

UNIVERSITAT POLITÈCNICA DE CATALUNYA- BARCELONATECH

Facultat d'Informàtica de Barcelona

Masters in Innovation and Research in Informatics

Master's Thesis

Data Mining in Learning Analytics

Joanna Sykurska

Supervisor:

Maria Ribera Sancho Samsó

Albert Obiols Vives

Barcelona, 04.02.2016

UNIVERSITAT POLYTÈCNICA DE CATALUNYA- BARCELONATECH
FACULTAT D'INFORMÀTICA DE BARCELONA

Joanna Sykurska

MSc THESIS

Data Mining in Learning Analytics

Barcelona, 2016

ABSTRACT

The goal of the Master's thesis was to perform the Data Mining part of a project "Learning Analytics for Secondary Schools" developed in inLab at Technical University of Catalonia. The tasks include introducing new indicators to an existing Learning Analytics project, perform the revision, validation and testing of the indicators. Finally, gathered indicators should result in notification report about student's motivation.

I have chosen this topic because it excited my curiosity and willingness to get to know the problems and possible solutions while working with data.

In the project for the ETL process Pentaho is used and for computation of indicators SQL as well as R. With help of Tableau inspection and revision of obtained data was possible.

As a final result I have implemented indicators which can assess students' motivation.

TABLE OF CONTENTS

1.	Introduction	3
1.1.	Motivation	3
1.2.	InLab FIB	3
2.	Learning Analytics	4
3.	Project context and formation.....	5
3.1.	Goal of LA4S	5
3.2.	Structure of LA4S.....	6
3.3.	Outcome of LA4S.....	6
3.4.	Project goal and outcome	7
3.5.	Gantt chart	8
3.6.	Budget.....	9
4.	Methodology.....	10
5.	Tools	12
5.1.	Pentaho	12
5.2.	SQL.....	12
5.3.	R	13
5.4.	Tableau	13
6.	Indicators	13
6.1.	Speed	14
6.2.	Intensity	15
6.3.	Persistence	16
6.4.	Choice.....	16
7.	Results	17
7.1.	Activity time and number of breaks	17
7.1.1.	First approach: from beginning till the end	17

7.1.2.	Second approach: interrupted time slots.....	19
7.1.3.	Justification of selected approach.....	23
7.2.	Starting Date	23
7.2.1.	First approach: quantiles of accesses	24
7.2.2.	Second approach: percentage of students	27
7.2.3.	Justification of selected approach.....	30
7.3.	Curiosity rate	31
7.4.	Forum activity	33
7.5.	Delivery Rate.....	36
7.6.	Validation	38
8.	Conclusions	38
9.	Summary.....	39
	References	40

1. Introduction

1.1. Motivation

I have chosen Data Mining as the topic of my Master's thesis because I got interested by its growing popularity. The importance of the information which can be gained from Data Mining is impressive. Computing abilities are being much easier with the use of cloud and obtained from it results are promising and encouraging. There are dozens of examples of possible applications of Big Data in the cloud and I am sure that this list will be extending very fast. Just as an example of application of Data Mining I can give two relatively different areas: business with targeted advertising or science with biochemical computations. In my opinion, Data Mining will get more and more popular in the future.

1.2. InLab FIB

My Master's thesis project has been carried out in inLab FIB at Technical University of Catalonia. InLab is an innovation and research laboratory gathering around 70 people of technical background including academics and students. From 2014 inLab started to be a member of CIT UPC (Centro de innovación) which is the largest technology centre in Catalonia. Projects developed by inLab FIB cover various topics, such as Smart City, Big Data and many more.

2. Learning Analytics

Learning Analytics is a new term which relates to usage of new technologies in the learning process. Research in this topic started in the last century with a set of conferences in 1998 where changes in the way of learning were analysed. This issues have been tackled due to application of Internet for learning purposes. In that time this trend was called Networked Learning [1]. The first time the topic Learning Analytics has been widely discussed was on first International Conference on Learning Analytics Knowledge (LAK) in 2011. Phil Long and George Siemens in their paper “Penetrating the Fog: Analytics in Learning and Education” have described the topic of Learning Analytics in details [2].

The exact definition of Learning Analytics is [2]:

„the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs.”

Traditional way of learning from books doesn't give us possibility to observe the learning process. However, due to growing popularity of online learning with use of various educational platforms there is a lot of data collected, which at first sight may not seem to carry extraordinary information. This data is called Big Data due to its unstructured form, big size and constant changes. These three factors describing Big Data are called 3V: variety, volume and velocity [3].

Huge amount of Big Data is being produced while we are online. Not only our final decisions for example in an online store are saved but also all previous actions which ended up in purchase of an item. How these data can be analyzed and what kind of advantage people take from it? There are a lot of examples, I will present a few. Thanks to Big Data telecommunication companies can predict if their client is about to churn. They can keep the undecided client, if they react fast and propose him/her a better offer. The Big Data analysis enables really deep insight into people's life, like for example shopping chain “Target”, which can estimate if a couple is expecting a baby [4]. Car insurance companies can assess from the activity on the Internet who is a good or a bad driver and thus offer the best insurance conditions. Other examples of applications [5]:

- Understanding and targeting customers
- Understanding and optimizing business processes

- Personal quantification and performance optimization
- Improving healthcare and public health
- Improving sports performance
- Improving science and research
- Optimizing machine and device performance
- Improving security and law enforcement
- Improving and optimizing cities and countries
- Financial trading

The last but not least application of Big Data can be also Learning Analytics. The advantages of applying Learning Analytics are broad. The teacher can react on students' learning process and change it according to his/her needs. Students who are endangered of dropping out can be spotted much earlier and some help for them can be provided. Studies can be tailored to students and improved according to their needs. However, not only teaching procedures can make use of LA. Also the administration can benefit by making their working procedures easier [2].

