Expert Systems with Applications xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Student data mining solution–knowledge management system related to higher education institutions

Srečko Natek [a,*], Moti Zwilling [b]

[a] School for Social and Business Studies, Slovenia
[b] Netanya Academic College, Israel

ARTICLE INFO

Keywords:
Data mining
Knowledge management system
Student's success rate
Data mining for small data set
Higher education institution
Educational data mining

ABSTRACT

Higher education institutions (HEIs) are often curious whether students will be successful or not during their study. Before or during their courses the academic institutions try to estimate the percentage of successful students. But is it possible to predict the success rate of students enrolled in their courses? Are there any specific student characteristics, which can be associated with the student success rate? Is there any relevant student data available to HEIs on the basis of which they could predict the student success rate? The answers to the above research questions can generally be obtained using data mining tools. Unfortunately, data mining algorithms work best with large data sets, while student data, available to HEIs, related to courses are limited and falls into the category of small data sets. Thus, the study focuses on data mining for small student data sets and aims to answer the above research questions by comparing two different data mining tools. The conclusions of this study are very promising and will encourage HEIs to incorporate data mining tools as an important part of their higher education knowledge management systems.

1. Introduction

In knowledge management process, data mining technique can be used to extract and discover the valuable and meaningful knowledge from a large amount of data. Nowadays, data mining has given a great deal of concern and attention in the information industry and in society as a whole. This technique is an approach that is currently receiving a great attention in data analysis and it has been recognized as a newly emerging analysis tool (Osei-Bryson, 2010; Park, 2001; Sinha & Zhao, 2008; Tso & Yau, 2007; Wan & Lei, 2009; Zanakis & Fernandez, 2005; Zhuang, Churilov, Burstein, & Sikaris, 2009).

In the literature, one may notice that they are many areas which adapted this approach to solve their problems such as in finance, medical, marketing, the stock market, telecommunication, manufacturing, health care and customer relationship. However, the data mining application has not attracted much attention from people in relation to educational systems, despite the majority of HEIs admit, that course's students success rates are very important for their prestige and senior management of the academic institution. Thus the possibilities to deeply understand the reasons and accurately predict the student success have become very valuable (Dermol & Čater, 2013; Marjetič & Lesjak, 2012; Natek & Lesjak, 2013; Rojko, Lesjak, & Vehovar, 2011; Trunk Širca, Babnik, & Breznik, 2013).

OLAP solutions and statistical methods are well established tool to analyze data but data mining enable fresh approach to understand hidden patterns and data prediction. In particular, data mining has turned to be a very popular among researches because many "approachable" standalone or desktop data mining tools are available in the market. From the observation of different tools, one can notice the following as an example: Microsoft Excel, SPSS, Weka, Protégé as Knowledge Acquisition System and Rapid Miner. Some of them (e.g. MS Excel Mining tool) are normally available to HEI's professor and they can benefit from existing knowledge of the Excel. That is the reason to include it in the research. Weka was choosed among other desktop DM tools because of supporting the data mining analysis in very different way, compared with MS Excel. Moreover, Weka is considered as a very legitimate tool analysis where education management data is involved (Cristóbal, Sebastián, & Enrique, 2008; Romero & Ventura, 2006).

HEIs recently promote knowledge management as encouraging data mining environment for their professors and researchers.

* Corresponding author. Tel.: +386 41630053.
E-mail addresses: srecko.natek@mfdps.si (S. Natek), Moti.Zviling@gmail.com (M. Zwilling).

http://dx.doi.org/10.1016/j.eswa.2014.04.024
0957-4174/

In the recent years, data mining has been recognized as a contemporary tool for building knowledge management systems (Jashapara, 2011). A knowledge management approach to data mining process, otherwise associated with business intelligence technology, brings some new synergy and added value to data mining solutions (Wang & Wang, 2008). This study was focused on HEI's knowledge management systems as part of their educational and research work (Baker & Yacef, 2009; Delavari, Somnuk, & Beikzadeh, 2008; Pena-Ayala, 2014; Tai-Chang, Anjin, & Li-Chen, 2008).

Unfortunately most data mining techniques work best with very large samples. Following that, Andonie (2010), reported that some data mining techniques, such as neural networks may not be able to accomplish the learning task (during the training phase) since small datasets cannot provide enough data to fill the gaps between too small samples. As a consequence, several authors concluded that small datasets limit the scope of data mining technique (Yuan & Fine, 1998).

In the real-life there are many academic situations, where relatively small data sets are normal, at least from the HEIs point of view. Therefore, the data which is collected on courses that students take is a good example of such situation. Even if the course is attended by a relatively large group of students the relevant data is considered to be a small dataset.

Andonie (2010) even suggested that there is no universal optimal solution to the problem of small datasets. So what possibilities are available to the HEIs if they want to explore the success rate of students in various courses using contemporary but widely available data mining tools? Student's data, available to HEIs is limited and clearly fall into the category of small data set (Table 1). The described approach in this study is very popular and well supported by data mining tools.

The research goal is to focus on small student's data set which is analyzed using data mining approaches, thus trying to answer various HEIs questions, using the student's courses data and the data mining tools for analysis (which are normally available to average HEIs). For HEIs the most important outcome are relatively reliable students' success rate prediction and user friendly experience. Bukowitz and Williams (2000) advocate how technology is blurring the distinction between types of knowledge from unknown knowledge to a known one. Their study supports the conceptual model as data mining on student's data sets which discovered some interesting key influencer and "Final grade" prediction to a chosen course. Moreover, as stated by the authors, small HEIs data sets can be used for analysis with data mining techniques. In these cases specific algorithms can be used to overcome the limit of data which is often not the case since most algorithms are working on a large scale databases (Xindong et al., 2008).

