Flavia CAIA, PhD Candidate
E-mail: flavia.caia@cig.ase.ro
Professor Pavel NĂSTASE, PhD
E-mail: nastase.pavel@ase.ro
Department of Management Information Systems
The Bucharest University of Economic Studies

MODELING A BUSINESS INTELLIGENCE SYSTEM FOR INVESTMENT PROJECTS

Abstract. *The phases of evaluation and selection of investment projects require marketing, technical and financial information analysis before the decisions are made. In order to meet strict legal environmental and fiscal regulations it is necessary to assess the environmental and social impact. These analysis may indicate alternative sites, models, technologies, and methods of implementation as solutions for executing the project, which often involve higher costs, and cause delays in the feasibility analysis and project selection.*

The purpose of this empirical research is to reduce the effort and duration of the selection and evaluation of projects and project portfolio optimization, increasing the efficiency and accuracy of budget implementation and controlling in accordance with the investment strategy of the company.

The results consist of a set of requirements for a specific business intelligence system, designed to improve the operational flow and the performance of the planning process, and an analysis of significant factors underlying performance differences between projected and actual achievements, to define the parameters of the model.

Keywords: *business intelligence, investments, budgeting, projects, requirements, model.*

JEL Classification: C61, D22, O22

1. INTRODUCTION

The ability to make assertive decisions in a highly competitive environment makes the difference between attaining and maintaining leader positions and losing market share. Managers can make decisions based on intuition, but the exponential increase of data volumes and sources makes automated tools essential for providing the necessary information (Menéndez and da Silva, 2014).

Business intelligence (BI) can be defined as a set of mathematical models and analysis methods to systematically exploit the available data to recover useful information and knowledge to support complex decision-making processes (Vercellis, 2009).Business intelligence refers to various concepts, methods, processes and software applications designed to improve business decisions by analyzing raw data, querying, making aggregations, forecasting and correlations, data mining, online analytical processing, and producing reports (Lahrmann, G., 2010).

The phases of evaluation and selection of investment projects require marketing, technical and financial information analysis before the decisions are made. In order to meet strict legal environmental and fiscal regulations it is necessary to assess the environmental and social impact. These analysis may indicate alternative sites, models, technologies, and methods of implementation as solutions for executing the project, which often involve higher costs, and cause delays in the feasibility analysis and project selection (Davenport and Harris, 2007).

In this context, this study proposes a model of a BI system for analyzing projects in an integrated manner, by taking into account a specified set of indicators used in the decision making process.

Management should evaluate alternative long-term strategies of investments, and then decide how to manage the investment projects in order to accomplish business objectives. The process of analysis and developing the capital budget is one of the most difficult, risky and important activities of which management is responsible. This task involves projections and estimates, and the management's decisions regarding the capital budget impact the company for many years (Kudyba, 2014).

Adopting low-cost technologies for massive data storage and wide availability of internet connection resulted in organizations accumulating and collecting large amounts of data over the years. Businesses that are capable of transforming data into information and knowledge can make more effective decisions faster, thus gaining a competitive advantage. This goal can be achieved through technology, with the assistance of competent minds using advanced methodologies of analysis (Rasmussen, Goldy&Solli, 2002).

2. CONSIDERATIONS REGARDING MODELING A BUSINESS INTELLIGENCE SYSTEM

There are many factors of key importance for the success of a BI solution, but the first step is to elicit requirements in accordance to the needs of the users. This process consists two parts: viability study and elicitation.

Performing the viability study implies collecting inputs like user needs and the documentation regarding the application domain for which the BI solution is developed.

The purpose of the requirements elicitation stage is to identify the initial aspects necessary for developing the BI system. The process comprises four tasks, performed in a cycle until stakeholders agree upon the requirements and the models created by requirements engineers (Menéndez and da Silva, 2014):

1) Discovery: collecting the documentation regarding the application field and identify the project needs as a list of items, for which the corresponding requirement will be formulated;
2) Classification and Organization: eliminating duplicate, additional, unjustified requirements, then organizing and classifying the requirements;
3) Identification of Priorities: setting the order of implementation and delivery for the requirements; and
4) Creation of models: created to validate the accuracy and comprehensiveness of elicited requirements.

The main issues that require the implementation of a BI system imply:

- Most companies have high volumes of data, but not always accurate data;
- Data is not information;
- Data is often spread over several heterogeneous systems and is not consistent.

BI consolidates data and presents them in ways that help users (managers, programmers, analysts, operational employees, suppliers and customers) to make informed decisions faster.

Business Intelligence methodologies are interdisciplinary and comprehensive, covering many fields of application. Characteristics to BI solutions are: they affect the representation and organization of the decision-making process, and implicitly the domain of decision theory; the collection and storage of data involves data warehousing technologies; for optimization and data mining, operational research and statistics are used mathematical models; and, BI connects several application areas such as marketing, logistics, accounting and control, finance, services and public administration (Loshin, 2012).

A data warehouse consists of a set a related tables populated with subject-oriented, integrated, denormalized, time-variant and non-volatile data from more sources, which can be viewed and analyzed using a variety of tools.

The tables can be of two types: dimensions (i.e. employees and products, they answer to questions like *who, what, when, where* and hold values that describe facts), and facts (refer to the dimensions, hold numerical measures to quantify and answer to the question *how much?*). Both require a surrogate key, a new code usually, used instead of any composite key as primary key (Roebuck, 2012).

Data warehouses (DW) typically use a design called OLAP (On-Line Analytical Processing) which implies simpler structures, easier and faster to work with. The number of tables and connections is reduced, and data is denormalized. In contrast, OLTP (On-Line Transaction Processing) is designed to work with one record at a time, being much slower, and contains highly normalized tables, implying duplicated data is removed. Obtaining data of a transaction can involve multiple links between tables, and ad-hoc reporting can become quite confusing. Also OLTP system is much slower (Kimball & Ross 2013).

DW combine data from multiple systems, homogenize differences between the systems, makes reporting easier, reduce the load on production systems, ensure consistency of transitions between systems and ensure long-term storage of data.

Business intelligence systems tend to promote a scientific and rational approach to the management of complex organizations and companies. Even using an electronic spreadsheet to evaluate the effects on the budget of updating the exchange rate, requires managers a mental representation of the financial flows (Vercellis, 2009).

A business intelligence environment provides decision makers information and knowledge derived from data processing by applying mathematical models and algorithms. In some cases they may consist simply of aggregations and percentages, while more developed analyses use advanced optimization models, inductive learning and prediction. A model refers to a selective abstraction of a real system, designed to analyze and comprehend the behavior of the real operational system from an abstract point of view (Vercellis, 2009).

Rapid access to information is a valuable asset, but BI tools are oriented towards passive analysis, based on the pre-defined criteria of the decision maker. To use the enormous strategic potential of business intelligence methodologies, companies should return to active forms of support for decision making, based on systematic adoption of mathematical models capable of converting data into knowledge, not only information, and to trigger a real competitive advantage (Baan and Homburg, 2013).

