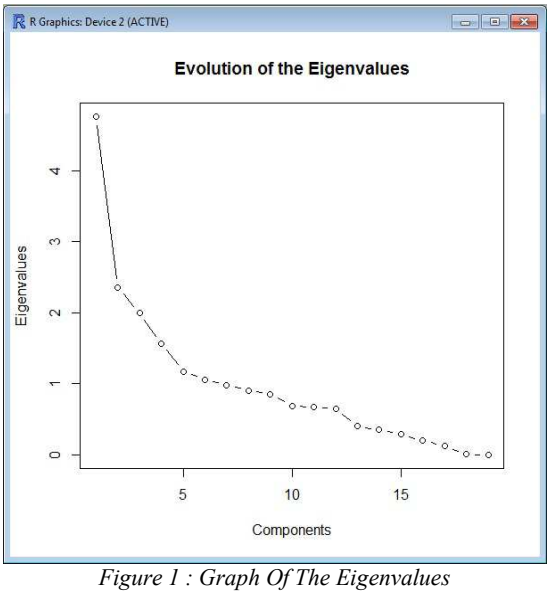

Figure 1 : Graph Of The Eigenvalues

3.3 Correlation Circle: Initial Variables - Compenents C1&C2

It is interesting to see how the old variables are related to new variables. That is why the correlations of the old variables with the new variables are calculated.

To represent the old variables, we take as coordinates of these variables, their correlation coefficients with the new variables. Then we obtain what is commonly called the "circle of correlation", a name which comes from the fact that a correlation coefficient ranging between -1 and +1 [10].

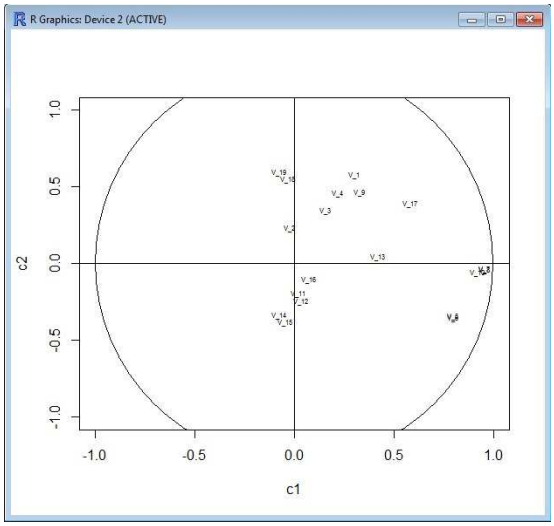

Figure 2: Correlation Circle

By examining the correlation circle, we can say that the variables (V_5, V_6, V_7, V_8 and V_10) are well represented on the plan (C1, C2) because they are near to the circle edge.

Variables V_5 and V_6 are highly correlated linearly and positively. Same for the variables V_7, V_8 and V_10. It should be noted that these results are consistent with those found on the inter-correlation matrix.

The other variables are not well represented (away from the circle edge), so we can not say anything about them.

3.4 Projection of Individuals on the First Principal Plan

The first principal plan is spanned by the first two principal components; it keeps 37.43% of the total inertia contained in the data table.

Journal of Theoretical and Applied Information Technology
31st May 2015. Vol.75. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

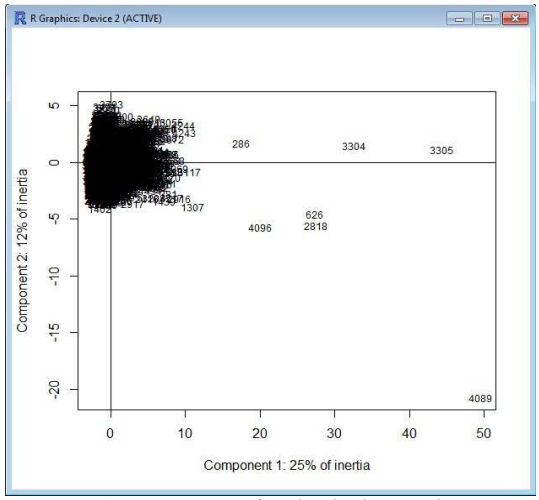

Figure 3: Projection Of Individuals On The First Principal Plane

This projection does appear a few outliers from the rest of the treated population, these are the following observations which deserve special attention: 286, 626, 2818 3304, 3305, 4089 and 4096.

