International Journal of Computer & Information Technologies (IJOCIT)
www.ijocit.ir
ISSN = 2345-3877

Factors affecting the Increase and Decrease student achievement in Primary School with Business Intelligence Approach

Lida Shams[1], Hassan Rashidi [*2]
Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran*[1]
Department of Mathematics and Computer Science, Allameh Tabataba'i University, Tehran, Iran[2]
shams.lida168@gmail.com

Abstract: Although the educational level of the Portuguese population has improved in the last decades, the statistics keep Portugal at Europe's tail end due to its high student failure rates. In particular, lack of success in the core classes of Mathematics and the Portuguese language is extremely serious. On the other hand, the fields of Business Intelligence (BI)/Data Mining (DM), which aim at extracting high-level knowledge from raw data, offer interesting automated tools that can aid the education domain. The present work intends to approach student achievement in secondary education using BI/DM techniques. Recent real-world data (e.g. student grades, demographic, social and school related features) was collected by using school reports and questionnaires. The two core classes (i.e. Mathematics and Portuguese) were modeled under binary/five-level classification and regression tasks. Also, four DM models (i.e. Decision Trees, Random Forest, Neural Networks and Support Vector Machines) and three input selections (e.g. with and without previous grades) were tested. The results show that a good predictive accuracy can be achieved, provided that the first and/or second school period grades are available. Although student achievement is highly influenced by past evaluations, an explanatory analysis has shown that there are also other relevant features (e.g. number of absences, parent's job and education, alcohol consumption). As a direct outcome of this research, more efficient student prediction tools can be developed, improving the quality of education and enhancing school resource management.

Keywords: Data Mining, Business Intelligence, Neural Network, Classification.

1. Introduction

Education is a key factor for achieving a long-term economic progress. During the last decades, the Portuguese educational level has improved.

However, the statistics keep the Portugal at Europe's tail end due to its high student failure and dropping out rates. For example, in 2006 the early school leaving rate in Portugal was 40% for 18 to 24 year olds, while the European Union average value was just 15% [3]. In particular, failure in the core classes of Mathematics and Portuguese (the native language) is extremely serious, since they provide fundamental knowledge for the success in the remaining school subjects (e.g. physics or history). On the other hand, the interest in Business Intelligence (BI)/Data Mining (DM), arose due to the advances of Information Technology, leading to an exponential growth of business and organizational databases. All this data holds valuable information, such as trends and patterns, which can be used to improve decision making and optimize success. Yet, human experts are limited and may overlook important details. Hence, the alternative is to use automated tools to analyze the raw data and extract interesting high-level information for the decision-maker. The education arena offers a fertile ground for BI applications, since there are multiple sources of data (e.g. traditional databases, online web pages) and diverse interest groups (e.g. students, teachers, administrators or alumni). This paper will focus in the last two questions. Modeling student performance is an important tool for both educators and students, since it can help a better understanding of this phenomenon and ultimately improve it. For instance, school professionals could perform corrective measures for weak students (e.g. remedial classes). In effect, several studies have addressed similar topics. Applied a DM approach based in Association Rules in order to select weak tertiary school students of Singapore for remedial classes. The input variables included demographic attributes (e.g. sex, region) and school performance over the past years and the proposed solution outperformed the traditional allocation procedure. In this work, we will analyze recent real-world data from two Portuguese secondary schools. Two

different sources were used: mark reports and questionnaires. Since the former contained scarce information (i.e. only the grades and number of absences were available), it was complemented with the latter, which allowed the collection of several demographic, social and school related attributes (e.g. student's age, alcohol consumption, mother's education). The aim is to predict student achievement and if possible to identify the key variables that affect educational success/failure. The two core classes (i.e. Mathematics and Portuguese) will be modeled under three DM goals:

- Binary classification (pass/fail);
- classification with five levels (from I very good or excellent to V - insufficient); and
- Regression, with a numeric output that ranges between zero (0%) and twenty (100%).