Data mining refers to the process of discovering knowledge, revealing patterns and relationships from large volumes of complex data sets (De Veaux, 2000). A data mining algorithm is a tuple: *{model structure, score function, search method, data management techniques}*. By combining different model structures, score functions, methods and techniques can be obtained an infinite number of new different algorithms. However, the most influential algorithms used in data mining are C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART (Wu et al., 2008).

C4.5 takes as input a collection of cases, described by their values for a fixed set of attributes, and generates classifiers that can accurately predict to which class a new case belongs, expressed as a decision trees. This process uses a measure of the disorder of the data called “Entropy”. The Entropy of $\vec{y}$is calculated using the formula (Korting, 2006):

$$Entropy(\vec{y}) = -\sum_{j=1}^{n} \frac{|y_j|}{\vec{y}} \log \frac{|y_j|}{|\vec{y}|}$$

And iterating through all values of $\vec{y}$. The conditionalEntropy is defined as follows:

$$Entropy(j|\vec{y}) = \frac{|y_j|}{|\vec{y}|}\log\frac{|y_j|}{|\vec{y}|}$$

The purpose is to maximize the gain, which is defined by:

$$Gain(\vec{y}, \mathrm{j}) = \mathrm{Entropy}(\vec{y} - Entropy(j|\vec{y}))$$

The k-means is an iterative algorithm for clustering that partitions n *d*-dimensional real vector observations (**x**$_1$, **x**$_2$, …, **x**$_n$) into k (≤ n) user-specified sets **S**={S_1,S_2,…,S_k}, where every observation is added to the cluster with the nearest mean so as to minimize the within-cluster sum of squares (Bishop, 1995):

$$\arg\min_{S}\sum_{i=1}^{k}\sum_{x\epsilon S_i}\left\|x-\mu_i\right\|^2$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i. When the mean no longer changes, the algorithm converges.

SVM aims to find the best classification function in a two-class learning task by distinguishing between the items of the two classes. A linear classification function for a linearly separable dataset corresponds to a hyperplane$f\ (x)$ that separates the two classes through the middle. A new data instance x$_n$ can be classified using the sign of the function $f\ (x_n)$; if $f\ (x_n)\ >\ 0$, x$_n$ belongs to the positive class. SVM additionally guarantee that the best linear hyperplane function is determined by maximizing the margin function between the two classes, represented as follows (Wu et al., 2008):

$$L_P = \frac{1}{2}\|\vec{w}\| - \sum_{i=1}^{t}\alpha_i\, y_i(\vec{w} * \vec{x_i} + b) + \sum_{i=1}^{t}\alpha_i$$

$L_P$is called the Lagrangian, $\alpha_i, i\ =\ \{1,\ldots,t\}$are the Lagrange multipliers, non-negative numbers for which the derivatives of$L_P$with respect to $\alpha_i$are zero, and *t* is the number of datasets. In this equation, the hyperplane is defined by the vectors $\vec{\boldsymbol{w}}$and constant *b*.

The Apriori algorithm is designed to operate on databases containing transactions, for learning association rules. The algorithm is used to obtain frequent item sets Fk of size *k*, using candidate Ck generation, from a transaction dataset and determine association rules, with a confidence level equal or above a specified minimum confidence. The smaller the size of the candidate sets, the better the performance. It begins with scanning the database for frequent item sets of size 1, incrementing the count for each item, and collecting them if the minimum support requirement is satisfied. Afterwards, it iterates the following three steps to extract all the frequent item sets (Wu et al., 2008):

1. Generate candidates of frequent item sets $Ck + 1$, from the frequent item sets Fk, of size *k*.
2. Scan the database and compute the support of each candidate.
3. Collect those item sets that fulfill the minimum support requirement to $Fk + 1$.

The algorithm can determine *Ck*+1 from *Fk* by using the following two steps:

1. Join step: Generate the initial candidates of frequent item sets R_{K+1}, of size *k+1*, based on the union of two frequent item sets *Pk* and *Qk*, of size *k*, having the first *k*–1 elements in common.
$$R_{K+1} = P_k \bigcup Q_k = \{item_1, \dots, item_{k-1}, item_k, item_{k'}\}$$
$$P_k = \{item_1, \dots, item_{k-1}, item_k\}$$
$$Q_k = \{item_1, \dots, item_{k-1}, item_{k'}\}$$
where $item_1 < item_2 < \cdots < item_k < item_{k'}$.
2. Prune step: Verifying if all the item sets of size *k* in *Rk*+1 are frequent and generate *Ck*+1 by eliminating those that fail to pass this requirement from *Rk*+1; any subset of size *k* of *Ck*+1 can be a subset of a frequent item set of size *k* + 1 only if it is frequent.

The EM algorithm aims to find the maximum likelihood parameters θ of a statistical model where the equations involve observable variables X, unobserved latent data points or accidental and unintended missing values Z, and a vector of unknown parameters θ, as well as a likelihood function$L(\theta; X, Z) = p(X, Z|\theta)$.

The maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data, which is extremely difficult to calculate.

$$L(\theta; X) = p(X|\theta) = \sum_{Z} p(X, Z|\theta)$$

The EM algorithm tries to find the MLE of the marginal likelihood by iteratively applying the expectation and the maximization steps (Xu, 2014).

The expectation step refers to calculating the estimated value of the log likelihood function, considering the conditional distribution of Z given X under the current estimate of the parameters$\theta^{(t)}$:

$$Q\left(\theta \middle| \theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)]$$

The maximization step refers to finding the parameter that maximizes the following quantity:

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q\left(\theta \middle| \theta^{(t)}\right)$$

Page Rank refers to a search ranking algorithm using hyperlinks that are interpreted as a vote of the source page for the destination page on the Web. Votes casted by high-rank pages weigh more and help other pages become more important. The value transferred from a given page through the outbound links to the target pages upon each iteration is divided equally by the number of outbound links (Wu et al., 2008).

If the Web was a directed graph$G = (V, E)$, where *V* is the set of all pages (nodes), and *E* is the set of hyperlinks (directed edges), the total number of pages $n = |V|$. The PageRank score P of page *i* can be computed as a system of *n* linear equations with *n* unknowns, according to the formula (Wu et al., 2008):

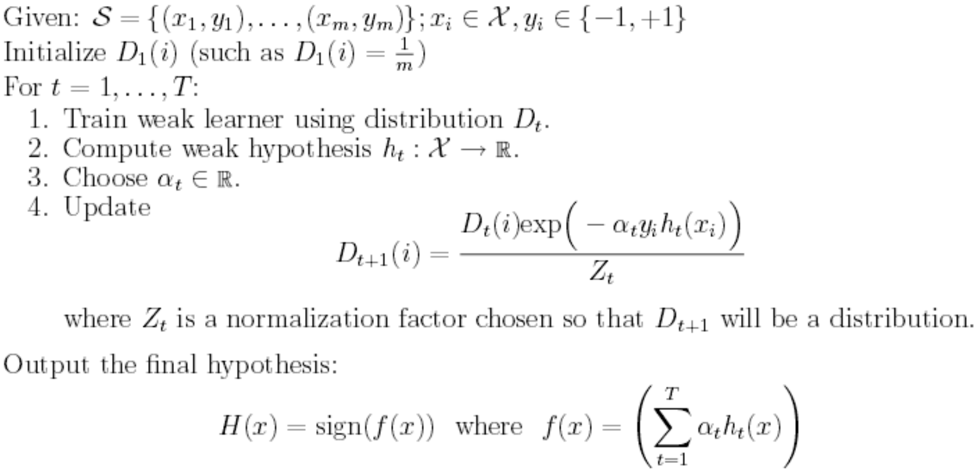
$$P(i) = (1-d) + d \sum_{(j,i)\epsilon E} \frac{P(j)}{Oj}$$

where *Oj* represents the number of outbound links of page *j*, and parameter *d* is the damping factor that can take a value between 0 and 1.