2. Methodology

2.1. Building student success rate data mining model using the MS Excel tool

Several authors suggest different steps for building data mining model, focusing mainly on data transformation process. Turban advocate data preparation model with data consolidation, data cleaning, data transformation and data reduction phases (Turban, Sharda, & Delen, 2011).

Berry and Linoff defined comprehensive data mining building model as an iterative process with several steps (Berry & Linoff, 2000). The adopted model was used to establish the student success rate data mining model related to International School for Social and Business Studies in Slovenia:

The first step was data requirements identification where HEIs require predicting the student's success rate related to the specific course.

Data acquisition focused on Informatics courses for Bachelor Study Programs – Economy in Contemporary Society for the academic years 2010/11 (42 students), 2011/12 (32 students) and 2012/13 (32 students). The data was imported from the web application 'Novis' – Higher Education Information System – Professor Module to MS Excel Table Tools with Data Mining add-in and MS SQL Server, which was installed on a laptop computer.

Validate, explore, clean and transpose data were the next activities where irrelevant columns were deleted (e.g. group, remark, individual activities points, registration number), translation from Slovene to English (columns title and content) was implemented and columns order were adjusted accordingly (logical order for success prediction, the last column was set as the predictable attribute: "Final grade"). For more transparent prediction the different statuses were transformed (from several status lists to yes/no value). To avoid personal data problems, the student names were transformed into numbers.

In addition, to increase the informative value of prediction some derived variables were added (additional columns): study year, gender (taken from the student name) and the type of study (by merging full time and part time students).

Creation model data sets is the central step of building data mining model with the following student data sets columns (Table 1): Study year (2010/11, 2011/12, 2012/13), Student (order number 1–106), Gender (female, male), Student year of birth (e.g. 1988), Employment (no, yes), Status, e.g. sport etc. (no, yes), Registration (first, repeat), Type of study (full time, part time), Exam condition (no, yes), Activities points (0–50), Exam points (0–50), Final points (0–50) and predictable Final grade (1–10). The data sets were divided into two groups. The first group of data sets

Table 1
Examples of students data sets (attributes taken are: year of study, Student #, gender, year of birth, employment, status – taken sport course, registration status, Type of study either full-time/part-time, exam condition, points for activity participation, exam points, final points, Final grade out of 10).

Study year	Student	Gender	Year of birth	Employment	Status (sport...)	Registration	Type of study	Exam condition	Activities points (50)	Exam points (50)	Final points (100)	Final grade (10)
2010–2011	1	Female	1988	No	No	First	Full time	Yes	46	46	92	10
2010–2011	2	Male	1990	No	No	First	Full time	Yes	38	33	71	7
2010–2011	3	Female	1990	No	No	First	Full time	Yes	39	30	69	7
2010–2011	4	Female	1990	No	No	First	Full time	Yes	47	35	82	8
2010–2011	5	Female	1989	No	No	First	Full time	Yes	39	36	75	7
2010–2011	6	Male	1990	No	No	First	Full time	Yes	38	30	68	7
2010–2011	7	Female	1990	No	No	First	Full time	Yes	39	36	75	7
2010–2011	8	Female	1990	No	Yes	First	Full time	Yes	39	33	72	7
2010–2011	9	Male	1990	No	No	First	Full time	Yes	39	38	77	8
⋮												
2012–2013	106	Female	1990	No	No	First	Part time	Yes	44	30	74	

Please cite this article in press as: Natek, S., & Zwilling, M. Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications* (2014), http://dx.doi.org/10.1016/j.eswa.2014.04.024

was determined as a training data taken from the courses of Informatics during the years 2010–11 and 2011–12 (74 rows). The second group of data sets was determined as predictable data taken from the courses of Informatics during the years 2012–13 (32 rows, without predictable "Final grade" column). In these courses, there 106 students who were included in the research (rows) participated. Regarding the Informatics 2012–13 course the actual "Final grade" was available for testing the results (32 row of actual "Final grade" data). The described data set from small number of examples but their content was estimated to be relevant for machine learning (Han, Kamber, & Pei, 2012).

In the next technological step the Data Mining Technology were chosen. Where, the Microsoft company generally offered three possibilities of data mining level for analysis. The basic level included MS Excel Table tools that include: data mining features analysis. The intermediate level included MS Excel Data Mining Adds-in features. The Expert level included MS SQL Server Data Mining capabilities. All levels of data mining are using algorithms from MS SQL Server Data mining, with different user interface, different technique and number of parameters, used to manage data mining process. For research purpose the basic level were chosen.

After this step where the Data Mining modeling Technique were chosen, MS Table tools were utilized to offer several Data Mining Techniques: Analyze key influencers, Detect Categories, Fill from Examples, Forecast, Highlight Exceptions, Scenario Analysis, Prediction Calculation and Shopping Basket Analysis. Some analysis are not suitable for a given student data sets, e.g. Forecast, Scenario Analysis and Shopping Basket due to data or content limitation. Among other techniques, the most usable technique is the Key Influencer and Fill from Examples, which were dynamically used for the following research.

Regarding the Model training, a Fill from example technique was used. Analysis was repeated several times, by choosing the "Final grade" as a prediction column (e.g. selecting column containing the examples). After choosing different combination of columns the authors found out the most relevant attributes by omitting student number (not relevant at all) and Exam points columns (directly defined the Final grade). The research dynamically explore the Microsoft Excel Data mining tool to build data mining model and thus relevant research datasets.