For each of these approaches, three input setups (e.g. with and without the school period grades) and four DM algorithms (e.g. Decision Trees, Random Forest) will be tested. Moreover, an explanatory analysis will be performed over the best models, in order to identify the most relevant features [5].

2. Business Intelligence

define Business intelligence as the process of taking large amounts of data, analyzing that data, and presenting a high-level set of reports that condense the essence of that data into the basis of business actions, enabling management to make fundamental daily business decisions. View BI as way and method of improving business performance by providing powerful assists for executive decision maker to enable them to have actionable information at hand. BI tools are seen as technology that enables the efficiency of business operation by providing an increased value to the enterprise information and hence the way this information is utilized. [6] Define BI as "The process of collection, treatment and diffusion of information that has an objective, the reduction of uncertainty in the making of all strategic decisions." Experts describe Business intelligence as a "business management term used to describe applications

and technologies which are used to gather, provide access to analyze data and information about an enterprise, in order to help them make better informed business decisions.' describes the basic characteristic for BI tool is that it is ability to collect data from heterogeneous source, to possess advance analytical methods, and the ability to support multi user's demands. Categorized BI technology based on the method of information delivery; reporting, statistical analysis, ad-hoc analysis and predicative analysis. The concept of Business Intelligence (BI) is brought up by Gartner Group since 1996. It is defined as the application of a set of methodologies and technologies, such as J2EE, DOTNET, Web Services, XML, data warehouse, OLAP, Data Mining, representation technologies, etc., to improve enterprise operation effectiveness, support management/decision to achieve competitive advantages. Business Intelligence by today is never a new technology instead of an integrated solution for companies, within which the business requirement is definitely the key factor that drives technology innovation. How to identify and creatively address key business issues is therefore always the major challenge of a BI application to achieve real business impact.

[9] Defined BI that includes effective data warehouse and also a reactive component capable of monitoring the time-critical operational processes to allow tactical and operational decision-makers to tune their actions according to the company strategy. Define BI as the result of in-depth analysis of detailed business data, including database and application technologies, as well as analysis practices. Widen the definition of BI as technically much broader tools that includes potentially encompassing knowledge management, enterprise resource planning, decision support systems and data mining.

3. Data Warehouse and data marts

The data warehouse is the significant component of business intelligence. It is subject oriented, integrated. The data warehouse supports the

physical propagation of data by handling the numerous enterprise records for integration, cleansing, aggregation and query tasks. It can also contain the operational data which can be defined as an updateable set of integrated data used for enterprise wide tactical decision-making of a particular subject area. It contains live data, not snapshots, and retains minimal history. Data sources can be operational databases, historical data, external data for example, from market research companies or from the Internet), or information from the already existing data warehouse environment. The data sources can be relational databases or any other data structure that supports the line of business applications. They also can reside on many different platforms and can contain structured information, such as tables or spreadsheets, or unstructured information, such as plaintext files or pictures and other multimedia information. A data mart as described is a collection of subject areas organized for decision support based on the needs of a given department [10]. Finance has their data mart, marketing has theirs, and sales have theirs and so on. And the data mart for marketing only faintly resembles anyone else's data mart. Perhaps most importantly, the individual departments own the hardware, software, data and programs that constitute the data mart. Each department has its own interpretation of what a data mart should look like and each department's data mart is peculiar to and specific to its own needs. Similar to data warehouses, data marts contain operational data that helps business experts to strategize based on analyses of past trends and experiences. The key difference is that the creation of a data mart is predicated on a specific, predefined need for a certain grouping and configuration of select data. There can be multiple data marts inside an enterprise. A data mart can support a particular business function, business process or business unit. A data mart as described is a collection of subject areas organized for decision support based on the needs of a given department. Finance has their data mart, marketing has theirs,

and sales have theirs and so on. And the data mart for marketing only faintly resembles anyone else's data mart [11].

