Student Data Analysis with RapidMiner

Jovica Krstevski¹, Dragan Mihajlov², Ivan Chorbev³

¹Innovaworks. ul. Partizanski Odredi 73, MK-1000 Skopje, Republic of Macedonia
jovicakrstev@yahoo.com

²,³Faculty of Electrical Engineering and Information Technologies, PO Box 574,
Skopje, Republic of Macedonia
{dragan, ivan}@feit.ukim.edu.mk

Abstract. The aim of this paper is to propose a methodology for modeling and implementing algorithms over data from student management systems with RapidMiner. RapidMiner is an environment for business analytics, predictive analytics, data mining, text mining, and machine learning. Traditional educational systems collect a lot of data about students. The collected data can be used to extract information and experiences. The extracted information can help school managers prepare better curriculum and manage the schools in an informed manner. Teachers can have advanced information about student progress and the potential weaknesses in courses. In this paper we describe how tools like RapidMiner can be used to extract information from raw student data. We have used student data from the Macedonian educational system, compared different algorithms and chose the most appropriate for the given model.

Keywords: Student Analysis, Academic Analytic, Educational Data Mining, RapidMiner.

1 Introduction

We are witnesses of how modern world education plays a significant role in people’s lives. The growing need for an educated workforce in the society requires education to be a leader in the reform of the economy. During the last decades an increasing number of educational organizations store their data in electronic format and the amount of stored information is increasing each day. In order to analyze collected data one needs a process where the data will be inspected, cleaned, transformed, and modeled. From the analysis one can extract information, propose conclusions and support decision making. Data analysis can be found in science, social and business domains. It has different approaches and includes versatile techniques with different names. This process in the educational area is called Educational Data Mining (EDM) [5], which converts student raw data from educational systems to information to help the educational process.
In this article we are comparing different analytical techniques for grouping students based on students’ data in a traditional educational system and finding rules for predicting student status and start program. We are using the RapidMiner tool for making this task easier for managers. The paper contains the following sections: Section 2 describes recent work in this area; Section 3 describes the background of the main analytic methods and algorithms in education; Section 4 describes the RapidMiner tool; Section 5 contains details from comparison of the analytic techniques; Section 6 contains the conclusions and further research.

2 Recent work

Presently, education institutions are just utilizing normal report with minimum knowledge exploration to analyze information regarding the performance of the institution such as student performance [16]. It is important for the educational managers to identify and analyze the relationships among different entities such as students, subjects, courses etc... Different areas of education are researched such as student performance [2], recommendations about difficulty and amount of work dictated for set of courses [19], enrollment management [20, 21], graduation [22], and retention [23]. Some authors propose solutions for software products which will help improve the educational system [2, 16, 17].

3 Background

An analytic process consists of data for analysis, algorithms and models which connect data, and algorithms.

The analytic process for decision support is based on the CRISP (Cross Industry Standard Process for Data Mining). The CRISP model provides a sequence of steps and directions to be followed in the data analysis process. The CRISP model consists of six phases: business and data understanding, data preparation, modeling, evaluation and deployment [25].

The business understanding phase begins with the setting up of goals for the project.

The data understanding phase starts with gathering data and continues with getting familiar with the data. Then one must identify data quality problems and lastly discover first insights into the data or detecting interesting subsets to form hypotheses for hidden information.

The data preparation phase includes all activities to create the final data set from the initial raw data. Data preparation tasks can be performed multiple times. Selection tables, records and attributes, and transforming and cleaning the data are tasks
performed in this phase. After these activities, data is prepared to be fed into the modeling tool.

The modeling phase contains selecting various modeling techniques, calibrating the parameters to optimal values, and applying these techniques. We can use a few techniques for the same problem type. Different techniques can have specific requirements on the form of data. In these cases going back to the data preparation phase is often required.

At the evaluation phase a model or models have been built, which are high quality, from a data analysis perspective. Before continuing with the next step (final deployment of the model), it is very important to thoroughly evaluate the model, and review all steps executed to create the model, in order to confirm all requirements.

The deployment phase depends on the project objectives; it can be simple as creating reports or as complex as implementing a recursive analysis process.

The classification technique is frequently used in educational analyses [15]. Classification is the technique in which the value of an attribute is predicted based on the value of other attributes, called predicting attributes. This technique contains a lot of different classifications methods, such as:

4 Data analysis with RapidMiner

4.1 RapidMiner Tool

RapidMiner, formerly YALE (Yet Another Learning Environment), is an environment for machine learning, data mining, text mining, predictive analytics, and business analytics. It is used for research, education, training, rapid prototyping, application development, and industrial applications [18]. We chose to use this tool for educational data because it is distributed under the AGPL open source license, and it is a data analytic tool used in real projects, ranked second in 2009 and first in 2010 according KDnuggets, a data-mining newspaper. RapidMiner provides a GUI to design an analytical process (reading data from source, transformations, applying algorithm). All GUI changes are stored in an XML (eXtensible Markup Language) file and then this file is read by RapidMiner to run the analyses.