Finally the best model was chosen and the results were evaluated. By comparison the results of prediction for student's during the year 2012/13 according to the prediction class attribute: "Final grade", with the actual Final grade of the same student the authors could choose the best parameters for the Fill from example data mining technique. The final Data mining Fill from example predictive model was ready for prediction as a 'future student Final grade ', which was based on already known instances. The model could be used by confidence, until several assumptions were obtained as True (Berry & Linoff, 2000). The first assumption assumed that the past results were considered as good predictors of the future (if the student generation is changed, therefore; their attitude and behavior may vary from the past and create different student pattern). The next assumption assumed that the data was available at hand. The researchers assumed that the data will be always available for HEIs. The last assumption assumed that the data contained what the researchers aimed to predict – the research data mining model as a result contains the relevant data for analysis.

2.2. Building student data mining model with Weka

The data was also built using the WEKA Data Mining tool, were the following steps were conducted: (Fig. 1). In general, the first step in working with data mining analysis tool is a creation of data sets (Step 1). In this work, the data was saved as described above, except from the last column ("the predicted class") – which was attributed as the – 'Final grade'. This attribute was transformed into an ordinal one using the following index: To values between 8 and 10 a "High" value was assigned, to values between 6 and 7 a "Medium" value was assigned and to lower values than 6 a "Low" value was assigned. (In the case of decision tree M5P the class attributes remained the same as numerical).

In the next step the Data Mining Technology were chosen, and the "Weka" tool was found as the one that could be used for the analysis phase (Step 2).

Next the specific technique which was utilized using the "Weka" tool was conducted (Step 3). In this phase three various decision trees techniques were evaluated: J48, M5P and RepTree. These techniques were chosen as best suited for data analysis that was implemented in the last phase (Step 4). (In the case of M5P no conversion of values related to the class attribute was necessary and the Final grade values were remained as numbers).

Finally, in the last phase (Step 5), the data was analyzed by each of the models, and the best one was chosen. In other words – the model that best predicted the class attribute value, with a high accuracy rate was selected. (Each model was evaluated on a training set and its accuracy was evaluated again on the test set respectively). As much as the 'gap' between the trained predicted value and the actual data was lower the model was found as a better predictor for analysis regarding the actual data.

3. Results

3.1. Key influencer for student Final grade

At the beginning of the study, the researches aimed to find out which data (variables) are most powerful influencers for the Final grade prediction. Key Influencer Analysis is the right and useful tool for the job. Table 2 exhibits the most powerful relative impact for different variables. For example, the last study year (2012–13), where "Final grade" is empty (the authors aimed to predict it) shows "100" relative impact to favor empty column. Similar, high final points strongly favor high "Final grade", and low "Activity points" or "Exam points" favor low "Final grade". The results suggested to eliminate some of these columns from final prediction or at least to try several combination of different columns to produce relevant results.

3.2. Fill from example – student "Final grade" prediction for 2012/13

The next step of study aimed to predict the Student "Final grade" for the missing 2012/13 data by using Fill from example

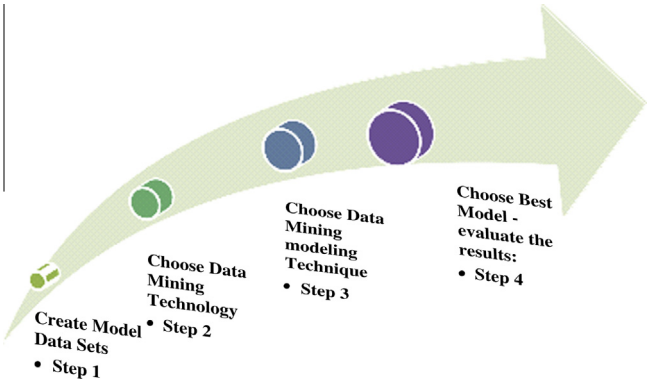

Fig. 1. Data mining process using the weka tool – the process is divided into 4 steps: model creation and data sets preparation (Step 1), choosing the technology for data mining analysis [in this research either excel/weka] (Step 2), choose of modeling technique – such as decision trees or regression (Step 3), choose the best model to evaluate the results (Step 4).

Please cite this article in press as: Natek, S., & Zwilling, M. Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications* (2014), http://dx.doi.org/10.1016/j.eswa.2014.04.024

data mining technique, based on machine learning using previous instances from 2010/11 and 2011/12. For more relevant result "Student", "Exam points" and "Final points" were eliminated from analysis. The Pattern Report in Table 3 shows a diversity of influences. The results justified the author's decision to eliminate powerful key influencer from analysis.

Prediction for student "Final grade" for 2012/13 was the final part of data mining analysis, using a Fill from Example Data Mining technique. The results for 32 students (rows) are shown in Table 4. The prediction based on data - knowledge, hidden in the student data. Different algorithms were used to exploit data from year 2010/11 and 2011/12 for data mining prediction model (training data) and were aimed to predict the missing student "Final grade" data for the year 2012/13. As noticed, the entire student's" Final grade" are positive (above 6) because the majority of previous results were found as positive too.

3.3. Data interpretation

Interpretation is considered as the final but vital phase of data mining (Negnevitsky, 2011). To evaluate the relevance of data mining student "Final grade" prediction the study compared the actual student's "Final grade" with the predicted one (Table 5). The following conclusions were calculated and discussed:

Table 2
Key influencer report for final grade (the table contains the following values: year of study, favors, relative impact and final/activities points).