BI tools are widely accepted as a new middleware between transactional applications and decision support applications, thereby decoupling systems tailored to an efficient handling of business transactions from systems tailored to an efficient support of business decisions. The capabilities of BI include decision support, online analytical processing, statistical analysis, forecasting, and data mining. The following are the major components that constitute BI.

4. Data Sources

Data sources can be operational databases, historical data, external data for example, from market research companies or from the Internet), or information from the already existing data warehouse environment. The data sources can be relational databases or any other data structure that supports the line of business applications. They also can reside on many different platforms and can contain structured information, such as tables or spreadsheets, or unstructured information, such as plaintext files or pictures and other multimedia information.

5. Data Warehouse

Data warehouse is one of the most important components in BI architecture. [12] Defines data warehouse as "a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process" (p. 29). The characteristics of a data warehouse are described as follows:

- **Subject-Oriented**

Data from various sources are organized into groups based on common subject areas that an organization would like to focus on, such as customers, sales, and products.

- **Integrated**

Data warehouse gathers data from various sources. All of these data must be consistent in terms of naming conventions, formats, and other related characteristics.

- **Time-Variant**

Each data stored in the data warehouse has time dimension to keep track of the changes or trends on the data. In other words, data warehouse will store historical changes on each piece of data.

- **Non-Volatile**

New data can be added into data warehouse regularly. But, all the data stored in data warehouse are read-only. This means users are not allowed to update, over-write or delete the stored data.

In summary, data warehouse is a central storage that collects and stores data from internal and external sources for strategic decision making, queries, and analysis (Bara et al., 2009; Imhoff et al., 2003). Data warehouse stores aggregated or summarized data. In addition, it also stores large amount of historical data for the purpose of long term analysis (Li et al., 2007). Data are stored in data warehouse longer (5 to 10 years) than in ODS (60 to 90 days) (Chan, 2005). Data in a data warehouse is updated regularly, for instance weekly or sometimes daily (Al- Noukari & Al-Hussan, 2008). As a result, it does not contain the latest data as in operational systems and ODS. Aside from that, data warehouses are designed to support OLAP (Online Analytical Processing) applications by storing and maintaining data in multi-dimensional structures for query, reporting, and analysis (Sen & Sinha, 2005).