The AdaBoost algorithm, illustrated in figure 1, has a solid theoretical foundation, is very accurate in prediction and simple, and deals with methods which employ multiple learners to construct an efficient classifier, which has a significantly better generalization ability to solve a problem than a single learner (Wu et al., 2008).

AdaBoost, unlike neural networks and SVMs, selects only those features recognized to improve the predictive capacity of the model, by reducing dimensionality and possibly the execution time as irrelevant features are not computed (Diddi and Jamge, 2014).

Figure 1. AdaBoost Algorithm presented by R. Schapire and Y. Singer, 1998



```
Given: $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}; x_i \in \mathcal{X}, y_i \in \{-1, +1\}$
Initialize $D_1(i)$ (such as $D_1(i) = \frac{1}{m}$)
For $t = 1, \ldots, T$:
  1. Train weak learner using distribution $D_t$.
  2. Compute weak hypothesis $h_t : \mathcal{X} \to \mathbb{R}$.
  3. Choose $\alpha_t \in \mathbb{R}$.
  4. Update
         $D_{t+1}(i) = \frac{D_t(i)\exp\Big(-\alpha_t y_i h_t(x_i)\Big)}{Z_t}$
     where $Z_t$ is a normalization factor chosen so that $D_{t+1}$ will be a distribution.
Output the final hypothesis:
         $H(x) = \mathrm{sign}(f(x))$  where  $f(x) = \left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$
```

The distribution is updated to make wrong classifications weight more than correct classifications. The classifier weight α*t* is selected to minimize a confirmed upper bound of training error, $\prod Z_t$. Thus, $\alpha t$is the root of the equation (Fan, Stolfo and Zhang, 1999):

$$Z_t^{'} = -\sum_{i}^{m} D_t(i) u_i \, e^{-\alpha_t u_i} = 0, where\ u_i = y_i h_t(x_i)$$

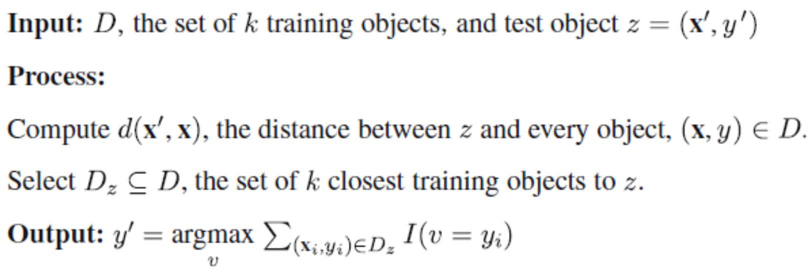
An approximation method for $0 \leq |h(x)| \leq 1$ lead to the following:

$$\alpha_t = \frac{1}{2} \ln \frac{1+r}{1-r}, where\ r = \sum_{i}^{m} D_t\,(i) u_i.$$

The k-Nearest Neighbor (kNN) algorithm represents a non-parametric pattern recognition method that is used for classification and regression (Altman, 1992).

The classification of an unlabeled test object using kNN first finds the closest k objects in the set, then computes the distance of the test object to the labeled objects and identifies the k-nearest neighbors and their class labels, and last, the class label of the test object is determined based on the predominant class labels of the kNN (Wu et al., 2008). The high-level summary of the kNN algorithm is presented in the figure 2.

Figure 2. The high-level summary of the k-nearest neighbor classification algorithm



Input: D, the set of k training objects, and test object $z = (\mathbf{x}', y')$

Process:

Compute $d(\mathbf{x}', \mathbf{x})$, the distance between z and every object, $(\mathbf{x}, y) \in D$.

Select $D_z \subseteq D$, the set of k closest training objects to z.

Output: $y' = \underset{v}{\text{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$

Source: Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H. & Steinberg, D. (2008)

The Naive Bayes (NB) classifiers refer to simple probabilistic classifiers based on applying Bayes' theorem, with strong naive assumptions of independence between each pair of features, to a set of objects belonging to a known class and vector of variables, with the purpose of constructing a rule that allows the assignment of future objects using only the vectors of variables (Collins, 2012).

The NB model uses *k*, an integer representing the total possible labels, *d*, integer stating the number of attributes, and parameters $q(y)$and $q_j(x|y)$, where $y\epsilon$ $\{1 \dots k\}$, j ϵ $\{1 \dots d\}$, x ϵ $\{-1, +1\}$.

The two parameters can be determined based on the Maximum A Posteriori (MAP) estimation. The naive Bayes classifiers versions differ generally because of the assumptions made concerning the distribution of $q_j(x|y)$ (Zhang, 2004).

The parameter $q(y)$ represents the probability of encountering the label y, where $q(y)$>0 and $\sum_{y-1}^{k} q(y) = 1$.

The parameter $q_j(x|y)$ represents the probability of attribute *j* having the value *x*, while the underlying label is *y*, where the $q_j(x|y) > 0$, and for all y, j, $\sum_{x\,\epsilon\,\{-1,+1\}} q_j(x|y) = 1$.

The probability for any $y, x_1 \dots x_d$ is defined as

$$p(y, x_1 \dots x_d) = q(y) \prod_{j=1}^{d} q_j\left(x_j \middle| y\right)$$

Once the parameters from training examples have been estimated, using a new test example $\underline{x}$=$\langle x_1, x_2, \dots, x_d \rangle$ the NB classifier has the following output:

$$arg \max_{y\in\{1\ldots k\}} p(y, x_1 \ldots x_d) = arg \max_{y\in\{1\ldots k\}} \left(q(y) \prod_{j=1}^{d} q_j \left(x_j | y\right) \right)$$

The Classification and Regression Tree (CART) refers to a binary recursive partitioning procedure that processes raw data, without binning necessary, and is capable of handling continuous and nominal attributes having both roles of targets and predictors. The decision trees are developed to a maximum size from the root node, without a stopping rule, and then pruned back split by split, to the root through cost-complexity pruning. That is the least contributing split(s) to the overall performance are pruned from the tree of training data (Wu et al., 2008).

CART rules for splitting at each child node are presented in the following way:

$$An\ instance\ goes\ left\ if\ CONDITION, or\ goes\ right\ otherwise$$

where the CONDITION has the expression $attribute\ X_i \leq C$ for continuous attributes, while for nominal attributes it appears as a member of a list of values. The quality of the split can be measured using the Least Squares (LS) or Least Absolute Deviation (LAD) for continuous targets, and using entropy, twoing criterion, ordered twoing criterion or the Gini measure of impurity for binary categorical targets.