3. Project context and formation

3.1. Goal of LA4S

The goal of the project "Learning Analytics for Secondary Schools" (LA4S) is providing teachers with information about students' performance and their motivation. Around 90% of secondary schools in Catalonia are using Moodle-based educational platform named Agora. Via this network teachers can give various tasks to do at home. Such homework can be given for example in a form of a quiz solved online or an assessment to be uploaded. This Learning Analytics project is being developed in cooperation with government of Catalonia. It is expected to be in use in all region reaching 1500 schools and 400 thousand students [6].

3.2. Structure of LA4S

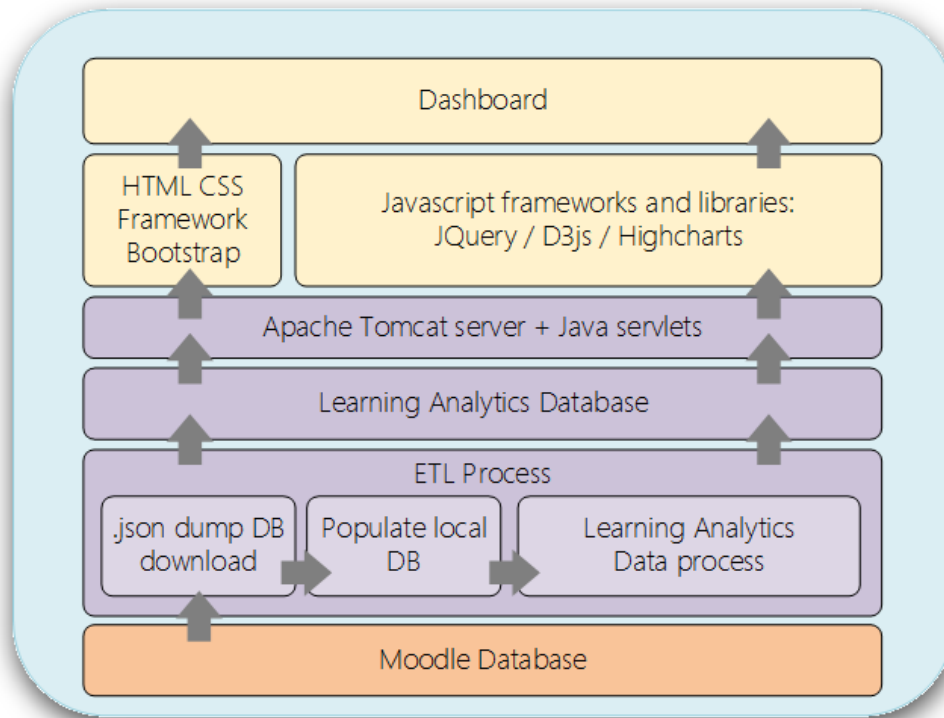


Figure 1: Architecture, author: J.Casanovas

The above presented architecture shows the data flow in the application. All the data used in the application come from educational platform Moodle. The ETL process (extraction, transformation, loading) is the first step on the server side. With the use of Perl scripts they are downloaded into local server database. After the computation of the indicators, obtained results are exported to Learning Analytics database. Finally, the visualization of the data is carried out by a web application.

3.3. Outcome of LA4S

In the picture below you can see an exemplary outcome of the calculations. The dashboard enables inspection of students' activity on Moodle. In the given example in the biggest plot we can see the accesses in time scale for English classes marked with colourful dots, representing different activities. At the bottom there are statistics about most common hours and weekdays of students' accesses. The navigation is possible with the use of calendar or with filters where we can select certain subject, group or student.



Figure 2: Exemplary outcome [6]

3.4. Project goal and outcome

The goal of my thesis was to perform the Data Mining part of a project "Learning Analytics" developed in inLab at Technical University of Catalonia. The tasks included introducing new indicators to an existing Learning Analytics project, perform the revision, validation and testing of the indicators. I was responsible for continuation of development of the Data Mining part which was started by previous graduate students. In my work I could base on their outputs which were mainly covering the theoretical part of the problem. My major task was to introduce practical ways to calculate indicators proposed by previous students. One of the biggest challenges was to asset whether calculated indicator fulfills expectations and brings valuable information about learning process. Finally, gathered indicators should result in notification report about student’s motivation. The output of my work should be given in such form that will be easy to present in web application.