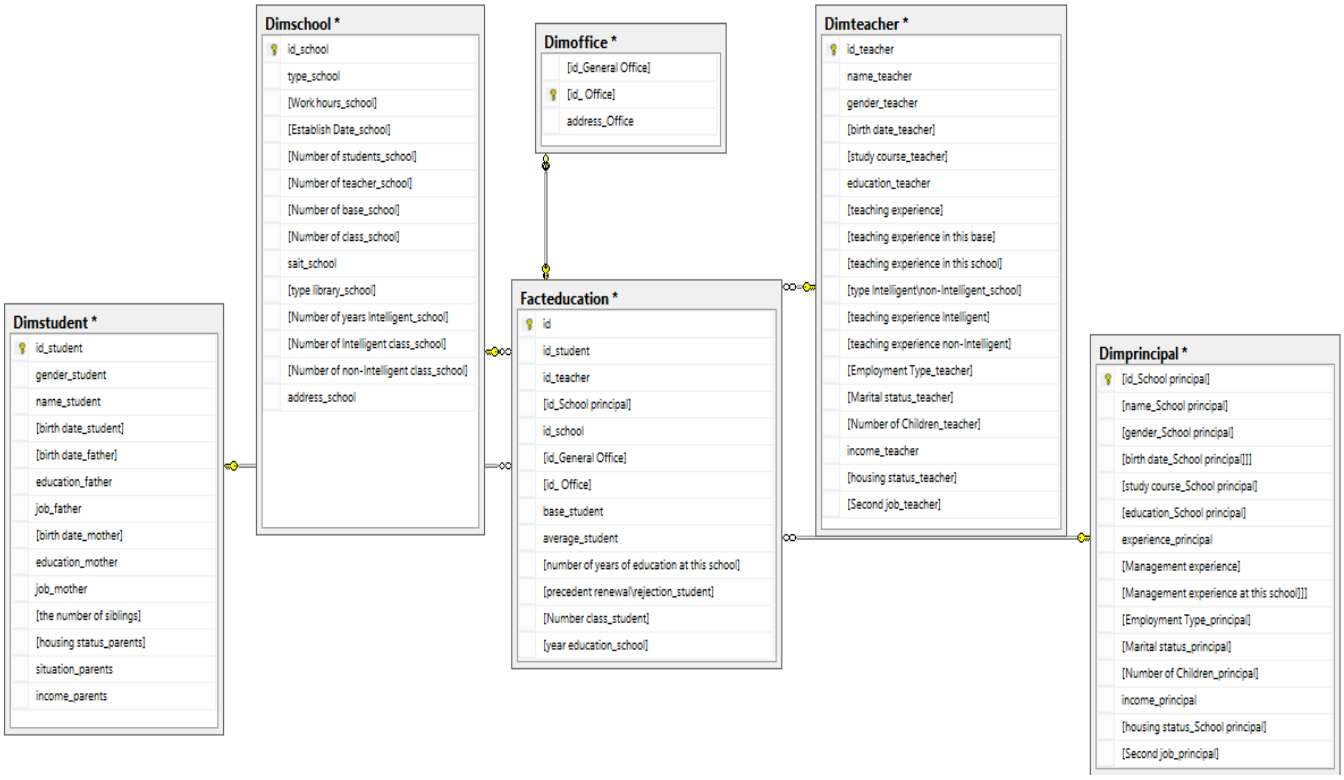

Figure 1: Scheme of Data Warehouse [11].

- **Data Mart**

While the data in a data warehouse is mainly used to support various needs across the whole organization, it is not equipped to support the needs and requirements of specific departments. Consequently, it is necessary to have data marts to support them. A data mart is a subset of the data warehouse that is used to support analytical needs of a particular business function or department. Like data Warehouses, it contains historical data that can help users to access and analyze different data trends [11]. However, it can only keep data for 60 to 90 days. Therefore, the amount of data stored in a data mart is much lesser than the data stored in a data warehouse. There can be many data marts inside an organization. Data warehouses and data marts are built based on multi-dimensional data model which consists of fact and dimension tables. Fact table contains quantitative data about business entities such as sales amount, quantity, and price.

Dimension table contains data (such as product, customer, data, and location) that describes facts [13].

- **Approaches in Data Warehousing**:

The main key to successful BI system is consolidating data from the many different enterprise operational systems into an enterprise data warehouse. Very few organizations have a full-fledged enterprise data warehouse. This is due to the vast scope of effort towards consolidating the entire enterprise data. Emphasizes that in view of emerging highly dynamic business environment, only the most competitive enterprises will achieve sustained market success. The organizations will distinguish themselves by the capability to leverage information about their market place, customers, and operations to capitalize on the business opportunities.

Several surveys including Gartner, Forrester and International Data Centre report that most of the firms throughout the globe are interested in investing in BI. It is to be noted that despite major investments in enterprise resource planning (ERP) and customer relationship management (CRM) over the last decade businesses are struggling to achieve competitive advantage. It is due to the information captured by these systems. Any corporate would look forward for one goal called 'right access to information quickly'. Hence, the firms need to support the analysis and application of information in order to make operational decisions. Say for marking seasonal merchandise or providing certain recommendations to customers, firms need right access to information quickly. Implementing smarter business processes is where business intelligence influences and influences the bottom line and returns value to any firm.

- **Association Rules and Regression**:

Association rules look for relationships between items. The most common example of this is market basket analysis. Market basket analysis studies retail purchases to determine which items tend to appear together in individual transactions.

Online retailers use market basket analysis on their websites to suggest additional items to purchase before a customer completes their order. The key to this type of analysis is the ability to find associations amongst the items in each analysis. This can include associations such as which items appear together the most frequently, and which items tend to increase the likelihood that other items will also appear in the same transaction.