3.5 Projection of Individuals on the 3D Vector Space Spanned by the First Three Principal Components

The projection of individuals on the 3D space returns to calculate the new coordinates of these vectors/individuals in the new basis of principal components.

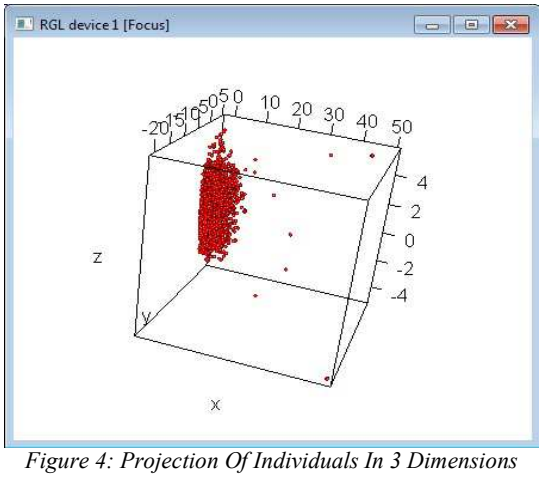

Figure 4: Projection Of Individuals In 3 Dimensions

The vector space spanned by the first three principal components captures 47.91 % of the total inertia contained in the data table. Visualization of the individuals projected onto this 3D space, allows us to get more information about their behavior and therefore a better decision compared to the 2D projection.

4. THE CONSTRUCTION OF PREDECTING MODEL

For trees decision induction and the construction of our prediction model, we worked on an array of 20 variables and 4285 observations. We used the software Revolution-R which has an interesting library of tree induction algorithms. It should be noted that interactive modules for trees induction that allow the user to manually intervene in the process of building, the model is always available in commercial software such as COGNOS, SPSS and STATISTICA; these programs have greatly helped to popularize decision trees with practitioners.

In this work, we have been forced to reduce the data volume to 20 variables and 4 285 observations given the limited computation power available on the machine. The use of the computing national grid "MaGrid" is provided for processing a larger volume of data.

It is clear that we have selected the CHAID method of learning and defines the problem to be addressed by selecting descriptors (explanatory variables/predictors) and the attribute to predict (the turnover variable); Then, we randomly divided the file in part learning 67% (2 871 individuals) and test 33% (1 414 individuals).

The following tables summarize the specifications and the results of the treatments we have executed. The three stop induction rules are:

- The maximum depth of the tree.
- The minimum size of the summit to segment.
- The minimum size of a child vertex generated.

Table 3: Induction With Maximum Depth Of 3 Levels

Specifications	**Maximum depth of the tree**	3	3	3	3	3
	Minimum observations of a parent node	100	50	20	10	5
	Minimum observations of a child node	50	25	10	5	2
Results	**Number of nodes in the tree**	18	25	43	34	43
	Number of terminal nodes (leaves)	13	18	27	22	29
	Effective depth of the tree	3	3	3	3	3
Overall percent correct	**Learning sample**	94,5	95,5	97	96,5	97,1
	Test sample	92,4	94,7	96,1	96	95,7

Table 4 : Induction with maximum depth of 4 levels

Specifications	**Maximum depth of the tree**	4	4	4	4	4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	Minimum observations of a parent node	100	50	20	10	5
	Minimum observations of a child node	50	25	10	5	2
Results	Number of nodes in the tree	20	27	41	49	54
	Number of terminal nodes (leaves)	14	18	26	32	34
	Effective depth of the tree	3	4	4	4	4
Overall percent correct	Learning sample	95	95,1	97	97,8	97,8
	Test sample	93,9	94,6	96,5	97,6	96,2

Table 5 : Induction With Maximum Depth Of 5 Levels

Specifications	Maximum depth of the tree	5	5	5	5	5
	Minimum observations of a parent node	100	50	20	10	5
	Minimum observations of a child node	50	25	10	5	2
Results	Number of nodes in the tree	23	29	48	44	73
	Number of terminal nodes (leaves)	15	19	31	28	46
	Effective depth of the tree	4	4	5	4	5
Overall percent correct	Learning sample	94,5	95,8	97	97,6	98,4
	Test sample	93,2	93,5	95,9	96,5	97,4

The experiments performed in this study have shown that the error rate calculated based on the training data constantly decreases as the number of nodes (the tree size) increases. Contrariwise, the error rate relied on the test shows first a decrease, then we observe that the error rate remains on a plateau before gradually rising while the tree is clearly oversized, we speak about a phenomenon of over-learning.