3.5. Gantt chart

Month	Sep		Oct				Nov				Dec				Jan		
Week	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3
Installation	■																
DM online course		■															
Reading documentation			■														
Activity time & no. of breaks				■	■	■											
Curiosity							■										
Forum activity & delivery rate								■	■								
Starting Date										■	■	■					
Final Report															■	■	■

Figure 3: Gantt chart

I have started my work in the middle of September 2015. Firstly, I had to acquaint myself with all the tools and synchronize them. It involved installation of R, Pentaho, Database Browser, Xampp etc. At the beginning I have done an online course on EDX platform about Data Mining in Learning Analytics. It gave me some insight into the topic. Before starting my work on indicators I had to know what were the results of previous students working on the project. After this, I have started with implementation of the first indicator: Activity Time. Getting familiar with new tools took some time as well as improvement of initial approach. Later I moved to another indicators, like Curiosity and Forum activity. Very important was the work on Starting Date indicator. It required a lot of time and analysis to obtain finally satisfying results. At the end of December I had 2 weeks of break due to Christmas and New Year, after coming back to work, I started preparing the final report and presentation.

What is interesting in the Gantt chart, you can assume the difficulty of calculating certain indicators. In this project true is the dependence between time spent on the implementation and its difficulty.

3.6. Budget

Before starting work on a project, very important to plan the expenses and the duration time, so the project will be finished without any unexpected costs. In the first table you can see an expected cost of hardware and maintenance costs such as electricity.

Hardware Component	Total Cost €	Maintenance €
Personal Computers	1000	30
Servers	470	10
Total	1470	40

Table 1: Hardware costs

The second table contains estimated costs connected with licenses for used software. The majority of used software is open source and for free, however costs even of a few software licenses is pretty high, as you can see in the table below.

Software	Total Cost €
Microsoft Windows 8.1	135
Tableau Desktop	730
Microsoft Word	70
Total	935

Table 2: Software costs

The last but not least issue are the human resources. I have estimated costs of my work during 18 weeks and two professors of UPC with whom I was consulting the project. Summing up all the cost, the initial cost for the first half of the year can be rounded to about 14 500 euros.

Vacancy	Cost per hour €	Hours per week	Number of weeks	Costs €
Data Scientist (student)	15	40	18	10 800
Professor Expert LAx2	35	1	18	1 260
			Total	12 060

Table 3: Human Resources costs

4. Methodology

As a methodology for my project I have chosen Cross Industry Standard Process for Data Mining (CRISP-DM) [7], [8] which is one of the most well-known and wide used methodology for data mining. The aim of using a CRISP is to analyse data faster, more efficiently and in more reliable way.

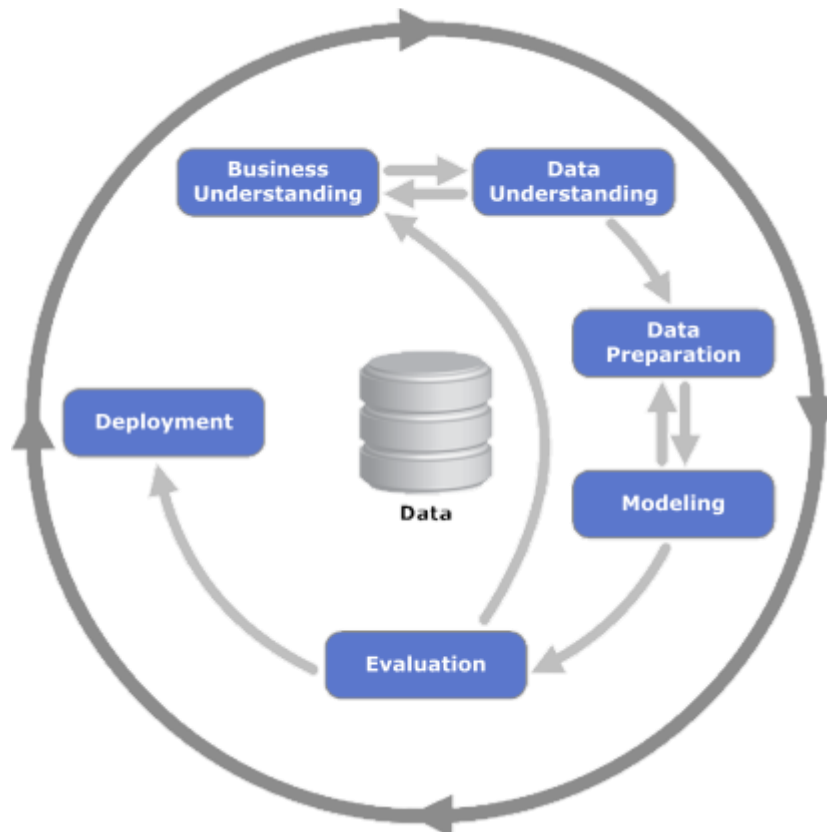


Figure 4: CRISP scheme [8]

First stage is problem understanding. What is your goal and why do we need some Data Mining to achieve it? What are the obstacles and what is our expected outcome? These are the basic questions which have to be answered at this stage.

In the case of our project, we have loaded data from Agora database which are extremely hard to analyse in the form as they are stored. Moreover, we would like to obtain some information about students' motivation basing on their activity on the platform which is stored in log table.

Second phase is data understanding which starts from gathering all data. However, in our project data from Moodle are already stored in relational database, so this was not a

problem. Important is to understand the structure of data, which was more challenging, since Moodle has more than 35 related tables.

Third stage is data preparation which is connected with its selection, cleaning or integration. In this project, I did this step with the use of SQL. It is important to select only relevant data, because it has huge impact on the speed of Data Mining process. In one example I observed a 30% improvement of execution speed only by making the SQL query more precise. In this project indicators are calculated independently and this step repeats every time some other indicator is being calculated.