6. Business Intelligence Architecture

This section proposes a framework of a five-layered BI architecture (see Figure 2), taking into consideration the value and quality of data as well as information flow in the system. The five layers are data source, ETL (Extract-Transform-Load), data warehouse, end user, and metadata layers. The rest of this section describes each of the layers. Nowadays, many application domains Require the use of structured data as well as unstructured and semi-structured data to make effective and timely decision [15]. All these data can be acquired from two types of sources: internal and external. Internal data source refers to data that is captured and maintained by operational systems inside an organization such as Customer Relationship Management and Enterprise Resource Planning systems. Internal data sources include the data related to business operations (i.e., customers, products, and sales data). These operational systems are also known as online transaction processing systems because they process large amount of transactions in real time and update data whenever it is needed. Operational systems contain only current data that is used to support daily business operations of an organization. Generally, operational systems are process-oriented as they focus mainly on specific business operations such as sales, accounting, and purchasing. External data source refers to those that originate outside an organization. This type of data can be collected from external sources such as business partners, syndicate data suppliers, the Internet,

governments, and market research organizations. These data are often related to competitors, market, environment (e.g., customer demographic and economic), and technology. It is important for organizations to clearly identify their data sources. Knowing where the required data can be obtained is useful in addressing specific business questions and requirements, thereby resulting in significant time savings and greater speed of information delivery. Furthermore, the knowledge can also be used to facilitate data replication, data cleansing, and data extraction this is because even though there are many existing data sources, some of them might be inaccessible, unreliable or irrelevant to current business needs. With correct identification of data sources, problems such as inconsistent information, difficulty in finding root causes, and issues of data isolation can be avoided.