Filter by 'column' or 'favors' to see how various columns influence 'Final grade 10'

Column	Value	Favors	Relative impact
Study year	2012–13	⟨Empty⟩	100
Final points _100_	92	10	100
Activities points _50_	39–42	7	100
Final points _100_	77	8	100
Final points _100_	85	9	100
Exam points _50_	9	2	100
Final points _100_	53	2	100
Exam points _50_	13	2	100
Final points _100_	51	2	100
Exam points _50_	8	4	100
Final points _100_	41	4	100
Exam points _50_	21	4	100
Final points _100_	66	6	100

Table 3
Fill from example analyze for student "Final grade" (the table contains several values/attributes taken for analysis such as: registration [first/repeat], activities points, etc.).

Filter by 'column' or 'favors' to see how various columns influence 'Final grade 10'

Column	Value	Favors	Relative impact
Type of study	Part time	10	13
Activities points _50_	43,539–47,000	10	13
Registration	Repeat	2	100
Study year	2010–11	2	13
Type of study	Part time	4	17
Gender	Male	6	33
Study year	2011–12	6	21
Activities points _50_	33,000–38,902	7	43
Status _sport___	Yes	7	36
Employment	Yes	7	31
Activities points _50_	38,902–41,221	7	18
Year of birth	1.967,148–1.982,240	8	25
Activities points _50_	43,539–47,000	8	21
Activities points _50_	43,539–47,000	9	36
Type of study	Part time	9	24
Employment	No	9	10
Activities points _50_	41,221–43,539	9	10

In the list of 32 students from Table 5, only 20 students already approached the exam (actual Final grade in third column). Using the data mining approaches the models predicted all 32 students "Final grade", included 12 (37.5%) without actual "Final grade". With zero "Final grade" tolerance, the prediction was successful for 7 "Final grade" (21.8%). With +– 1 "Final grade" tolerance the prediction was successful for 11 "Final grade" (43.3%). The positive "Final grade" prediction was successful for 13 of 32 students (40.6%). And finally if 12 students (without actual "Final grade") were omitted from the data, the positive "Final grade" prediction was considered as successful for 13 of 20 students (65%), which could be evaluated as generally positive results.

The study's results are not very clear. From observation of the final interpretation, it seems that the positive "Final grade" prediction was successful for 13 of 20 students (65%) indicating that the potential of data mining technique even for small student data sets is still of importance. However, if a larger data set or even more relevant student data was available the result relevance could be improved, particularly when parameters and columns are carefully combined in data mining analysis pre-process.

3.4. Key influencer for student "Final grade attribute" using the Weka tool

Designing and offering new courses to students is a very important task to higher education decision makers, especially in dynamic and changed environment, where data sets regarding students are limited and does not contain substantial information. In this case "weka" was found by the authors to be used as a data mining analysis tool with no obstacle. In addition, this tool was found (during analysis) as a good predication on the success of students in their studies along the years, where the Final grade attribute could be used as a key influencer for prediction alongside other attributes such as: Type of studies (Part time/Full time), Age, Employment and the Type of study taken by the student. This finding, that using only several attributes for analysis, where the Final grade is used as a key player in the over whole prediction process was found as very interesting for implication in the higher education system from the aspect of design, offer and budget planning for higher educational programs.

3.5. REPTree model

Training phase indicated 97.0588% where 66 instances were correctly classified and 2.9412% (2 instances) incorrectly classified (Fig. 2).

Testing phase indicated 100% (20/20 instances) correctly classified Instances.

3.6. J48 model

Training phase indicated 98.5294% (67 instances) which were correctly classified where 1.4706% (1 instance) was incorrectly classified, as presented in Fig. 3.

Testing phase indicated 90% (18/20 instances) correctly classified instances (see Fig. 4). Q4

3.7. M5P model

M5P model indicated correlation coefficient of 0.9358 and relative absolute error of 32.2884%.

Evaluation of results related to Table 6, revealed that 18 out of 20 instances were correctly classified by the method J48 and fully classified (20 of 20) where the REPtree method was taken. Therefore when using a small data set of data along with structured data

Please cite this article in press as: Natek, S., & Zwilling, M. Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications* (2014), http://dx.doi.org/10.1016/j.eswa.2014.04.024

Table 4
"Final grade" prediction for 2012/13 (the table contains the following attributes: study year, student #, Final grade out of 10, and Final grade prediction received by data mining analysis using the Excel tool).