RapidMiner contains all tools for data analysis from data processing (ETL), data modeling, data and result visualization. RapidMiner allows connections to the most varied of data sources such as Oracle, Microsoft SQL Server, MySQL, and access to Excel, Access as well as numerous other data formats.

4.1.1 RapidMiner Terminology

In the following section we will describe some of the terminology used by RapidMiner.

<table>
<thead>
<tr>
<th>Value Type</th>
<th>RapidMiner Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Nominal</td>
</tr>
</tbody>
</table>
### Numerical values

<table>
<thead>
<tr>
<th>Category</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integers</td>
<td>Integer</td>
</tr>
<tr>
<td>Real numbers</td>
<td>Real</td>
</tr>
<tr>
<td>Text</td>
<td>Text</td>
</tr>
<tr>
<td>2-value nominal</td>
<td>Binominal</td>
</tr>
<tr>
<td>Multi-value nominal</td>
<td>Polynomial</td>
</tr>
<tr>
<td>Date Time</td>
<td>date_time</td>
</tr>
</tbody>
</table>

Table 4.1 Value types in RapidMiner

- **Operator**: presents the super-class of any implemented operator in RapidMiner. Operator object can accept an array of input objects, does some operations or process and returns an array of output objects. An operator contains ports which are used for output, input objects and connects with other operators.
- **ExampleSet**: presents the data set; it can be passed to operator as input or delivered by operator as output.

### 4.2 Data Analyses Process

The data used in the analyses in this paper are stored in a Microsoft SQL Server database and RapidMiner features are used to connect with Microsoft SQL Server and retrieve student data. RapidMiner UI allows executing SQL Queries in SQL server for filtering table data, joining with other tables, or using existing functions and views.

The analyses process consists of the following steps:

#### 4.2.1 Data Preparation

The data preparation step in this analysis consists of finding the table which contains the data, relations between them and selecting the columns which are used for analyses. The following tables and columns were chosen:

<table>
<thead>
<tr>
<th>Table</th>
<th>Columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>Birth Date, Sex, Living Place, Previous Education, Previous Education Grade, Use Scholarship, Nationality, Status, Quota, Study Cycles, Start Year, Quota ID, Mother Education, Father Education, Birth Country, Living Country</td>
</tr>
<tr>
<td>Program</td>
<td>Program Name</td>
</tr>
<tr>
<td>Quota</td>
<td>Quota name</td>
</tr>
</tbody>
</table>

Table 4.2 Student tables with used columns
4.2.2 Data Selection and Transformation

This step includes creating tables or flat files which can be used from RapidMiner as import data for the read operator. Because RapidMiner can work with databases we used the table approach. This approach allows easily changing import columns while testing different models and algorithms. The table contains all the data queried from the tables and transformed. Transformation is required for some of the selected columns e.g. previous education column contains plenty of misspelled values. The next table shows some of this columns value before transformation and after transformation:

<table>
<thead>
<tr>
<th>Column:</th>
<th>Old Values</th>
<th>Transformed Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous Education</td>
<td>ХЕМИСКО</td>
<td>умарски фак.</td>
</tr>
<tr>
<td>Previous Education Grade</td>
<td>3.63</td>
<td>3.49</td>
</tr>
<tr>
<td>Program Name</td>
<td>Сметководство и Ревизија</td>
<td>Сметководство и ревизија 2008</td>
</tr>
</tbody>
</table>

Table 4.3 Column value transformations

The missing values in all the transformed columns were checked and populated with the NA element. After transformation and data surveys (MotherEducation, FatherEducation, BirthCountry, LivingCountry etc...) some columns were removed because they contained a lot of NA values or single values which do not have valuable information for further processing.

4.2.3 Implementation of Model

After transformation and cleaning we were ready to use the data from the sql table. Figure 4.1 shows the RapidMiner UI design flow.
Figure 4.1 RapidMiner UI design flow workplace

The first step is to add the read database operator, create a connection to our database server, and build the query which returns the data. After the read operator, we added the Set Role operator to convert one of the attribute roles from regular, which is an input variable for learning tasks, to label, which is a target learning attribute. We did this because all the returned attributes from the database operator are assigned a regular role. The next operator is cross-validation in order to estimate the performance of a learning operator. The figure 4.2 shows the components of Cross-validation.

![Cross-Validation component](image)

The cross-validation component contains both training and testing sections. The training section contains algorithms and all other necessary operators depending on the algorithm and the example set. The testing section must return a performance vector. This vector is generated by applying the model and measuring the model’s performance.

A very important thing to note here is the ability of RapidMiner to notify the user if there are missing operators or if the input example set has wrong attribute types. It allows the operator to fix the problem. After running the process, the result is shown in the workplace where the example set data can be examined, data can be viewed with various plots, and output values from the process can be examined (trees, rule inductions etc.).