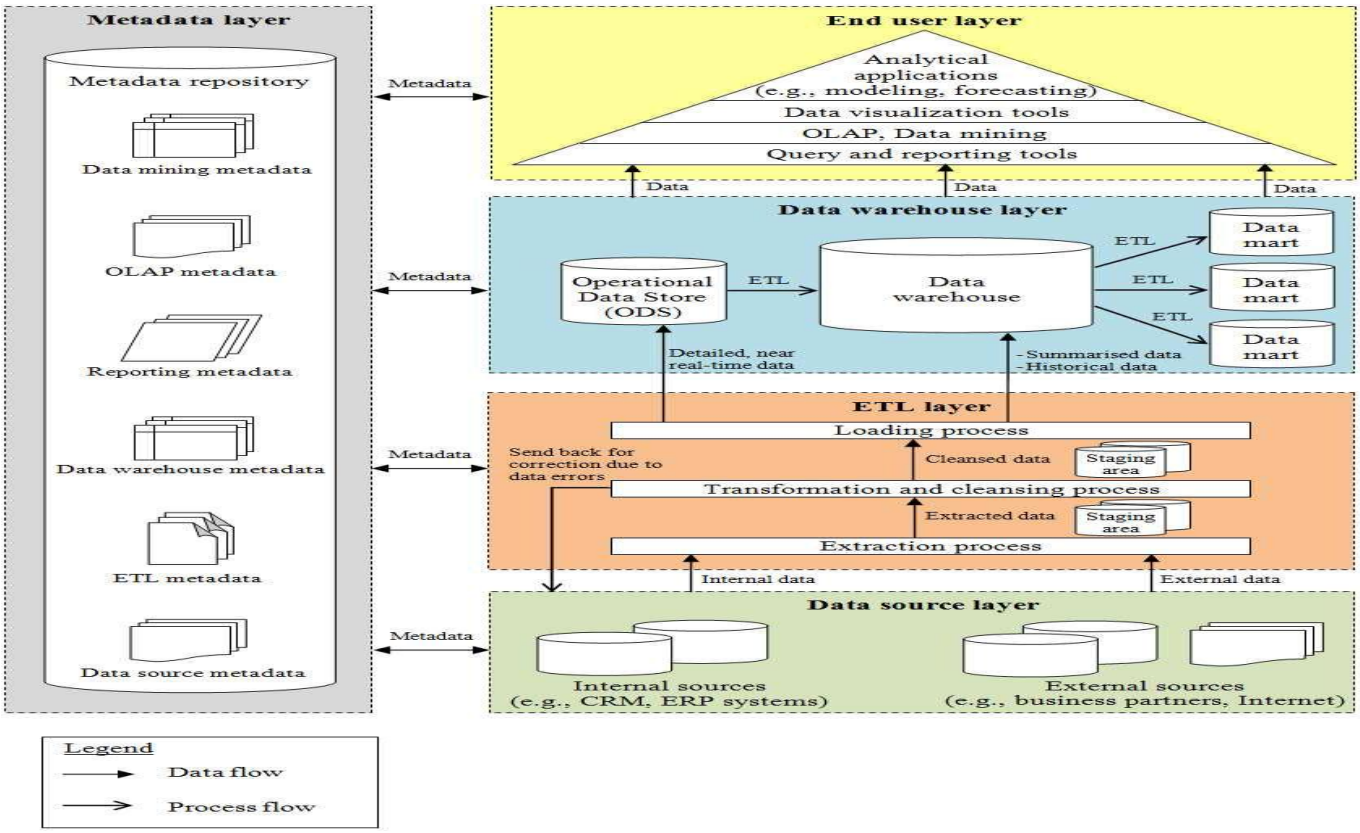

Figure2: BI Architecture [15]

7. Conclusion

Education is a crucial element in our society. Business Intelligence (BI)/Data Mining (DM) techniques, which allow a high level extraction of knowledge from raw data, offer interesting possibilities for the education domain. In particular, several studies have used BI/DM methods to improve the quality of education and enhance school resource management. In this paper, we have addressed the prediction of secondary student grades of two core classes (Mathematics and Portuguese) by using past school grades (first and second periods), demographic, social and other school related data. Three different DM goals (i.e. binary/5-level classification and regression) and four DM methods, i.e. Decision Trees (DT), Random Forests (RF), Neural Networks (NN) and Support Vector Machines (SVM), were tested. Also, distinct input selections (e.g. with or without past grades) were explored. The obtained Results reveal that it is possible to achieve a high predictive accuracy, provided that the first and/or second school period grades are known. This confirms the conclusion found in [16]: student achievement is highly affected by previous performances. Nevertheless, an analysis to knowledge provided by the best predictive models has shown that, in some cases, there are other relevant features, such as: school related (e.g. number of absences, reason to choose school, extra educational school support), demographic (e.g. student's age, parent's job and education) and social (e.g. going out with friends, alcohol consumption) variables. This study was based on an off-line learning, since the DM-techniques were applied after the data was collected. However, there is a potential for an automatic on-line learning environment, by using a student prediction engine as part of a school management support system. This will allow the collection of additional features (e.g. grades from previous school years) and also to obtain a valuable feedback from the school professionals. Furthermore, we intent to enlarge the experiments to more schools and school years, in

order to enrich the student databases. Automatic feature selection methods (e.g. filtering or wrapper) will also be explored, since only a small portion of the input variables considered seem to be relevant. In particular, this is expected to benefit the nonlinear function methods (e.g. NN and SVM), which are more sensitive to irrelevant inputs. More research is also needed (e.g. sociological studies) in order to understand why and how some variables (e.g. reason to choose school, parent's job or alcohol consumption) affect student performance.

References:

[1] Duan, y., et al., (2010), "Intelligent Student Engagement Management - Applying Business Intelligence in Higher Education", International Conference on Information and Social Science (ISS) & MLB, Vol. 24-26, pp.12-20.

[2] Foley, E., Guillemette, M.G., (2010), "What is Business Intelligence?', International Journal of Business Intelligence Research, VOL. 1, NO.4, PP.1-28.