An important step after the construction of the tree is to extract decision rules by assigning a conclusion to each leaf of the tree. The path from a leaf to the root of the tree can be read as a prediction rule such as "If Condition... Then Conclusion".

When the leaf is pure, assign the conclusion corresponding to the single modality has seems natural. Contrariwise, when several modality are present in the leaf, we must use a rule of efficient allocation. The rule most often used is the majority rule: we assign to the leaf the modality of the variable to predict who has the actual greatest.

5. EVALUATION OF THE PREDICTION MODEL

A classic way to evaluate the quality of learning is to compare the model predictions with observed values on a sample of the population. This comparison is summarized in a table called confusion matrix. It is possible to extract synthetic indicators, the most famous is the error rate or misclassification rate. It is possible to interpret it as an average cost of misclassification, it is also possible to interpret it as an estimator of the probability to make a wrong prediction using the decision tree [13].

The main interest of the error rate is that it is objective; it is typically used to compare learning methods on a given problem. For an unbiased indicator, it is imperative not to measure the sample used to develop the model. For that, the practitioner often set aside a sample (test sample), said test will be used to evaluate and compare models.

The following table shows the confusion matrix relative to the last tree induced characterized by a maximum depth of 5 levels, a minimum size of the parent node is equal to 5 and a minimum size of a child node is equal to 2.

Table 6 : Confusin Matrix

Sample	Observations	Previsions		
		0	1	Correct percentage (%)
Learning	0	1 858	37	98,0
	1	9	964	99,1
	Overall percentage (%)	65,1	34,9	98,4
Test	0	923	28	97,1
	1	9	457	98,1
	Overall percentage (%)	65,8	34,2	97,4

6. OPTIMIZATION OF THE TREE SIZE

The performance of a decision tree based primarily on determining the optimal size [4]. Trees

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

tend to produce a "classifier" too complex, overly sticky data; this is the phenomenon of over-learning. The leaves, even if they are pure, are composed of many few individuals to be reliable in the prediction. It was also demonstrated that the tree size tends to increase with the number of observations in the training set [14]. Hence, the need to establish a sufficiently effective rule to ensure the best performance of the decision tree.

The issue of finding the optimal size is to stop (pre-pruning) or reduce (post-pruning) tree to get a good classifier. Got the subject importance, we are going to detail the pre-pruning method implemented by CHAID together with post-pruning.

6.1 Pre-Pruning

It involves setting a stop rule which allows to stop the construction of the tree during the construction phase. A very simple approach is to set a local stop criterion on the node which is being treated, that evaluates the informational contribution of segmentation that we are going to initiate. This evaluation approach is local to a node because it ignores the overall behavior of the tree. However, it gives good results [13].

Other more empirical criteria to the leaf size can be implemented. The objective is to avoid the appearance of nodes in numbers too small to expect a reliable prediction. They rely largely on intuition of the practitioner. They can also be fixed by performing the tests: the first strategy is to set a size from the node which we no longer attempt to realize segmentation; the second is equivalent to setting an effective eligibility: If one of the produced leaves by the segmentation is smaller than a threshold that can be set, the operation is refused.

Rather empirical nature, these rules stop proved practices in the implementation of decision trees in actual studies [13].

6.2 Post-Pruning

This approach appeared with CART method [4]. It was very widely adopted in various forms thereafter. The idea is to build the tree in two phases: first phase of expansion, which attempts to produce all possible pure trees, therefore we accept all segmentations even if they are not relevant; in a second phase, we try to reduce the tree using another criterion for comparing trees with different sizes. Construction time of the tree is of course higher; it may be disadvantageous when the database is very large; in return, the objective is to obtain a more efficient tree ranking.

Two opposing approaches exist in the literature. The first, based on Bayesian formulation which transforms the learning problem to an optimization problem. The criterion reflects the compromise between the complexity of the tree and its ability to stick to the data [15]. Inspite of the used formulation elegance, it must be recognized that these methods are not well known; they are also implemented in some programs [16],[17].

The methods of the second approach are based on an unbiased estimate of the classification error rate during the pruning phase, these methods are more widespread. Some methods use a calculated estimation on the same learning sample, but penalized by the size of the actual treated node (case of C4.5) [7]; other methods use an evaluation of the error rate with a second sample, called "pruning set" (CART) [18].