Study year	Student	Gender	Year of birth	Employment	Status (sport...)	Registration	Type of study	Exam Condition	Activities points (50)	Exam points (50)	Final Points (100)	Final grade (10)	Final grade _10__Extended
2012–13	75	Female	1990	No	No	First	Full time	Yes	38	22	60		7
2012–13	76	Male	1990	No	No	First	Full time	Yes	38				7
2012–13	77	Male	1992	No	No	First	Full time	Yes	44				8
2012–13	78	Male	1992	No	No	First	Full time	Yes	38	30	68		7
2012–13	79	Male	1990	No	No	First	Full time	Yes	43	30	73		8
2012–13	80	Male	1983	No	No	First	Full time	Yes	43	19	62		8
2012–13	81	Female	1991	No	No	First	Full time	Yes	43	49	92		7
2012–13	82	Female	1973	Yes	No	First	Full time	Yes	44	42	86		8
2012–13	83	Female	1989	No	No	First	Full time	Yes	42				7
2012–13	84	Female	1989	No	No	First	Full time	Yes	38				7
2012–13	85	Female	1990	No	No	First	Full time	Yes	38				7
2012–13	86	Female	1976	Yes	No	First	Full time	Yes	44	36	80		8
2012–13	87	Female	1987	No	No	First	Full time	Yes	44	32	76		8
2012–13	88	Male	1992	No	No	First	Full time	Yes	38	39	77		7
2012–13	89	Female	1992	No	No	First	Full time	Yes	43	24	67		7
2012–13	90	Female	1992	No	No	First	Full time	Yes	44				9
2012–13	91	Male	1992	No	No	First	Full time	Yes	38	30	68		7
2012–13	92	Male	1992	No	No	First	Full time	Yes	38	6	44		7
2012–13	93	Male	1992	No	No	First	Full time	Yes	38	24	62		7
2012–13	94	Female	1992	No	No	First	Full time	Yes	38				7
2012–13	95	Male	1991	No	No	First	Full time	Yes	43	36	79		8
2012–13	96	Male	1987	No	No	First	Full time	Yes	38	32	70		7
2012–13	97	Female	1990	No	No	Repeat	Full time	Yes	41				7
2012–13	98	Female	1991	No	No	First	Full time	Yes	38				7
2012–13	99	Female	1968	Yes	No	First	Full time	Yes	44	38	82		8
2012–13	100	Female	1986	No	No	First	Full time	Yes	44	46	90		8
2012–13	101	Male	1989	No	No	First	Full time	Yes	38	17	55		7
2012–13	102	Male	1990	No	No	First	Part time	No					7
2012–13	103	Female	1988	No	No	First	Part time	Yes	38	25	63		7
2012–13	104	Female	1969	No	No	First	Part time	Yes	44				8
2012–13	105	Female	1988	No	No	First	Part time	Yes	44				9
2012–13	106	Female	1990	No	No	First	Part time	Yes	44	30	74		9

Table 5
Actual and predicted "Final grade" comparison (study year 2012–13) using 2 decision tree methods. Analysis was conducted by "Weka". The Final grade prediction (out of 10) is compared with the actual Final grade. Other attributers taken are: Student # & year of study.

Study year	Student	Final grade (10)	Prediction
2012–13	75	4	7
2012–13	76		7
2012–13	77		8
2012–13	78	7	7
2012–13	79	7	8
2012–13	80	3	8
2012–13	81	10	7
2012–13	82	9	8
2012–13	83		7
2012–13	84		7
2012–13	85		7
2012–13	86	8	8
2012–13	87	8	8
2012–13	88	8	7
2012–13	89	4	7
2012–13	90		9
2012–13	91	7	7
2012–13	92	1	7
2012–13	93	4	7
2012–13	94		7
2012–13	95	8	8
2012–13	96	7	7
2012–13	97		7
2012–13	98		7
2012–13	99	8	8
2012–13	100	9	8
2012–13	101	3	7
2012–13	102		7
2012–13	103	4	7
2012–13	104		8
2012–13	105		9
2012–13	106	7	9

mining algorithms (decision trees) one can notice that the over whole prediction rate is found as high (above 80%).

4. Discussion

The study explores the possibility to predict the success rate of students enrolled to an academic course using a contemporary data mining tools normally available to HEIs. The research clearly exhibits that available desktop data mining tools have matured in terms of their usability and ease of use, and provide usable results without extensive investment.

The results of the study yield that student data, available to higher education decision makers (such as professors) via export/import features, carries enough student-specific characteristics (or in other words – information) in the sense of hidden knowledge which can be successfully associated with student success rates.

Despite the well-known fact that data mining algorithms work best on large data sets, the study focused on student data available to HEIs which is limited and clearly falls into the category of a small data set. Results show that small student data sets in the specific data mining analysis did not limit the use of data mining tools.

Moreover the Weka tool, that was used as a comparative analysis to MS Excel, showed that by using decision tree models a high prediction accuracy (especially with the REPTree model) is obtained (the accuracy was verified during the test phase). Moreover, M5P regression tree did not perform as well as the other decision trees since it required more assumptions than the formers. In fact, The REPTree was found as a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. The model only sorted values for numeric attributes once. Missing values were dealt with using C4.5's method of using fractional

Please cite this article in press as: Natek, S., & Zwilling, M. Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications* (2014), http://dx.doi.org/10.1016/j.eswa.2014.04.024