The twoing criterion is known to perform better on multi-class targets and binary targets that are difficult to predict, and is based on directly comparing the distribution of a target attribute in two child nodes (Wu et al., 2008):

$$I(split) = \left[.25\left(q(1-q)\right)^u \sum_k |p_L(k) - p_R(k)| \right]^2$$

where k is the target class index, $p_L()$ and $p_R()$ are the probabilities of the target distributions in child nodes, u is the penalty imposed by the user on splits and q is the fraction of splits to the left.

When the binary target node t has a small number of categories, the “Gini measure of impurity” is recommended and the aim is to minimize it:

$$G(t) = 1 - p(t)^2 - \left(1 - p(t)\right)^2$$

where $p()$ in the node t represents the relative frequency of class 1. The improvement made by splitting the parent node P into children nodes L (left) and R (right) is:

$$I(P) = G(P) - qG(L) - (1-q)G(R)$$

The pruning mechanism begins with a cost-complexity measure and uses only training data:

$$R_a(T) = R(T) + a|T|$$

where $R(T)$ represents the training sample cost of the tree, $|T|$ is the total terminal nodes, and a represents a progressively increased penalty charged on every node to prune all splits (Wu et al., 2008).

3. RESEARCH METHODOLOGY

The purpose of this empirical research is to reduce the effort and duration of the selection and evaluation of projects and project portfolio optimization, increasing the success rate of budget implementation in accordance with the investment strategy of the company.

The research comprises an empirical study on investment projects, based on which the requirements elicitation process will be conducted. The model is designed for a BI system dedicated to the capital investments budgeting and control.

Research issues addressed during the empirical study are:

1. How are investments planned in the short, medium and long term?
2. What are the criteria for selecting and approving projects that will be included in the capital budget?
3. How are decisions and the evolution of investments projects reflected in the flexible budget?

In this empirical study the qualitative research was used for analyzing the established company procedures regarding investment planning and changes in the approval of projects and project budgets incurred during the year; for post-investment evaluation; identifying the events and issues that caused deviations from the original budget; analyzing deviations between static and flexible budget, as well as studying reporting procedures and the reporting format.

Research methods used are exploratory case study, case study methods based on explicit and archives. The research techniques used in this paper fall into the following categories: structured techniques (structured interview), semi-structured open techniques (conversation, individual semi-structured interview) and other techniques (grid analysis, observational techniques).

Exploratory study was used for understanding internal procedures for the selection of investment projects and capital budgeting. In this case, had been used semi-structured interviews with people from the Investment Controlling department, responsible for reporting and analysis, and other research techniques like observational techniques, by taking notes of internal aspects observed.

The case study was used explicitly for understanding the causes and effects of deviations of the flexible budget from the original approved budget. In this case were used semi-structured interviews with people in the Investment Control department responsible for controlling and analyzing the evolution of the investment projects and reporting the results; and other research techniques, like observational techniques, by taking notes regarding the internal aspects observed.

Methods based on archives are used through reviewing approval motions related to investment projects, capital expenditure motions; motions for project budget overrun, project acceleration, fund reallocation and delaying projects; reports and internal documents regarding the performance of investment projects; documentation on the indicators used; the company's directive on capital expenditures and the controlling manual. Structured research techniques were used, such as structured interviews with professionals from the Investments Controlling department that are responsible for monitoring budgets of investment projects and reporting.

The results consist of a set of requirements for a business intelligence model dedicated to investment projects, and an analysis of significant factors underlying performance differences between projected and actual achievements, to define the structure and the methods compatible in developing the solution.

4. EMPIRICAL STUDY ON INVESTMENT PROJECTS

The research was conducted in a multinational company in the oil industry, having well regulated investment activities, dedicated system for budgeting, reporting and controlling, and internally developed definition regarding the capital expenditure budget and the budgets of investment projects. The company has several business units, with different functions.

The company's investments include all procurement and production processes that result in increasing the value of tangible and intangible assets, as well as of participating interests:

- Capital expenditures (CAPEX) for tangible and intangible assets;
- CAPEX for equity participation or increasing the share capital;
- The acquisition of operations or interests;
- Foundation or acquisition of companies;
- Rental leases, classified as finance leases according to IAS;
- Advance Payments for long term rentals.

Investment Management

Investments are an essential contribution to the growth of the company, and in the case of the studied company, investments budgeting focuses on three time horizons. Thus, investment projects are planned for the short term (capital budget), medium term (mid-term planning) and long term (strategic planning). The main criterion for the selection of investment projects is the internal rate of return, but investments can be made only if there are positive free cash flows available and sufficient.

The company distinguishes between investment programs and investment budgets. The *investment program* consists of all the new capital investments requests submitted for approval as part of the Medium Term Planning (MTP). As a general rule, the submitted proposals are expected to be approved during the financial year. The program sets out a list of all projects and the way of scheduling them over the years.

Annual capital budget, on the other hand, consists in the investment program for the respective year, which was approved in that year and in previous years (total investments in tangible / intangible assets and participating interests for the financial year).

According to the policies of the company, investments plans are developed by considering four major objectives:

- Strategic objectives (ex. the rate of integration, production volumes, market share);
- Objectives related to capital structure (return on equity, gearing ratio - the company has defined a minimum rate of return on equity and a maximum gearing ratio permitted, if rates remain within the defined limits, it means that the capital structure is sound);
- The objective of profitability (ROACE, payout ratio, dividend payment rate);
- Credit score required (external rating–aim to achieve an A score, which indicates a high quality of the loan).

These elements are constantly checked to determine whether the investments project helps achieving the strategic objectives.

For all investment projects a discounted cash flow model of the future free cash flows must be calculated. The investments with excess net cash flows may be accepted in special situations, for a short term period, and only within the limits of an amount that can be brought on the medium-term in the parameters set for the capital structure. If these conditions are not met, then the investments plans of the business segments should be reviewed.

Evaluation and selection of investments projects

For the evaluation and selection of investment projects, decisions must be driven by increased revenue and business value. Project evaluation is based on the weighted average cost of capital (WACC), which represents the minimum rate of return on equity employed to calculate the net present value (NPV). The rate used in the valuation calculations is 10% for all business units, unless otherwise recommended.

For individual projects, the main criterion for selection and decision will be the internal rate of return of cash flows after tax (outflow and inflows), “IRR”, which must be greater than 13%.

Another selection criterion taken into account and presented is the payback period. The maximum allowed period is of 7 years for investments in the current activity, and 10 years for investments in the growth / development of the company. Complex interdependencies between the individual investment programs of the business units will be explicit and carefully considered, evaluated in economic terms and made transparent through decision scenarios by the Executive Council before finalizing the planning cycle.

For all investment projects there has to be established a cash flow model for the expected future free cash flows. To start the investment project, the internal rate of return of cash flows must be equal to or greater than the minimum rate of return required. The minimum rate of return represents the minimum required profit derived from capital invested.