Following step, modelling, is connected with selecting proper modelling technique. In our project tools for this part are Pentaho and R. Before designing the transformation I firstly thought about execution steps which I will have to implement. According to these, I was choosing proper Pentaho steps or writing R code.

Later follows evaluation of results, revision and also it is decided what should be the next step, moving to implementation or going back to problem understanding which is the beginning of a process to make some improvements. In this project I verified results in various ways, depending on the type of output. I checked with SQL some conditions which results should satisfy. I have also used visualization to observe obtained results and by inspection decide whether they are relevant or not.

The final step, implementation, is connected with putting the solution into production, make documentation from work and assure proper maintenance. This thesis can be treated as a documentation, because later I will in details explain everything I have done in this project.

I have chosen this methodology for my project because it is one of the most commonly used in Data Mining for the last 20 years [7]. Around 50% of Data Mining processes are performed with its use. Alternative to CRISP are usually personally adapted procedures. In my work I wanted to use some formal methodology to systematise my work and follow its guidelines. I came to a conclusion, that CRISP methodology is very intuitive schema of work and at some point I wasn't aware of the fact that I was following some CRISP rules.

5. Tools

5.1. Pentaho

Pentaho is an open source-based platform for diverse big data deployments [9].

During my work on the project I firstly used only computational steps provided by Pentaho. However, later I focused more on R which is also able to execute within Pentaho Kettle. The advantages of DM implemented with Pentaho steps are that even a person who doesn't know the project can easily understand how the data is manipulated step by step. Icons of steps give first information about operation, it can be for example: filter, group by, calculator etc. Seeing these icons we know the sequence and objective of change in the data. When we click inside the step we can exactly see what and how has been changed. Another positive thing about Pentaho and the transformation step-by-step is that you can see the output from each step just by one click, without going into debugging mode, you can see how the data has changed after certain action. The disadvantages of such solution is that sometimes for simple change we need to use a lot of steps. One example you will see below, while I had changing the format of date from UNIX timestamp to real date. The final transformation itself can be done very easily, however preparation took a lot of steps, like adding new constant, multiply timestamp by this constant (the conversion was done from milliseconds) etc. The problem with Pentaho arose when I had to introduce some iterations in my calculations. Then I used R executor, where you define the input data and write the R code. This solution seemed to be faster than steps, however Pentaho compiler wasn't very useful in detecting R errors.

5.2. SQL

SQL (Structured Query Language) is the standard language for management of relational database systems. SQL statements are used to perform operations on data such as: insertion, update, deletion or retrieval. The most commonly known SQL database management systems SQL are: Oracle, Microsoft SQL Server, Access, etc. In our project I was working with MySQL and for initial inspection of the data I used Database Browser. Extraction of data was the first step in each Pentaho transformation and like in the case of

R scripts, the messages about any errors in SQL from Pentaho weren't helpful and firstly I evaluated SQL queries in Database Browser.

5.3. R

R is a language for statistical computing. It is one of the most commonly used languages for Data Mining along with Python and SQL. R's roots come from S language [10]. R executor is implemented in Pentaho, however I primarily used also RStudio, an IDE for R, to obtain properly running code, because compiling errors given in Pentaho weren't very helpful. It was my first contact with R, I expected it to be similar to Matlab, however there are some basic concepts in R which I don't know from other programming languages.

5.4. Tableau

Tableau is a software for visualizing the data from relational databases. It is very easy to use. You only have to connect with the desired database and select the table you wish to visualize. To put data on the graph you just need to drag and drop selected columns. You can adjust the type of graph, colours, get sum, maximum or minimum of a field. In our project the final results are presented in web application using Highcharts, which is a JavaScript library. However, for my needs Tableau is better solution, because I see the results very fast, just by refreshing the page and I can focus on various columns to inspect the reliability of the data. Some indicators I was able to validate by inspection of the output table, however majority of calculations needed graphical representation to evaluate their accuracy. Using Tableau is very intuitive, however I needed some time to obtain the desired plot. It happened to me that I wanted to put two variables on one axis and I did so, but the output wasn't correct. It turned out that each axis had different scaling and the results weren't synchronised.

6. Indicators

The indicators I am going to present below and their calculations are basing on the theoretical investigation of the problem of indicators for learning analytics done by Ivan Vukić presented in his Master's Thesis: "Measurement of motivation of high school

students for real-time tracking from the Virtual learning environment” defended in June 2014 at Technical University of Catalonia. Because indicators are calculated with data taken from Moodle I want to mention some key words used in this database and my calculations:

- Userid: is an unique number for each student, with which his/her name and surname can be obtained
- Course/Courseid: id of a subject
- Cmid (course-module id): is an id of certain task (module). The dependence is such that one subject has many tasks for example course mathematics can have modules such as: integrals quiz, area of a cube assessment etc.

	course	cmid	userid	time	action
1	2863023	2860	28668	1410378766	assign
2	2863023	2860	28668	1410378767	view
3	2863023	2860	28668	1410378775	view
4	2863023	2860	28668	1410378775	view
5	2863023	2860	28668	1410378936	editsection
6	2863023	2860	28668	1410378936	view
7	2863023	2860	28668	1410378980	editsection
8	2863023	2860	28668	1410378980	view
9	2863023	2860	28668	1410379033	editsection
10	2863023	2860	28668	1410379034	view

Figure 5: Exemplary output from log table

In the presented above table you can see how the log table looks like. In the given example logs are for the same course, module and user. We can see the actions of the user and time of his/her work. Most probably this is a teacher who is editing some module.