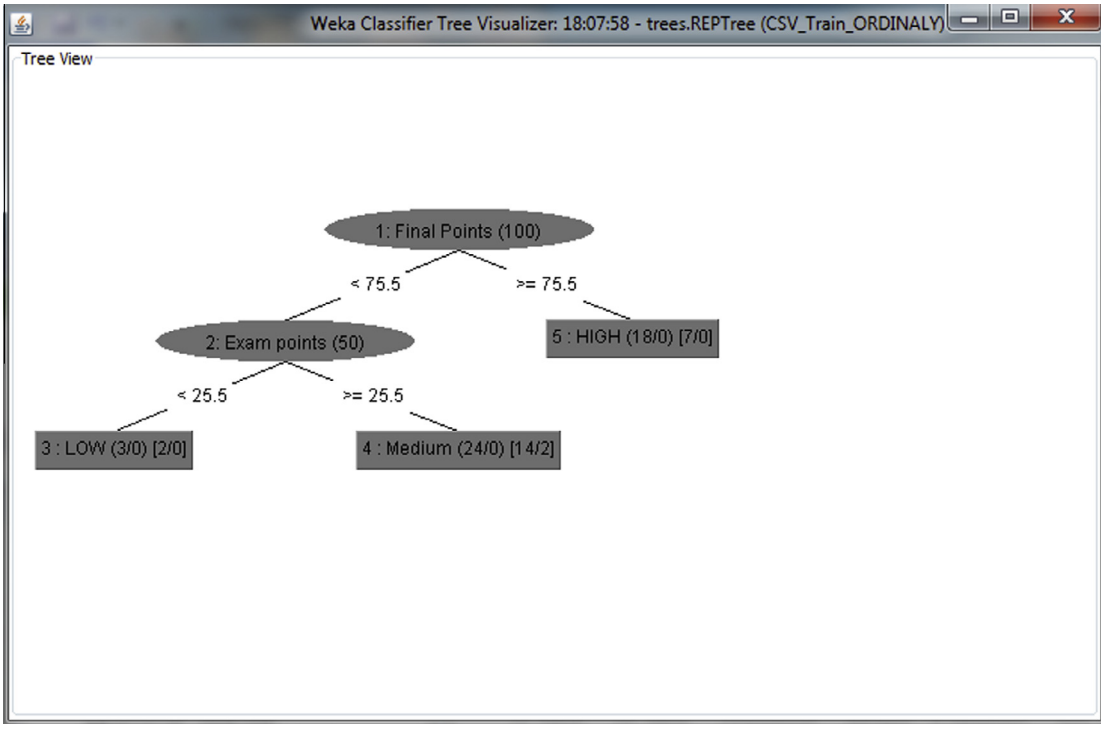

Fig. 2. "REPTree Decision Tree Model (constructed in the train phase) – the root contains the final points attribute where the classification of Exam points and final points (low, medium, high) is represented in the leafs.

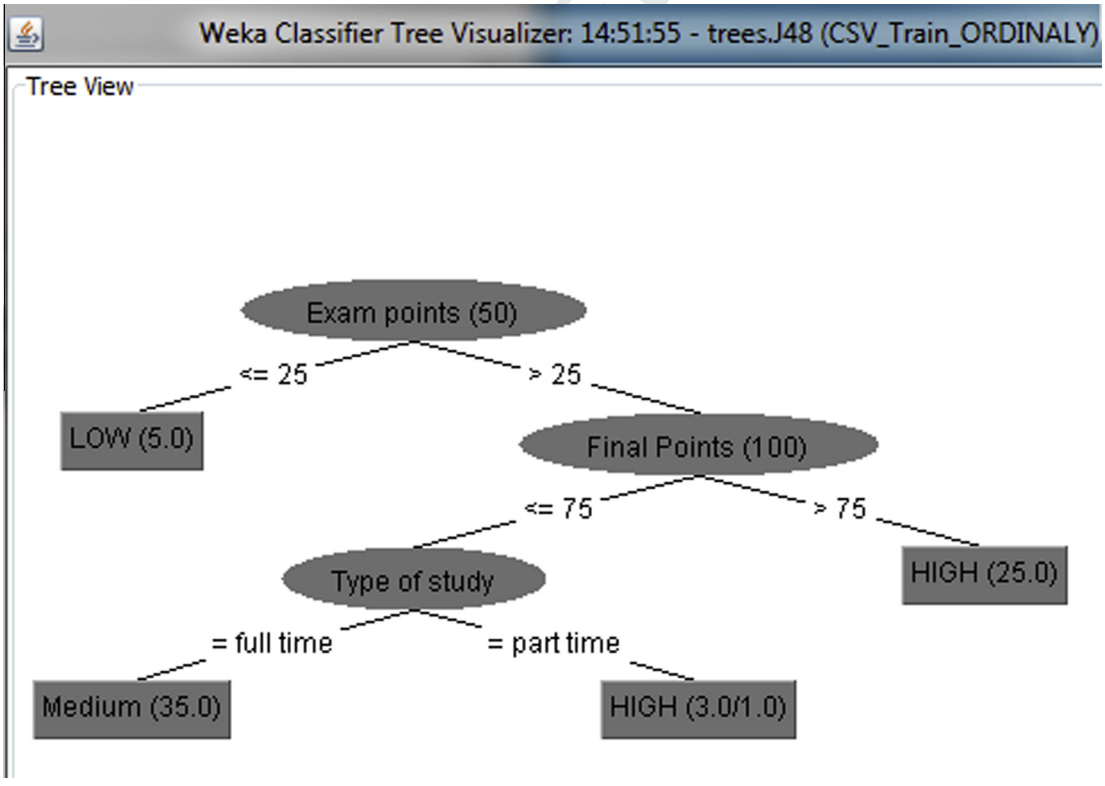

Fig. 3. "J48 Decision Tree Model (constructed in the train phase) – the root contains the exam points attribute where the classification of final points and type of study (low, medium, high) is represented in the leafs.

instances. From the above, it may be deduced that since the predicted attribute was transformed into an ordinary value, REPTree was less sensitive to missing values than J48 thus prediction accuracy on the test set was better performed, however on the training set it was shown that a slightly less performance than J48 (~97% vs. ~98%) prediction rate is achieved. Hence, J48 is a slightly modified C4.5 algorithm that generates a classification-decision tree for the given data-set by recursive partitioning of data (the decision is

Please cite this article in press as: Natek, S., & Zwilling, M. Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications* (2014), http://dx.doi.org/10.1016/j.eswa.2014.04.024