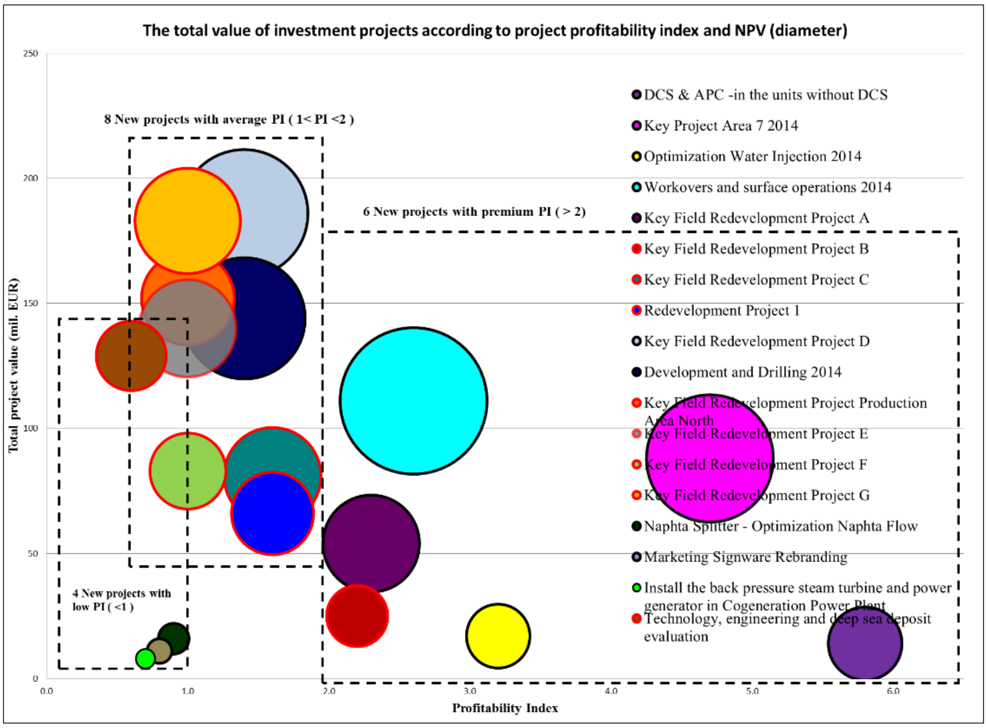
$$Internal\ Rate\ of\ Return >= \ minimum\ Rate\ of\ Return$$

Next, the net present value and the discounted payback period are determined. Strategic objectives that imply risks and the necessary related investments should take into account the established limits and the capital structure desired for the company. If the minimum rate of return on equity and gearing ratio remain within defined maximum allowable limits, it means that the company has a solid capital structure. Capital expenditures are made only if there are positive free cash flows available and sufficient.

If for a project there is insufficient evidence on the profitability of the investment, it may be included in the planning, but the project should be treated as unauthorized and shall be subject to approval before starting.

Figure 3 presents the annual activities program for 2014, which includes the main investment projects by status of their approval.

Figure 3. The total value of investment projects according to project profitability index and NPV (diameter)



Source: Projections based on the internal documents of the company

Projects that did not get the final investment approval are marked with a red outline. Total capital expenditure amounts to 1.507 mil. EUR, of which 500 mil. EUR refer to new project included in the 2014 annual investment program. The total net present value is 1.786 mil. EUR corresponding to projects for which the profitability index can be computed.

The analysis of the deviations from the consolidated budget

The update of actual results compared to budgeted values is computed periodically, as an activity related to capital expenditures reporting. The responsibility for this process lies with the Investment Controlling departments of each business unit.

For the full year 2014, the actual capital expenditure according to the company's Investments Controlling Directive, did not exceed the budgeted amounts for any of the divisions. In the case of the Refining and Marketing Division, the budget of the Marketing Subdivision was exceeded by about 2 mil. EUR, due to the acceleration of a project to be completed in 2015, but the project budget is not exceeded (Table 1 and 2).

Table 1. The quarterly budget and the actual investments realized in 2014

	Actual	Actual	Actual	Actual	Budget	Budget	Budget	Budget
CAPEX mil. EUR	Q 1	Q 2	Q 3	Q 4	Q 1	Q 2	Q 3	Q 4
	2014	2014	2014	2014	2014	2014	2014	2014
EXPLOITATION & PRODUCTION	120	155	207	323	176	203	216	215
REFINING & MARKETING	17	34	53	81	26	45	59	67
Refining	*16*	*33*	*48*	*61*	*23*	*38*	*52*	*60*
Marketing	*0*	*2*	*5*	*20*	*3*	*7*	*7*	*7*
GAS & ENERGY	3	13	9	5	14	8	18	0
CORPORATE	3	3	2	5	7	8	5	7
TOTAL COMPANY	142	205	271	413	223	265	297	289

Source: Projection based on the reports and budgets presented by the company

Table 2. The quarterly and annual differences between actual and forecasted budget

CAPEX mil. EUR	Δ	Δ	Δ	Δ	Actual	Budget	Δ	Actual %
	Q 1	Q 2	Q 3	Q 4	2014	2014	2014	2014
EXPLOITATION & PRODUCTION	-57	-48	-9	108	806	810	-4	99%
REFINING & MARKETING	-9	-11	-7	14	184	198	-14	93%
Refining	*-7*	*-5*	*-4*	*1*	*157*	*173*	*-16*	91%
Marketing	*-2*	*-6*	*-3*	*13*	*27*	*25*	*2*	108%
GAS & ENERGY	-11	5	-9	4	30	40	-11	74%
CORPORATE	-4	-6	-3	-2	12	26	-14	46%
TOTAL COMPANY	-81	-60	-27	124	1,032	1,075	-43	96%

Source: Projection based on the reports and budgets presented by the company

The company ended the financial year 2014 with a budget execution of 96%, according to the Investments Controlling Directive, and realization of 98% according to international accounting standards in EUR. Considering the acceleration of the modernization and rebranding project, the marketing budget has an achievement of 108%.

The investments in E&P activities accounted for 73% of the total amount invested in 2014 and focused on drilling development wells, intervention works, modernization and depth / surface operations, field redevelopment projects, production equipment, waste infrastructure and equipment necessary to the Exploitation and Production Services sub-division (Table 3).

The Gas&Energy and Corporate Divisions have significant degrees of under achievement, with a negative effect on the degree of achievement of the general budget. Investments in G&E decreased compared to the previous year because the evolution of the final stages of the Power Plant and Wind Power Park projects determined a delay of about 9 months. The plans for the Wind Power Park project were recently modified to include 3 additional turbines of 9MW. The budget will not be exceeded for any of the projects (Table 3).