5 Results from data analysis

In this part of the paper the result generated by the RapidMiner tool are presented. The goal was to build a model and select an algorithm by analyzing algorithm accuracy for the model. For this analysis we introduced a model that can give predictions for the student’s status and start program. The following attributes are part of the model: Status, Living Place, Nationality, Previous Education, Previous Education Grade, Sex, Quota Name, Start Program and Start Age. After all the preparatory steps, which were described in the previous sections of the document, the total number of samples that we used for training the model was 10181.
Table 5.1 shows values for each attribute.

<table>
<thead>
<tr>
<th>Status</th>
<th>Вонреден (1002), Редовен (9173), Во Мирување (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Living Place</td>
<td>Скопје (5331), Друго (1206), Прилеп (61), Кавадарци (203), Гостивар (213), Кичево (94), Берово (70), Куманово (46), Велес (66), Охрид (312), Илинден (125), Неготино (59), Битола (217), Тетово (290), Гостивар (213)</td>
</tr>
<tr>
<td>Nationality</td>
<td>Македонец (9113), Ром (31), Албанец (483), Турчин (139), Србин (202), * Друго (81)</td>
</tr>
<tr>
<td>Previous Education</td>
<td>Гимназија (5588), Економско (4188), Електро (102), Медицинско (100), Чемпирско (39), …</td>
</tr>
<tr>
<td>Previous Education Grade</td>
<td>3.00 (475), 5.00 (8103), 4.00 (1553), 2.00 (50)</td>
</tr>
<tr>
<td>Sex</td>
<td>male (3763), female (6418)</td>
</tr>
<tr>
<td>Quota Name</td>
<td>Не плаќа (3819), Државна Квота (3239), Кофинансирање (3123)</td>
</tr>
<tr>
<td>Start Program</td>
<td>финансиен менеджмент (2000), ва год општа (705), надворешна трговија (1130), маркетинг (1081), сметководство и ревизија (1278), економија (1427), менеджмент (1730)</td>
</tr>
<tr>
<td>Start Age</td>
<td>[0.000 ; 47.000]</td>
</tr>
</tbody>
</table>

Table 5.2 Operator accuracy

<table>
<thead>
<tr>
<th>Operator</th>
<th>Accuracy - Status</th>
<th>Accuracy - Start Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>RapidMiner - Decision Tree</td>
<td>90.10% +/- 0.04%</td>
<td>19.64% +/- 0.04%</td>
</tr>
<tr>
<td>RapidMiner - Naive Bayes</td>
<td>89.27% +/- 0.56%</td>
<td>22.80% +/- 1.11%</td>
</tr>
<tr>
<td>RapidMiner - Neural Net</td>
<td>90.20% +/- 0.14%</td>
<td>22.01% +/- 1.48%</td>
</tr>
<tr>
<td>RapidMiner - Rule Induction</td>
<td>90.10% +/- 0.04%</td>
<td>22.06% +/- 0.53%</td>
</tr>
<tr>
<td>Weka - W-J48</td>
<td>90.26% +/- 0.21%</td>
<td>21.96% +/- 1.32%</td>
</tr>
</tbody>
</table>

Table 5.2 Operator accuracy

The first accuracy column in the table 5.2 shows the accuracy results received when student status was predicted. According to the results, all algorithms gave similar results, but the best results were produced by W-J48 tree operator, which is C4.5 decision tree, and this trained model can be used for student status predictions. For better result understanding the RapidMiner tool offers visualizations for each algorithm.
Figure 5.3 gives a visual presentation of W-J48 decision tree for the status attribute. From the tree we can create following rules:
- If student start age <= 22 then status is ‘редовен’
- If student start age > 22 and Quota is ‘не плаќа’ then status is ‘редовен’
- If student start age > 22 and Quota is ‘кофинансирање’ then status is ‘вонреден’
- If student start age > 22 and Quota is ‘Државна квота’ and Previous Education Grade is = 2 or 3 or 4 then status is ‘вонреден’
- If student start age > 22 and Quota is ‘Државна квота’ and Previous Education Grade is = 5 then status is ‘редовен’

Figure 5.3 Status tree with W-J48

The second accuracy column in table 5.2 shows measurements received when student start program is predicted. For this attribute all retrieved results have low accuracy, and the trained model cannot be used for predictions.

6 Conclusion

This research is an attempt to use RapidMiner tool’s functions to analyze student academic data and introduce some of the possibilities which are offered to analyze data. We investigated only the key features of the tool and used it in one model to predict student status and start program. We hope that the created model and findings will be useful for further usage of RapidMinner in student data analysis. This work may help faculty managers to easily predict resources for the next academic year.

In our future work, we plan to extend the model with other student data related to the study cycle (courses, grades, etc.).
References

[5] A.N. Rafferty, M. Yudelson, Applying Learning Factors Analysis to Build Stereotypic Student Models, Stanford University, University of Pittsburgh
[10] P. Baepler, C. J. Murdoch, Academic Analytics and Data Mining in Higher Education, International Journal for the Scholarship of Teaching and Learning, Vol. 4, No. 2 (July 2010), ISSN 1931-4744 @ Georgia Southern University