6.1. Speed

Speed indicators are giving us information about time in general. It can be information about first access, time spent on task etc. Having such information we can analyse the way of student’s work and its duration.

Activity time indicator gives information about total time spent on a certain task. Thanks to this information we can observe if student is dedicated to his/her work. Having information about time which would be longer than expected, we can say that student has problems in solving the task, needs some extra classes or more teacher’s attention. On the other hand, if the solving time was so short, a student might be either bored or the course is too easy for

him/her and teacher needs to propose other activities. Activity time can be calculated for tasks which are filled online, like quizzes or hotpots (other type of quizzes, not relevant in understanding the case). Activity cannot be calculated for resources like links or books, because we don't have information how long student has worked with this resource.

Another indicator is Starting Date. In the Moodle teachers can specify the time in which task will be opened and should be submitted. Most commonly this option is used in the case of obligatory quizzes. However, there are often no dates within which student can access some task. So theoretically, it can be put at the beginning of the semester and wait several months until the moment when a teacher presents certain topic and gives homework. The aim of calculating Starting Date is to obtain a date which is most probably date of giving homework basing on number of accesses for all tasks. In the end the whole ETL process will be run in our application every 24 hours to make updates according to changes given in the last time.

6.2. Intensity

Intensity indicators provide us with information about measurable results of a student. It can be either number of submitted assessments, quizzes etc. In this indicator we concentrate in measurable efficiency of student's work which may be presented with number of solved tasks or number of written posts on forum.

First indicator is Delivery Rate which is quote of finished tasks to unfinished ones. This indicator is designed for obligatory tasks. By obligatory tasks are meant quizzes, hotpots (other kind of quiz) and assignment. These modules need students' interaction, either solving the module online in the case of quiz or uploading solution for assignment. For these modules it is possible to put the deadline which makes them obligatory to make.

$$\frac{finishedTasks}{allTasks}$$

The second intensity indicator is Forum activity Rate. Motivated student is active, in case of doubts he/she asks about his/her problems and helps others in need. The desire to exchange opinions and problems is a very important part of learning process. In this

indicator I am monitoring the amount of comments, which give information about involvement and willingness to deeply understand the topic or helping his/her peers in it. Also important is the number of entrees, because student may not take active part in discussion but follow it on regular basis. These two measures are not equal and more important are the comments, thus there are parameters which will differentiate them.

$$\alpha \cdot \text{comments} + \beta \cdot \text{entires}$$

6.3. Persistence

Persistence in Learning Analytics puts focus on continuity of a learning process. It is important is to know whether student is able to focus for a longer time or get easily bored or distracted.

In this indicator I am calculating number of breaks while solving a task. With this indicator we can see that student has worked on some quiz in total for 40 minutes but also had 4 breaks, when his/her peers solved the same task in one go.

6.4. Choice

Indicator of Choice which is Curiosity Rate should give us information about student's interests apart from obligatory tasks. Teachers give more quizzes to do or some additional books or links to learn from. Knowing the tasks and resources that student accesses on his/her own, we observe his/her preferences and stimulate his/her work with additional materials.

$$\frac{\alpha \cdot \text{fulfilled} + \beta \cdot \text{accessed}}{\text{allActivities}}$$

7. Results

7.1. Activity time and number of breaks

In the previous chapter I have presented theoretical approach to the indicators. In this section I will focus on the practical analysis and description of my solution. Importance of activity time indicator can be given by exemplary student Mark who is obliged to solve quiz after every biology class. Mark is always very focused and solves quizzes relatively fast in one go. However, at some point Mark started to have some problems and was getting distracted very easily. Sometimes happened that he started a quiz but after one minute he switched the tab and started chatting with his friends or watching some videos. Sometimes his break was so long that he wasn't able to finish the quiz the same day because he had other obligations to make. Another time he was unable to solve the quiz and all the time he had to check previous sections of the educational platform to search for the answer, he was constantly jumping in and out from the quiz. Thanks to indicator which is presented in this section, Mark's teacher is able to spot the difference in his performance and find out what causes such change in Mark's learning process.

7.1.1. First approach: from beginning till the end

In the below presented figure you can see the Pentaho transformation of this indicator. After first step which is getting all data from log table follows mapping of action names. It means that instead of actions described by words I gave them numbers. It means that for example "attempt" has number one, "continue attempt" number 2 etc. Thanks to such mapping later the R script executes much easier. The next phase is the conversion of number type from long to double. I encountered this error in data format while executing the R script and to prevent from it I added this step. After R script which will be later described in details, follows filtering of empty rows. The output of the R script is a joined table which may have some empty rows. These results are not complete, thus irrelevant in our calculation and taken out from the results table. The last step is summation of time and attempts for each date. Finally, obtained table is saved in the database.



Figure 6: Activity scheme

I started calculating Activity Time from accessing the database. The main table we are interested in is the log table. However, to take only those log rows which indicate operations of quizzes and hotpots, I had to join two extra tables to be able to write “where” clause.