Table 3. The flexible budget comprising the major investment projects

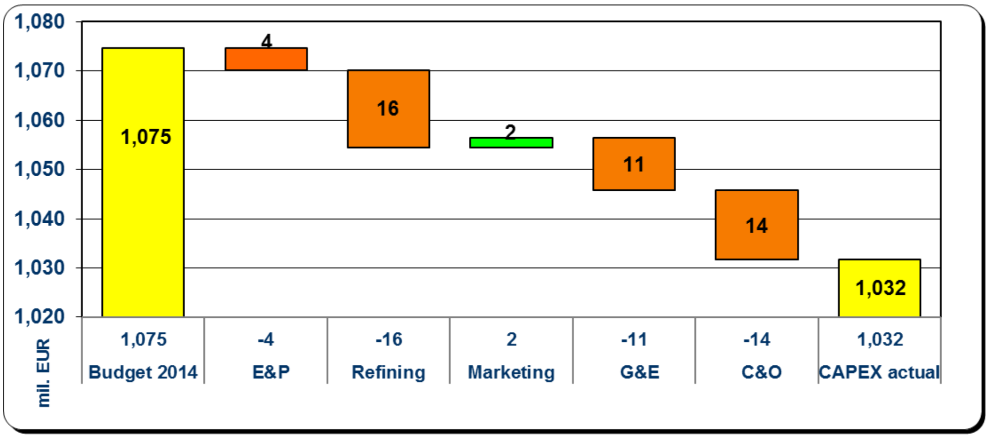
In mil. EUR		Total value	Budget adjustments	Total budgeted value	Actual to date	Actual 2014	Budget 2014	Variance 2014
Continued Projects								
E&P	Development Wells and Seismic	238	-	238	214	16	24	-8
E&P	Field Redevelopments and Optimization Water Injection	1	-	1	-	1	1	0
E&P	Surface Facility Modernization	83	19	102	57	45	26	19
E&P	Waste Infrastructure	58	-	58	26	23	10	13
E&P	Integrity and energy efficiency	26	-	26	8	5	5	0
E&P	Other investments	89	-	89	81	8	8	0
E&P	Investments in Business Units	75	-	75	-	41	41	0
G&E	Wind Power Park	100	-	100	89	7	11	-4
G&E	POWER PLANT	535	-	535	508	21	27	-6
G&E	Other small projects	-	-	-	1	2	2	0
R&M	Refinery modernization	-	-	-	72	19	68	-49
R&M	Storage projects	-	-	-	26	7	30	-23
R&M	Systematization, optimization and upgrading loading / unloading stations	24	-	24	4	20	7	13
R&M	Other continued projects - R&M	67	136	203	125	123	78	45
C&O	Other continued projects - C&O	-	-	-	7	5	5	0
TOTAL CONTINUED PROJECTS		1,296	155	1,451	1,218	342	343	-1
New Projects								
E&P	Production equipment	37	-	37	-	37	37	0
E&P	Work over / well modernization	111	-	111	-	111	111	0
E&P	Field Redevelopments and Optimization Water Injection	740	-8	732	17	161	179	-18
E&P	Capitalized exploration and appraisal	144	-6	138	-	134	144	-10
E&P	Integrity and energy efficiency	36	-	36	-	36	36	0
E&P	Surface facility modernization	51	-	51	-	51	51	0
E&P	Technology, engineering and deep sea deposit evaluation	43	53	96	-	95	43	53
E&P	Other investments	82	-	82	-	23	76	-53
E&P	Intervention work capitalized and modernization of depth wells	87	-	87	-	16	16	0
E&P	New projects starting in 2015	2,656	-	2,656	-	3	3	0

	In mil. EUR	Total value	Budget adjustments	Total budgeted value	Actual to date	Actual 2014	Budget 2014	Variance 2014
New Projects (continued)								
R&M	New projects R&M	139	-	139	2	15	15	0
C&A	New projects C&O	110	-	110	8	8	22	-14
TOTAL NEW PROJECTS		4,734	38	4,772	47	1,767	1,880	-113
TOTAL CAPITAL EXPENDITURE 2014						1,783	1,896	-113

Business investments also declined during the year, as a result of completing the headquarter construction project, and mainly relate to IT projects (Figure 4).

Values above or below the budgeted, the deviations between the actual, the forecasted and the initial budget may be caused by different costs than anticipated and changes in the schedule of the projects.

Figure 4. The evolution of investments compared to the Capital Expenditure Budget



Source: Projection based on the reports and budgets presented by the company

Cash flows from investment activities mainly include payments for investments in tangible and intangible assets. According to the Investments Controlling Directive, in 2014 the company invested 1.032 mil. EUR, while the recorded external cash flow was of 1.079 mil. EUR. This difference requires a reconciliation between capital expenditures and cash outflows corresponding to investments in intangible and tangible assets.

Actual cash outflows are calculated by the Treasury Department based on actual payments realized from the account of each project, and estimated cash outflows are calculated within each division with regard to technical and operational reasoning. The estimates are reported to the Treasury Department, and it manages the funds for the next period.

Project evaluation should be conducted based on the value of the cash flow in EUR. Values should be calculated based on EUR / USD value, according to the rate provided for in the assumptions underlying the Medium Term Planning (MTP) of the company.

These assessment criteria do not apply to infrastructure investments, investments in the mandatory Health, Safety and Occupational Safety, Environment, Quality Management or mandatory investment requirements imposed by other legal requirements.

The maximum time period that can pass between project approval and the beginning of the project activities (the first record of CAPEX) is 18 months. If the project does not start during this period, the project approval shall automatically expire without exception. If the project is further supported, it will be resubmitted for approval or information to the relevant management board, so the entire approval process and information is restarted.

If when a project is submitted, there is insufficient evidence on its profitability, the investment value may be included in planning, but the project should be treated as unauthorized and shall be subject to approval before it is started.

The main factors that determine deviations from the budget

For evaluation of projects and budget planning and review, several factors are considered:

- Brent oil price;
- Refining margin;
- The exchange rate EUR / USD and other exchange rates;
- Sales prices;
- Prices of raw materials and synthetic compounds;
- Macroeconomic variables (GDP developments, consumer price index);
- Energy consumption;
- Level of taxes and excise duties;
- Legal framework - limiting or banning the export of certain products lower prices and reduce the profitability of projects;
- Prices of equipment and facilities, especially for new technologies, etc.

The exceeds of the budgets of investment projects are determined by comparing, on the one hand, the latest forecast for the total value of the project during its entire length, in USD or EUR, and, on the other hand, total value approved (total discounted value of the project).

A planned reallocation between investment projects is interpreted as a deliberate budget exceeding for a project and, in parallel, a reduction for another project. Reassignment is only accepted if the total budget of the respective business unit is maintained and the updated documentation of these projects is accepted by management.

Reallocations between business units (“transfer budget”) generally are not expected during the year, for reasons of accuracy and governance, but can be made during the preparation of annual investment program and require separate approval.

5. CONCEPTUAL BI MODEL AND APPLICABLE METHODS

Implementing a Business Intelligence system must take into account the following features and specifications:

- The peculiarities of the activity require a hybrid solution for budgeting: zero-based budgeting method for new projects, and reviewing the current results and the predicted behavior for the next year in the case of continued projects.
- In the annual budgeting process, for short-term investments is made a distinction between fixed costs and variable costs, while all long-term investment expenses are considered variable. This peculiarity of considering all short-term expenses as variable costs is specific to activity-based budgeting, used for investments because it provides more detailed information and allows a better control of costs.
- During the year, projects can be accelerated, deferred or canceled. Certain projects that have low priority or no longer meet the selection criteria can be canceled.
- The budget of certain projects may be reallocated for other high priority or better performing projects, both within and between divisions.
- If during the year are identified some opportunities that were not foreseen when the budget was approved, or the budget increase of some large projects is approved, which lead to exceeding the investment budget of the division, the difference is presented and justified separately to shareholders.