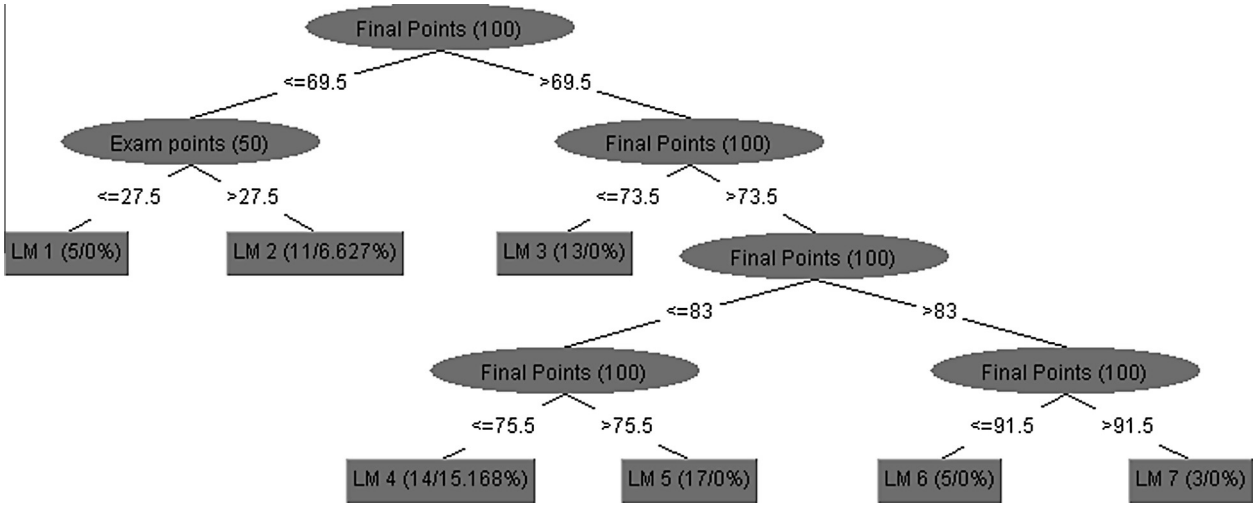

Fig. 4. "M5P Decision Tree Model (constructed in the train phase) – the root contains the final points attribute where the classification of final points is represented in the leafs.

Table 6
"Final grade" prediction for 2012/13.

Study year	Student	Final grade (10)	Prediction-J48	Prediction REPTree
2012–13	75	4 (low)	Low	Low
2012–13	76			
2012–13	77			
2012–13	78	7 (medium)	Medium	Medium
2012–13	79	7 (medium)	Medium	Medium
2012–13	80	3 (low)	Low	Low
2012–13	81	10 (high)	High	High
2012–13	82	9 (high)	High	High
2012–13	83			
2012–13	84			
2012–13	85			
2012–13	86	8 (high)	High	High
2012–13	87	8 (high)	High	High
2012–13	88	8 (high)	High	High
2012–13	89	4 (low)	Low	Low
2012–13	90			
2012–13	91	7 (medium)	Medium	Medium
2012–13	92	1 (low)	Low	Low
2012–13	93	4 (low)	Low	Low
2012–13	94			
2012–13	95	8 (high)	Low	High
2012–13	96	7 (medium)	Medium	Medium
2012–13	97			
2012–13	98			
2012–13	99	8 (high)	High	High
2012–13	100	9 (high)	High	High
2012–13	101	3 (low)	Low	Low
2012–13	102			
2012–13	103	4 (low)	Low	Low
2012–13	104			
2012–13	105			
2012–13	106	7 (medium)	High	Medium

grown using Depth-first strategy). The algorithm considered all the possible tests that can split the data set and select a test that yields the best information gain. Concerning the evaluated data, J48 was less accurate than REPTree, which could be postulated since it is considered as a more sensitive algorithm than REPTree.

5. Conclusions and future work

The study answers the research questions and supports the promising conclusions that data mining has a real potential to become a serious part of HEI's knowledge management systems. According to Nonaka's process model of a knowledge-based firm (Nonaka, Toyama, & Hirata, 2008) it is very important that organizations will adopt critical decisions based on data analysis rather than relying on their employees' heads. In other words, sometimes it is more important to explore the hidden knowledge in an organization's data and transform it into an explicit knowledge to improve the decision making process rather than relying on experience or rule of thumbs. Concerning the discussed study the quality of the educational process is very important.

The research results clearly support the conclusion that data mining is generally not limited to large data sets as the majority but not all of the authors and previous research anticipated this finding. Specific use of data mining techniques for structured small data set can also serve the usable results. Data mining for small student data sets from research is relevant example of effective use of data mining technology to develop the HEI's knowledge management systems in the education domain. In the past, there were several attempts to predict the student educational success with various successes. The research focused on small student data sets and the results were obtained by comparable using of two different data mining technologies (Microsoft and Weka). Both results show a relative high prediction success and thus strength the conclusion that they can be relevant tool for developing HEI's knowledge management systems as a main contribution of the paper.

Another important conclusion is that several decision trees algorithms which are common to be used in data mining are in fact very practical to work on small data sets. These algorithms yield a high prediction score. Weka as a "free" and open source data mining analysis tool is shown as a promising tool to evaluate and analyze small data sets to various problems and is found as a promise tool to analyze and interpret data to improve knowledge management in organizations such as the one that is presented in this study.

The research conclusions can serve the HEI's management to discuss the HEI and academic reactionaries about problems regarding student's educational success. Some of these problems are hard to solve simply because usual statistical methods are not deep enough to discover the hidden patterns and knowledge, useful for education processes planning and organizing. By knowing the student success rates in advance HEI's management can be contributed effectively by planning the strategy and policy of the institution quite a long time before even the enrolment process of the students. This contribution can be achieved as a substantial asset to invest into the data mining technology and solutions.