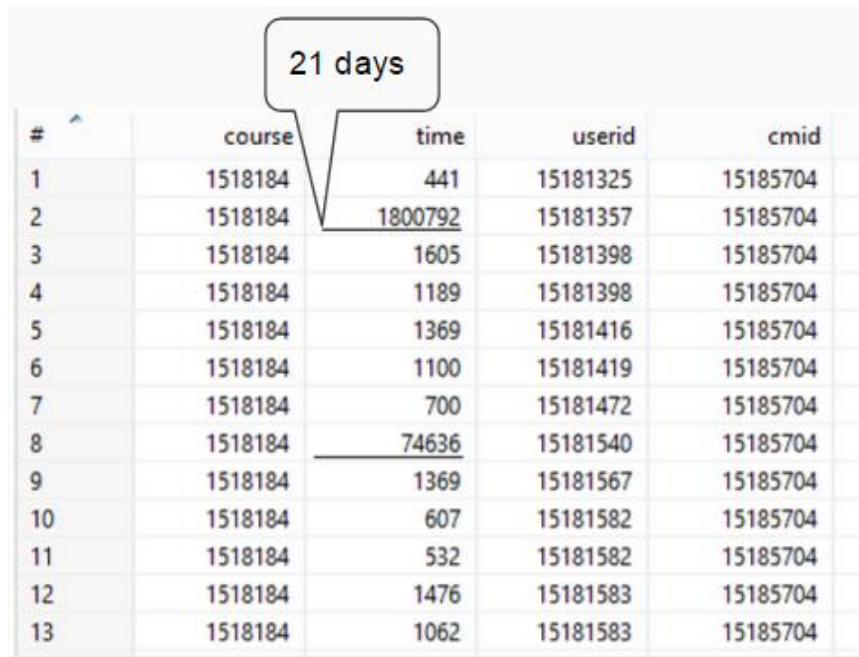
```
SELECT
lo.*,
CONCAT('mdl_',mo.name) as `table`
FROM
mdl_log lo
INNER JOIN mdl_course_modules cm ON lo.cmid = cm.id
INNER JOIN mdl_modules mo ON cm.module = mo.id
WHERE mo.name='quiz' OR mo.name='hotpot'
ORDER BY `time`, userid
```

Source Code 1: SQL code in log table

The biggest challenge in calculating the activity time was deciding how to find the beginning and the end of an activity. For this purpose I had to analyse the log tables and the imagine possible way of using the platform. The most important thing was to understand the meaning of action types, to know his way of using the platform:

- attempt: indicates start of work on task
- continue attempt: indicated continuation of work
- close attempt: tells about closing a task.

The problem in the logs, is that we only know time of an action, so if student has abandoned the task, in the log file we won't see it. In the first version of the algorithm I calculated time between action "attempt" and "close attempt", so the time distance between opening and closing task. The results however were only partially satisfying, because some of time slots were incredibly long. You can see them in the Figure 6 depicted below. The majority of calculated times (given in seconds) doesn't exceed 30 minutes, however in the presented table there are at least two results which don't match. The result form the second row has value of 21 days, which surely is not possible to happen in real life.



#	course	time	userid	cmid
1	1518184	441	15181325	15185704
2	1518184	1800792	15181357	15185704
3	1518184	1605	15181398	15185704
4	1518184	1189	15181398	15185704
5	1518184	1369	15181416	15185704
6	1518184	1100	15181419	15185704
7	1518184	700	15181472	15185704
8	1518184	74636	15181540	15185704
9	1518184	1369	15181567	15185704
10	1518184	607	15181582	15185704
11	1518184	532	15181582	15185704
12	1518184	1476	15181583	15185704
13	1518184	1062	15181583	15185704

Figure 7: Activity time in seconds

7.1.2. Second approach: interrupted time slots

The new approach was based on a concept of time slots between action "continue attempt" and any other action which interrupted the student. Sum of these time slots gave an exact time spent on a task. I left the concept of using "attempt" as initial activity because "attempt" was always followed by "continue attempt". That is why in the analysis I have focused only on the "continue attempt" action. However, still the problem of abandoned tasks wasn't solved. It was a problem which couldn't be solved with data which were given. There were two possibilities, either omit such cases and not have these entries registered or make an assumption that one session has to be limited in time. I have chosen the second option and decided that when between "continue attempt" and the next action is more than 2400 sec. (40 minutes) the final time for this slot will be also 2400 seconds.

In the R script presented below I am iterating over the logs and looking for an action "attempt" followed by "continue attempt" or only "continue attempt" for the same user. If I find one, the time of the following row for the same user and task will be the final time of a working session. Every such pair I subtract and obtain a single time slot within which student has worked. As I have mentioned before, if this single time slot exceeds 40 minutes, automatically time of 40 minutes is being set.

```
numMods <- nrow(values)

mdat <- matrix( nrow = numMods , ncol = 4, byrow = TRUE)
colnames(mdat) <- c("course","cmid", "userid", "time")
for (i in 1:numMods-2) {

  if (isTRUE(all.equal(values$userid[i], values$userid[i+1])) &&
(isTRUE(all.equal(as.numeric(as.character(values$action[i])), 2)) ||
isTRUE(all.equal(as.numeric(as.character(values$action[i])), 1))) &&
isTRUE(all.equal(values$cmid[i], values$cmid[i+1])) )
  {
    if(isTRUE(all.equal(as.numeric(as.character(values$action[i+1])), 2)))
    {i<- i+1}
    mdat[i,1] <- values$course[i]
    mdat[i,2] <- values$cmid[i]
    mdat[i,3] <- values$userid[i]
    mdat[i,4] <- values$time[i+1]- values$time[i]
    if (mdat[i,4] >2400)
      mdat[i,4] <-2400
  }
}
mdat <- data.frame(mdat)
mdat

# 1 attempt
# 2 continue attempt
# 3 close attempt
# 4 view
# 5 review
# 6 view summary
# 7 preview
# 8 update
# 9 report
# 10 editquestions
# 11 manualgrade
```