The parallel between these two approaches was carried out in an article published by two important authors in the field of trees [19]. The first method is known in the world of machine learning; the second is more famous among statisticians.

7. ADVANTAGES AND DISADVANTAGES OF DECISION TREES

The decision tree induction is a matured technique. Trees have comparable performance to other supervised methods; many empirical comparisons have shown enough [20],[21]. The method is non-parametric; it does not postulate a priori hypothesis on data distribution; it is resistant to outliers; The prediction model is non-linear. When the training set is large, it has properties similar to nearest neighbor algorithms [4].

However, the first complaint that we can send to decision trees is their inability, with conventional algorithms (C4.5, CART, CHAID, etc..) To detect combinations of variables; This is due to the principle of step by step construction of the tree that drives a certain "myopia". The second complaint is the need for a large sample of learning. The trees can certainly reproduce approximately all forms of borders, but at the cost of rapid fragmentation of data, with the danger of producing leaves with very few individuals. Corollary to this, the trees are generally unstable; terminals discretization especially in the lower parts of the tree are linked to a high variability. Thus, some researchers recommend proceeding with the prior discretization of variables before the construction of the tree [22].

The decision tree induction is capable of handling indiscriminately continuous and discrete data. It has more of a natural selection mechanism variables. It should be preferred when working in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

areas where the number of descriptors is high, some in large numbers, are irrelevant. We also need to put this assertion. In fact, work in the field of variable selection showed that the prior reduction of the descriptors in very noisy areas significantly improves the performance of decision trees [23].

Finally, the last point of differentiation, which provides much of the popularity charts among practitioners: their ability to produce a simple and directly usable knowledge, to the scope of the uninitiated. A decision tree can be read and interpreted directly; it is possible to translate basic rules without losing information. This quality is enhanced by the ability of the expert to intervene directly in the process of creating the prediction model. The use of the tool by experts in the field at the same time provides a better interpretation and comprehensibility of the results [24].

8. THE "GRID COMPUTING" TECHNOLOGY

The word grid was chosen by analogy with the American electricity distribution system (electric power grid), this term was widespread in 1998 by the work of Ian Foster and Carl Kesselman [25]. Indeed, a grid can be seen as an instrument that provides the computational power and/or the storage capacity in the same manner as the electric network of the electrical power supplies. The inventors vision of this term is that it will be possible eventually to plug into a computer grid for the computational power and/or data storage without knowing where or how that power is provided to picture of what happens to electricity. The analogy with the electricity distribution system identifies the grid vision of a user view.

Grids are increasingly used, both in the world of research and in industry. It is indeed a powerful tool for addressing difficult problems involving many large data. However, designing applications for this system type is complex. Applications are made to geographically distributed running on the nodes belonging to the grid process.

Scientists use these data to model the environment. Analyzing models and simulating their evolution. It becomes possible to understand complex phenomena and systems or inventing new ones, and even anticipate their possible developments. Simulate the evolution of such models therefore requires computing power and storage growing.

Faced with increasing power demand, the Moroccan Computing Grid ‘’MaGrid’’ and in the context of national policy support for scientific and technical research, CNRST was charged by the government through the plan 2000-2004 to install a "data center" which goal is to provide the Moroccan scientific community a platform for intensive computing and processing large amounts of data to enable researchers teachers, PhD and Master students, to improve their opportunities for cooperation and be more qualified to integrate high-level projects.

In this context, CNRST chose the technology called "Grid Computing" and set up the National Grid computing "MaGrid" by deploying its first cluster in 2006 with 20 CPUs and 10 Tb of storage.

9. CONCLUSION

We have shown in this paper the interest of conduct statistical studies and develop predictive models applied to productive sectors in Morocco as the estate area. At first, we conducted a descriptive statistical study before moving on to the construction of our model based decision trees and especially the CHAID method. The evaluation of the models obtained showed the robustness and reliability of the method since the error rate obtained on the training and testing samples are respectively around 2% and 3% to a maximum tree depth of 5 levels.

In this work, a particular attention was paid to search for the optimal size decision trees and to avoid two uncomfortable situations: Too small trees that are too conservative and have lower performance. And trees too large to learn training data and have performances collapsing on test data.

It should be noted that the model developed in this paper does not take into account external factors, that can come to reverse the trends towards rising as in the direction of lower (for example: tax incentives, government policies, bank loans, etc ...).