Please cite this article in press as: Natek, S., & Zwilling, M. Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications* (2014), http://dx.doi.org/10.1016/j.eswa.2014.04.024

The contribution of this study is expressed in new possibilities for HEI's management to early predict the student's success rate and prepare the measures to encourage students to increase their workload if the prediction does not meet the HEI's requirements:

- On the theoretical level, this study advocate small data sets data mining possibilities, supported by contemporary desktop data mining information technology.
- On the implementation level, the research data and results proved that the prediction is significant successful by both technologies, used in the research.
- Additionally, our findings constitute a practical guide for decision makers in HEI and higher education professors to effectively identify problems regarding student success, which is sometimes hidden in professor's data about their students.
- Finally, this model makes it possible to react as early as possible to improve student success rate, immediately after their enrolment or at any time during the study.

As we considered, the research can be further expanded to international comparable research of student data sets taken from different countries utilizing different data mining technologies. A comparable study of student success rates for different size of student data sets could greatly further expand the research potential and forming the new view how to effectively use the data mining technology and accompanied methods (algorithms) as an important tool for the development of HEI's knowledge management systems.

The authors hope that these conclusions will encourage HEIs management to improve the student's success rates also by incorporate data mining tools into their daily work.

6. Uncited reference

Q5 Becerra-Fernandez, Gonzales, and Sabherwal (2004).

Acknowledgment

This work has been supported by International School for Social and Business Studies.

References

Andonie, R. (2010). Extreme data mining: Inference from small datasets. *International Journal of Computers, Communications & Control, V*(3).

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining application in 2009: A review and future vision (Vol. 11, no. 1, pp. 3–17). Q6

Becerra-Fernandez, I., Gonzales, A., & Sabherwal, R. (2004). *Knowledge management, challenges, solutions, and technologies*. Prentice Hall: Pearson.

Berry, M., & Linoff, G. (2000). *Mastering data mining. The art and science of customer relationship management*. Wiley.

Bukowitz, W. R., & Williams, R. L. (2000). *The knowledge management fieldbook*. Prentice Hall.

Cristóbal, R., Sebastián, V., & Enrique, G. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368–384.

Delavari, N., Somnuk, P. A., & Beikzadeh, M. R. (2008). Data mining application in higher learning institutions. *Informatics in Education, 71*(1), 31–54.

Dermol, V., & Čater, T. (2013). The influence of training and training transfer factors on organisational learning and performance. *Personnel Review, 42*(3), 324–348.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining, concepts and techniques*. Morgan Kaufmann.

Jashapara, A. (2011). *Knowledge management, an integrated approach* (2nd ed.). Prentice Hall.

Marjetič, D., & Lesjak, D. (2012). Financing of higher education and the role and dilemmas of tariff groups. *International Journal of Management in Education, 6*(1/2), 56–72.

Natek, S., & Lesjak, D. (2013). Improving knowledge management by integrating HEI process and data models. *The Journal of Computer Information Systems, 53*(4), 81–86.

Negnevitsky, M. (2011). *Artificial intelligence, a guide to intelligent systems* (3rd ed.). Pearson Education Limited.

Nonaka, I., Toyama, R., & Hirata, T. (2008). *Managing flow, a process theory of knowledge-based firm*. Palgrave Macmillan.

Osei-Bryson, K.-M. (2010). Towards supporting expert evaluation of clustering results using a data mining process model. *Information Sciences, 180*(3), 414–431.

Rojko, K., Lesjak, D., & Vehovar, V. (2011). Information communication technology spending in (2008-) economic crisis. *Industrial Management and Data Systems, 111*(3), 391–409.

Romero, C., & Ventura, S. (2006). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135–146.

Sinha, A. P., & Zhao, H. (2008). Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decision Support System, 46*(1), 287–299.

Trunk Širca, N., Babnik, K., & Breznik, K. (2013). Towards organisational performance: Understanding human resource management climate. *Industrial Management and Data Systems, 113*(3), 367–384.

Tso, G. K. F., & Yau, K. K. W. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy, 32*, 1761–1768.

Turban, E., Sharda, R., & Delen, D. (2011). *Decision support and business intelligent systems* (9th ed.). Pearson.

Wan, S., & Lei, T. C. (2009). A knowledge-based decision support system to analyze the debris-flow problems at Chen-Yu-Lan River, Taiwan. *Knowledge-Based System, 22*(8), 580–588.

Wang, H., & Wang, S. (2008). A knowledge management approach to data mining process for business intelligence. *Industrial Management and Data Systems, 108*(5).

Xindong, W., Vipin, K., Ross Quinlan, J., Joydeep, G., Qiang, Y., Hiroshi, M., et al. (2008). Top 10 algorithms in data mining. *Knowledge Information Systems, 14*, 1–37. Q7

Yuan, J. L., & Fine, T. (1998). Neural-network design for small training sets of high dimension. *IEEE Transactions on Neural Networks, 9*.

Zanakis, S. H., & Fernandez, I. B. (2005). Competitiveness of nations: A knowledge discovery examination. *European Journal of Operational Research, 166*(1), 185–211.

Zhuang, Z. Y., Churilov, L., Burstein, F., & Sikaris, K. (2009). Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. *European Journal of Operational Research, 195*, 662–675.

Please cite this article in press as: Natek, S., & Zwilling, M. Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications* (2014), http://dx.doi.org/10.1016/j.eswa.2014.04.024