Source Code 2: R Script for Activity Time

The output obtained from such calculations is given in the table, where we know user, course and task, total activity time for a given task and in how many attempts the work was done. The number of attempts can be also considered as number of breaks (number of attempts -1)

course	cmid	Userid	timetotal	attempts
399502	3993523	3991004	559	1
1518184	15185708	151827	2400	1
1518184	15185707	151827	8	1
1518184	15185704	151827	87	1
1518184	15185705	151827	305	5
1518184	15185742	151827	2400	1
1518184	15185761	151827	536	3
1518184	15185782	151827	2400	1
1518184	151812001	151827	35	1
1518184	151812002	151827	2400	1
1518184	151811981	151827	60	1
1518184	151812142	151827	128	1
486922	48678834	48641	2400	1
1518184	151812001	151827	2400	1
1152204	11527065	115221	17	1
1152204	11527066	115221	16	1
1152204	11527065	115221	2400	1

Table 4: Activity Time and number of attempts

The results from the table are presented in the Figure 8. On the axis we have students and the modules they solved and on the other axis total time spent on each module and number of attempts which lead to solving a task. Worth mentioning is fact that amount of lines depicting time and attempts for given student and module are equal.

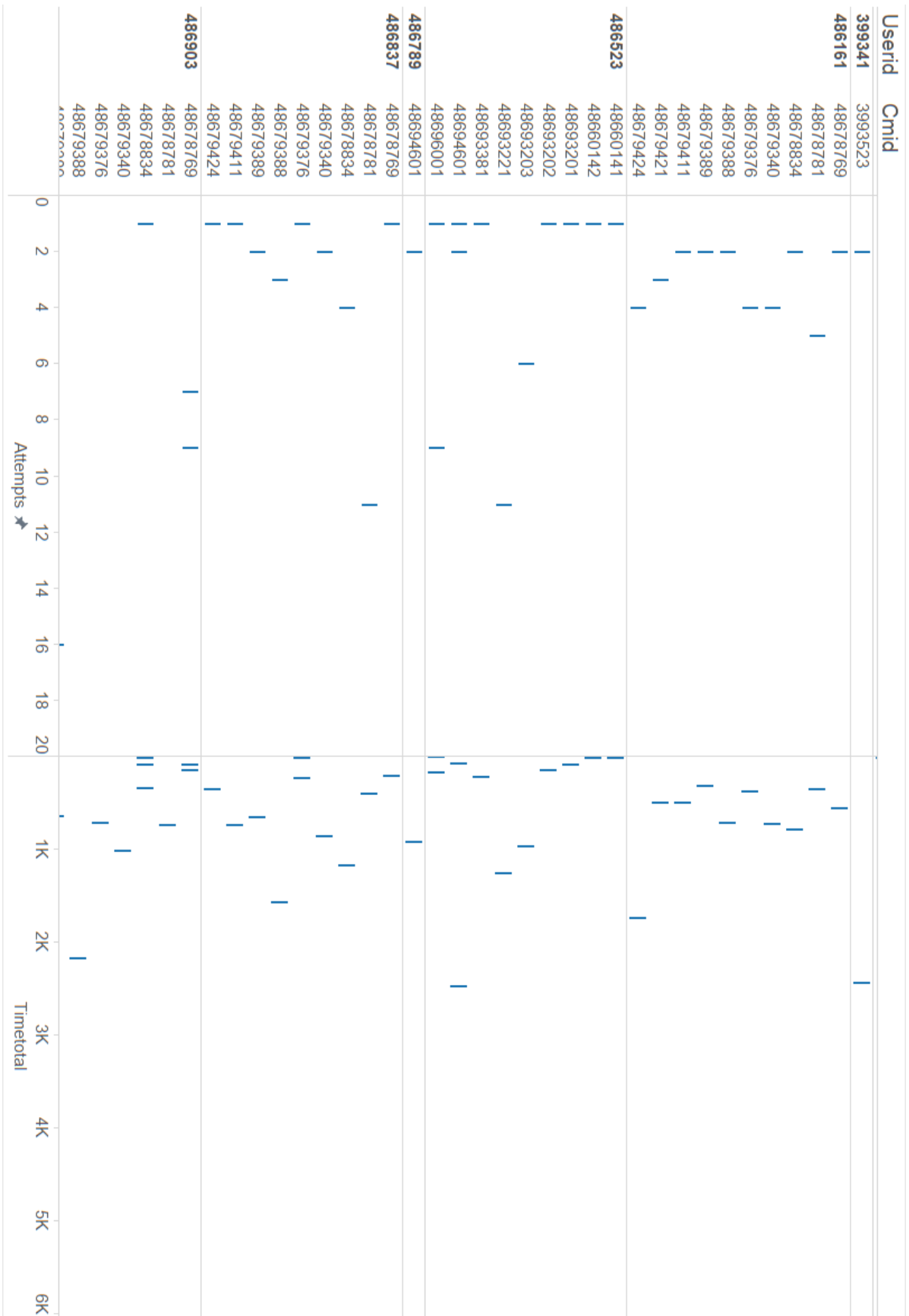


Figure 8: Activity time and number of attempts

7.1.3. Justification of selected approach

As you can see even looking at the results, better approach is the second one. In this solution I take into consideration possibility of interrupted work. I also solved the problem of abandoned task, which leads to reliable results which can give teacher important feedback about way of work of his/her students.

7.2. Starting Date

In the previous chapter I have defined the principle of Starting Date indicator. In this section I will present the way of its calculation.

Imagine a situation when teacher Tom has uploaded quiz for his class at the beginning of semester, in September. The scheduled time for the lesson for which the quiz was prepared in December. In the meantime some curious students have entered the quiz to check what is it about. In December Tom would like to monitor actions of his students but he forgot to put the dates within which it was scheduled. In classical set of registered accesses Tom would get a lot of irrelevant data caused by students who entered the quiz before December. It would be also hard for Tom to estimate himself the Starting Date because he is teaching different groups, where number of students varies. Sometimes there is a class of 10 students where 6 is a majority and if they access the platform in such number it may not be a coincidence. On the other hand, Tom has also groups of 36 students where 6 pupils are only small part of the class and their registered access might be just a mistake. The aim of this indicator is to eliminate such situations and make it possible to set the Starting Date only basing on students accesses.