[3] Zhang, y., et al., (2010), "Use data mining to improve student retention in higher education- a case study", ICEIS Use Data Mining Ying. Vol., 38, No. 5, pp.488-501.

[4] Beatriz Piedade, m., and Yasmina Santos, m., (2010), "Business Intelligence in Higher Education", Business Intelligence in Higher Education, Vol. 16-19, pp. 1-5.

[5] Golfarelli,M. , & Rizii,S. (2009), "Data Warehouse Design: Modern Principles and Methodologies". McGraw-Hill. Vol. 5, pp. 12-19.

[6] Wixom, B, 2010, "The BI-Based Organization". International Journal of Business Intelligence, Vol.1, No.4, pp., 13-28.

[7] Olszak, C., & Ziemba, E. (2007). "Approach to Building and Implementing Business Intelligence Systems". Interdisciplinary Journal of Information Knowledge and Management, Vol. 6, pp. 126-132.

[8] Golfarelli, M. , & Rizii, S. (2009), "Data Warehouse Design: Modern Principles and Methodologies" . McGraw-Hill. Vol. 5, pp. 26-39.

[9] Williames, S., Williams, N., 2007, "The Profit Impact of Business Intelligence", Morgan Kaufmann Publishers, Vol. 6, pp. 457–464

[10] Azma, F., & Mostafapour, 2012, "M. Business intelligence as a key strategy for development organizations". Procedia Technology, Vol. 8. Pp. 102-106.

[11] KayvanJoo A.H, Ebrahimi M, Haqshenas G. 2014, "Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms". BMC research notes; Vol. 7, No. 1, pp. 565.

[12] Ogan M. Yigitbasioglu and Oana Velcu, 2012 "A review of dashboards in performance management: Implications for design and research," International

Journal of Accounting Information Systems, Vol. 13, pp. 41 - 59.

[13] WAYNE W. ECKERSON, 2011, "Performance Dashboards Measuring, Monitoring, and Managing Your Business", New jersey: John Wiley & Sons, Vol. 7, pp.13-18.

[14] Presthus, w., and Bygstad, b., 2012 "Business Intelligence in College: A Teaching Case with Real Life Puzzles", Journal of Information Technology Education: Innovations in Practice, Vol 11, pp.1-17.

[15] Manjunath T. N, et al., (2012), "Realistic Analysis of Data Warehousing and Data Mining Application in Education Domain", International Journal of Machine Learning and Computing, Vol. 2, No. 4, pp.419-422.

[16] Breiman L., (2001), "Random Forests. Machine Learning", IEEE Frontiers in Education , Vol. 45, No. 1, pp. 5–32.

[17] Breiman L.; Friedman J.; Ohlsen R.; and Stone C.,. (1984), "Classification and Regression Trees. Wadsworth, Monterey, CA". Vol. 1, pp. 18-28.

[18] Cortez P., In press. RMiner: (2007), "Data Mining with Neural Networks and Support Vector Machines using R. In R. Rajesh (Ed.)" , Introduction to Advanced Scientific Softwares and Toolboxes. Eurostat,. Early school-leavers. http://epp.eurostat.ec.europa.eu/ [Accessed on September 2016]..

[19] Flexer A., (1996), "Statistical Evaluation of Neural Networks Experiments: Minimum Requirements and Current Practice". In Proceedings of the 13th Euro- pean Meeting on Cybernetics and Systems Research. Vienna, Austria, Vol. 2, pp. 1005–1008.

[20] Hastie T.; Tibshirani R.; and Friedman J., 2001. "The Elements of Statistical Learning: Data Mining" , Inference, and Prediction. Springer-Verlag, NY, USA, Vol. 6, pp.56-68.

[21] Kotsiantis S.; Pierrakeas C.; & Pintelas P., (2004), "Predicting Students' Performance in Distance Learning Using Machine Learning Techniques". Applied Artificial Intelligence (AAI), Vol. 18, No. 5, pp. 411–426.