An interesting perspective for this work is to develop other prediction models with other techniques, and make a comparison of the performance of different models. We recommend also that policymakers use these techniques to try to understand more and control the current and future situations of the phenomena that affect the urban scenery of the Moroccan cities such as the phenomenon of slums. However, these problematic phenomenon deserve further poly-disciplinary study.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES:

[1] J. Morgan, J.A. Sonquist, "Problems in the Analysis of Survey Data, and a Proposal", Journal of the American Statistical Association, 58, 1963, pp. 415-435.

[2] J. Morgan, R. Messenger, "THAID-a sequential analysis program for the analysis of nominal scale dependent variables", Survey Research Center, University of Michigan, 1973.

[3] G. Kass, "An exploratory technique for investigating large quantities of categorical data", Applied Statistics, 29(2), 1980, pp. 119-127.

[4] L. Breiman, J. Friedman, R. Olshen, C. Stone, "Classification and Regression Tree", Wadsworth International Group, 1984.

[5] R. Quinlan, "Discovering rules by induction from large collections of examples", D. Michie ed., Expert Systems in the Microelectronic age, 1979, pp. 168-201.

[6] E. B. Hunt, "Concept Learning: An Information Processing Problem", Wiley, 1962.

[7] R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufman Publishers, 1993.

[8] L. Torgo, "Inductive Learning of Tree-Based Regression Models", PhD Thesis, Department of Computer Science, University of Porto, 1999.

[9] M. Chavent, C. Guinot, Y. Lechevallier, M. Tenenhaus, "Méthodes divisives de classification et segmentation non supervisée : recherche d'une typologie de la peau humaine saine", Revue de Statistiques Appliquées, XLVII (4), 1999, pp. 87-99.

[10] C. Duby, S. Robin. "Analyse en Composantes Principales", Institut National Agronomique Paris-Grignon. Département O.M.I.P. 10 Juillet 2006.

[11] A. Bouchier, "L'analyse des données à l'usage des non mathématiciens", INRA. Formation permanente. Janvier 2006.

[12] R. Elmansouri, E. Ziyati, D. Aboutajdine, and O. Elbeqqali, "The fragmentation of datawarehouses: an approach based on principal components analysis", ICMCS'12. Tanger. 10-12 Maï 2012.

[13] R. Rakotomalala, "Arbres de décision", Revue MODULAD, 33, 2005, pp. 163-187.

[14] T. Oates, D. Jensen D, "The effects of Training Set Size on Decision Tree Complexity", Proceedings of 14th International Conference on Machine Learning, 1997, pp. 254-262.

[15] C. Wallace, J. Patrick, "Coding Decision Trees", Machine Learning, 11(1), 1993, pp. 7-22.

[16] W. Buntine, "About the IND tree package", Technical Report, NASA Ames Research Center, Moffet Field, California, September 1991.

[17] R. Kohavi, D. Sommerfield, MLC++. In Will Klosgen and Jan M. Zytkow, editors, "Handbook of Data Mining and Knowledge Discovery", chapter 24.1.2, Oxford University Press, 2002, pp. 548-553.

[18] L. Breiman, "Bagging Predictors", Machine Learning Journal, 24(2), 1996, pp. 123-140.

[19] R. Kohavi, J. Quinlan, "Decision-tree Discovery", in Handbook of Data Mining and Knowledge Discovery, Klösgen and Zytkow Editors, 2002, pp. 267-276.

[20] D. Zighed, R. Rakotomalala, "Graphes d'Induction : Apprentissage et Data Mining", Hermes Science Publications, 2000.

[21] T. Lim, W. Loh, Y. Shih, "A comparison of prediction accuracy, complexity and training of thirty-three old and new classification algorithms", Machine Learning Journal, 40(3), 2000, pp. 203-228.

[22] J. Dougherty, R. Kohavi, M. Sahami, "Supervised and unsupervised discretization of continuous attributes", Proceedings of 12th International Conference on Machine Learning, 1995, pp. 194-202.

[23] L. Yu, H. Liu, "Efficiently Handling Feature Redundancy in High-Dimensional Data", Proceedings of International Conference on Knowledge Discovery and Data Mining, 2003, pp. 685-690.

[24] R. Rakotomalala, "Graphes d'Induction", PhD Thesis, Université Claude Bernard Lyon 1, 1997.

[25] C. Kesselman and I. Foster. "The Grid: Blueprint for a New Computing Infrastructure", Morgan Kaufmann Publishers, November 1998.