Deviations between the total investment budget approved and the amount of actual investments are generally due to purchases that are made ad hoc, as the entity identifies on the market companies that can help it achieve its objectives. Purchases are not part of the scope of the company, so are most often the cause of these differences. BI system must keep a log of the budget, budget deviations to correlate transactions that caused differences, forecast and integrate these developments into reports for controlling and management teams.

Every year, in addition to establishing the annual investment program, the medium term plan and strategic budget for the next interval of 3 and 5 years are revised. The main drivers of changes in the planned capital expenses for the projects are the exchange rates and the oil and gas prices. Medium-term plan and strategic budget are for guidance, they are not approved by the Executive Board, but are the basis for the annual investment program of the company.

In this context, it is necessary that the BI system manages the planning history on the three time horizons, contains a group of tables of measures for key factors determining changes in the budgets of investment projects, provides real-time budget estimates based on these measures, allows hierarchical navigation along multiple dimensions, such as time horizons, the investment program, the project budgets, etc.

A periodic investments report containing the statuses of the projects is developed quarterly and it is required for each major project. As the values of the projects included in the capital budget for the year cannot be broken down by months or quarters in a clear way and without undue expenses, no actual budgetary comparisons exist during the year. The assessment of the investments is based on a forecast for the full year, which is developed from the actual costs incurred up to the date of the report.

These issues require the implementation of a BI system updated in real time, integrating computer systems used for accounting, taxation, legal activities, financial management, production, performance management, purchasing and controlling. Based on the correlations between the stage of the project planning and the activities performed, the BI solution should facilitate the realization of status reports for the projects, interpretations of the results and identification of risk factors.

The analysis shows that status reports are susceptible to errors due to delays in the key moments that lead to inadequate decisions. For large projects such delays cause postponement of next stages and financial losses.

Corresponding to the type of requirements and specifications collected through the study, the conceptual model developed has a multidimensional hierarchy structure, a constellation architectural model around the planned, forecasted and actual facts regarding investment projects, uses the Top-Down implementation perspective and is user-requirement and goal oriented (Figure 5).

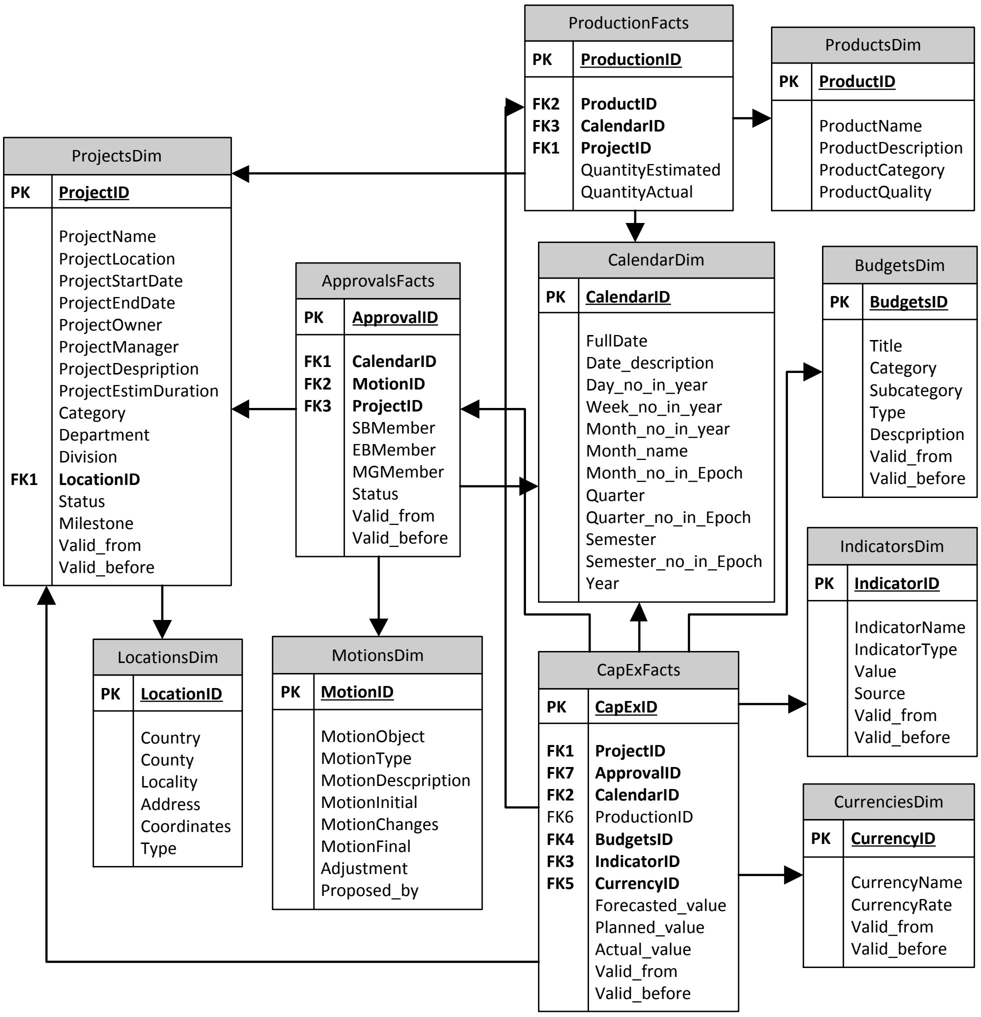
The company can combine different data mining tools in order to optimize its decisional process. Thus, the C4.5 algorithm can be used for improving decision-making based on investment scenarios or project estimations, as it can easily handle missing values.

Considering the budget for each division and the continued projects, the company can analyze the production and capital expenses using the k-means algorithm, in order to optimize the investment strategy and determine the budget for growth or business continuation projects.

SVMs can be used to solve various problems like text and hypertext categorization and information retrieval, for example in the case of project prospects investment opportunities, production support documentation that previously could not be used directly for analysis, classification of production fields using geological expertise images, classifying exploitation products and by-products, and many other business areas where raking is used.

Apriori algorithm can be used to determine what reports and information are most accessed or requested and by whom, in order to optimize the reporting system, and modify or eliminate useless analysis.

Figure 5. DW logical model of the BI system for investment projects



Source: Projection based on the requirements elicitated throughout the study

EM algorithm offers the ability to work with missing data and unidentified variable and is a useful tool for determining prices, and managing the risk of an investments portfolio (McNeil et al., 2005).

PageRank uses general mathematics and is applicable to any graph, in bibliometrics, social and information network analysis, and for link prediction and recommendation (Gleich, 2014). The company can use the algorithm for prioritizing and optimizing its reporting activity and archiving, by recommending related analysis and documents for faster retrieval and easier access.

The company can use the AdaBoost algorithm for fraud detection, minimizing exponential loss, maximizing the margin and estimating the time span of the project's activity stages.

kNN can be used to verify approval signatures on scanned documents, text mining through motions and project prospects, evaluating field resources for estimating daily production based on engineering documentation and reports, computing correlations between variables, forecasting the evolution of the scheduled activities for each project (Imandoust et al., 2013), planning investment strategies, uncovering oil prices trend, predicting currency exchange rate, classifying projects and establishing statuses.

Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents. This algorithm can be used for classifying logs and organizing archived documents.

The company can use the CART for classifying capital expenses in the appropriate category for building investments programs, budgets and reports, explore the outcomes of different project planning solutions based on specific coefficients and indicators, classifying risk levels and improving strategic investments results. The data has to be continuously updated according to market evolution. The results of the investment budgets may be negatively influenced by the instability of the trees.

6. CONCLUSIONS

Management should evaluate alternative strategies and long-term investments, and then to decide how to implement investment projects so as to achieve business objectives. This process of analysis and capital budgeting is one of the most difficult, risky and important activities that management is responsible of. This task involves computations, projections and estimates, and the decisions of the management regarding the capital budget have a long-term impact on the company.

In order to improve the budgeting and controlling processes, a synthetic set of specifications was collected through the empirical study and a BI system model was proposed. The first stages in building a BI solution are the viability study and the elicitation of requirements for the elaboration of the DWH model. These stages are essential for the success of the BI project, they are the most time-consuming processes, require very good communication with the stakeholders and a very good understanding of their needs in order to develop a high-quality model of the BI system.

Some of the benefits associated to the proposed BI system regarding the efficiency and accuracy of budgeting and controlling refer to:

- Clarifying the situation of project managers and project owners in order to avoid double budgeting or failure of the project budget;
- Automatic alignment within the departments of divisions regarding capital expenses;
- Transparency of costs, the activities of planning, forecasting, budgeting, and deviation analysis require much less effort, are more effective and available on real-time;

- Standardization of reporting at division level, respecting the same reporting requirements agreed in the company, so controlling teams can easily support and guide the reporting tasks;
- When planning, allowed categories of expenditure are respected: where only administrative costs are allowed on cost center of the department, only those costs can be allocated;
- Automatic review of forecasts and accurate estimates by project and department, relevant for the calculation of taxes and forecasts at consolidated level;
- Documenting assumptions for each individual amount estimated;
- Timeliness of delivery of reports scheduled at fixed dates.

Further research directions on the subject regard the impact on the financial performance and the decisional process. The improvement of the operational processes is expected to be reflected in the success and flexibility of decision-making, the adaptability to the changing market conditions and the capacity of mitigating risk factors. Another aspect for investigation is the ability of the professionals to grasp and interpret the new information available and employ it in gaining competitive advantage and if a BI system can trigger the proper behavior.

Acknowledgement:
This work was co-financed from the European Social Fund through Sectorial Operational Programme Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/142115 „Performance and excellence in doctoral and postdoctoral research in Romanian economics science domain”.

REFERENCES

[1] **Altman, N. S. (1992), *An Introduction to Kernel and Nearest-neighbor Nonparametric Regression*.** The American Statistician, Vol. 46 (3), 175–185;
[2] **Baan, P. & Homburg, R. (2013), *Information Productivity: An Introduction to Enterprise Information Management*;** Enterprise Information Management, *Springer, New York*. 1-42;
[3]**Bishop, C. M. (1995), *Neural Networks for Pattern Recognition***; Oxford, England: *Oxford University Press*;
[4]**Collins, M. (2012), *The Naive Bayes Model, Maximum-Likelihood Estimation, and the EM Algorithm*;** Course notes, Columbia University, accessed online at http://www.cs.columbia.edu/~mcollins/em.pdf;
[5]**Davenport, T. and Harris, J.(2007), *Competing on Analytics: The New Science of Winning*;** *Harvard Business School Press*;

[6]**De Veaux, R. (2000),** ***Data Mining: What's New, What's Not***. Presentation at a Data Mining Workshop, Long Beach, California;
[7]**Diddi, V. K. and Jamge, S. B. (2014),** ***Head Pose and Eye State Monitoring (HEM) for Driver Drowsiness Detection: Overview***. International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 9;
[8]**Fan, W., Stolfo, S. J. & Zhang, J. (1999),** ***The Application of AdaBoost for Distributed, Scalable and On-line Learning***. In Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 362-366;
[9]**Gleich, D. F. (2014),** ***Page Rank beyond the Web****;* arXiv preprint arXiv:1407.5107;
[10]**Imandoust, S. B. & Bolandraftar, M. (2013),** ***Application of K-Nearest Neighbor (KNN) Approach for Predicting "Economic Events: Theoretical Background".*** Int. Journal of Engineering Research and Applications, 3(5), 605-610;
[11]**Kimball, R. & Ross, M. (2013),** ***The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling***. *John Wiley & Sons*;
[12]**Korting, T. S. (2006),** ***C4. 5 Algorithm and Multivariate Decision Trees*.** Image Processing Division; National Institute for Space Research–INPE Sao Jose dos Campos–SP, Brazil;
[13]**Kudyba, S.(2014),** ***Big Data, Mining and Analytics: Components of Strategic Decision Making****; CRC Press*;
[14]**Lahrmann, G., Marx, F., Winter, R. & Wortmann, F. (2010),** ***Business Intelligence Maturity Models: An Overview;*** The VII conference of the Italian chapter of AIS (itAIS 2010). Italian chapter of AIS, Naples;
[15]**Loshin, D. (2012),** ***Business Intelligence: The Savvy Manager's Guide*.** Newnes;
[16]**McNeil, A., Frey, R. and Embrechts P. (2005),** ***Quantitative Risk Management: Concepts and Tools.*** *Princeton University Press;*
[17]**Menéndez, D.A. and da Silva P.C. (2014),** ***A Requirement Elicitation Process for BI Projects****;* Lecture Notes on Software Engineering; Vol. 4, No. 1;
[18]**Rasmussen, N.,Goldy, P. and Solli P.(2002),** ***Financial Business Intelligence. Trends, Technology, Software Selection, and Implementation***; *Wiley*;
[19]**Roebuck, K. (2012),** ***Big Data: High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors***; *Emereo Publishing,* 154-209;
[20]**Schapire R. and Singer Y. (1998),** ***Improved Boosting Algorithms Using Confidence-rated Predictions***. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory;
[21]**Vercellis, C. (2009),** ***Business Intelligence: Data Mining and Optimization for Decision Making;*** Take Edition, *Publisher John Wiley& Sons, West Sussex;*

[22]**Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H. & Steinberg, D. (2008),** ***Top 10 Algorithms in Data Mining. Knowledge and Information Systems***; 14(1), 1-37;
[23]**Xu, H. (2014),** ***Probabilistic Atlas Statistical Estimation with Multimodal Data Sets and its Application to Atlas Based Segmentation***. Statistics. Ecole Polytechnique X, accessed online at https://tel.archives-ouvertes.fr/pastel-00969176/document;
[24]**Zhang, H. (2004),** ***The Optimality of Naive Bayes****;* Proceedings of FLAIRS.