Figure 9 shows the scheme of Pentaho transformation which isn't very complicated because the majority of transformation is performed within R code. The first step is the database input, followed by R script, third step is the conversion from UNIX timestamp to real date and finally writing down the results into database.



Figure 9: Starting Date scheme

7.2.1. First approach: quantiles of accesses

The first approach of calculating Starting Date was basing on total number and quantiles of accesses. At first I was looking for a time difference between the quantile of 25% and 50% of accesses (depicted as dq), later I moved this tripled time difference and subtracted seven hours. Obtained point was the desired Starting Date.

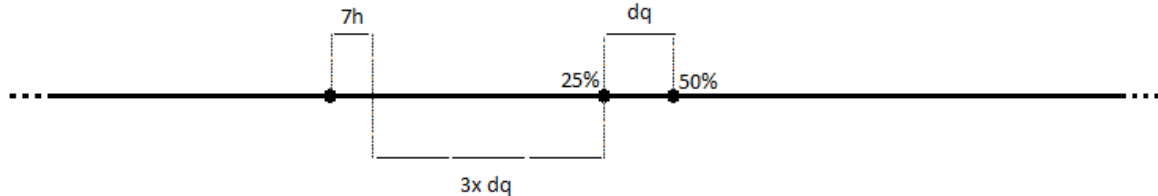


Figure 10: Scheme of calculation Starting Date

```

minTimes <- aggregate(x = input, by = list(input$cmid, input$userid), FUN =
"min")
quantile25 <- with (minTimes, tapply(time, cmid, quantile, probs=0.25))
quantile50 <- with (minTimes, tapply(time, cmid, quantile, probs=0.5))
quantiles <- quantile50 - quantile25
numMods=nrow(quantiles)
mdat <- matrix(nrow = numMods, ncol = 2, byrow = TRUE)
colnames(mdat) <- c("cmid", "dateToStart")

for (i in 1:numMods) {
  dateToStart <- quantile25[i] - 3*quantiles[i]
  mdat[i,1] <- minTimes$cmid[i]
  mdat[i,2] <- dateToStart
}

f <- function(time) {time= min(time)}
merged <- merge(mdat, minTimes,by="cmid")
filtered <- merged[merged$time>merged$dateToStart, ]
start3q <- aggregate(x = filtered, by = list(cmid=filtered$cmid), FUN ="f")

start3q$dateToStart <- start3q$time - 25200

minTimes$Group.1 <- NULL
minTimes$Group.2 <- NULL
start3q$Group.2 <- NULL
start3q$Group.1 <- NULL
start3q$userid <- NULL
start3q$cmid <- NULL
start3q$time <- NULL
start3q$userid <- NULL
start3q$course <- NULL

final <- merge(start3q, input,by=c("cmid"))
filteredBeforeStart <- final[final$time>final$dateToStart, ]

newdat <- data.frame(final)
newdat

```

Source Code 3: R Code for Starting Time

After analysis of results shown in the graphic, I came to conclusion that obtained outcome is not satisfying, because there are several modules in which Starting Date was calculated before the expected moment, for example in the holidays period, in July or August, which you can see in the figure below (Figure 11). The entrees are depicted with blue lines and calculated Starting Date with green dots. The problem with calculating the indicator this way was that I was basing on the log table collected within several months. It means that we were analysing logs from closed semester. Only when we had final number of entries we could calculate quantiles of 25% or 50%. Another issue is that we never knew the actual percentage of students which actually accessed the platform.

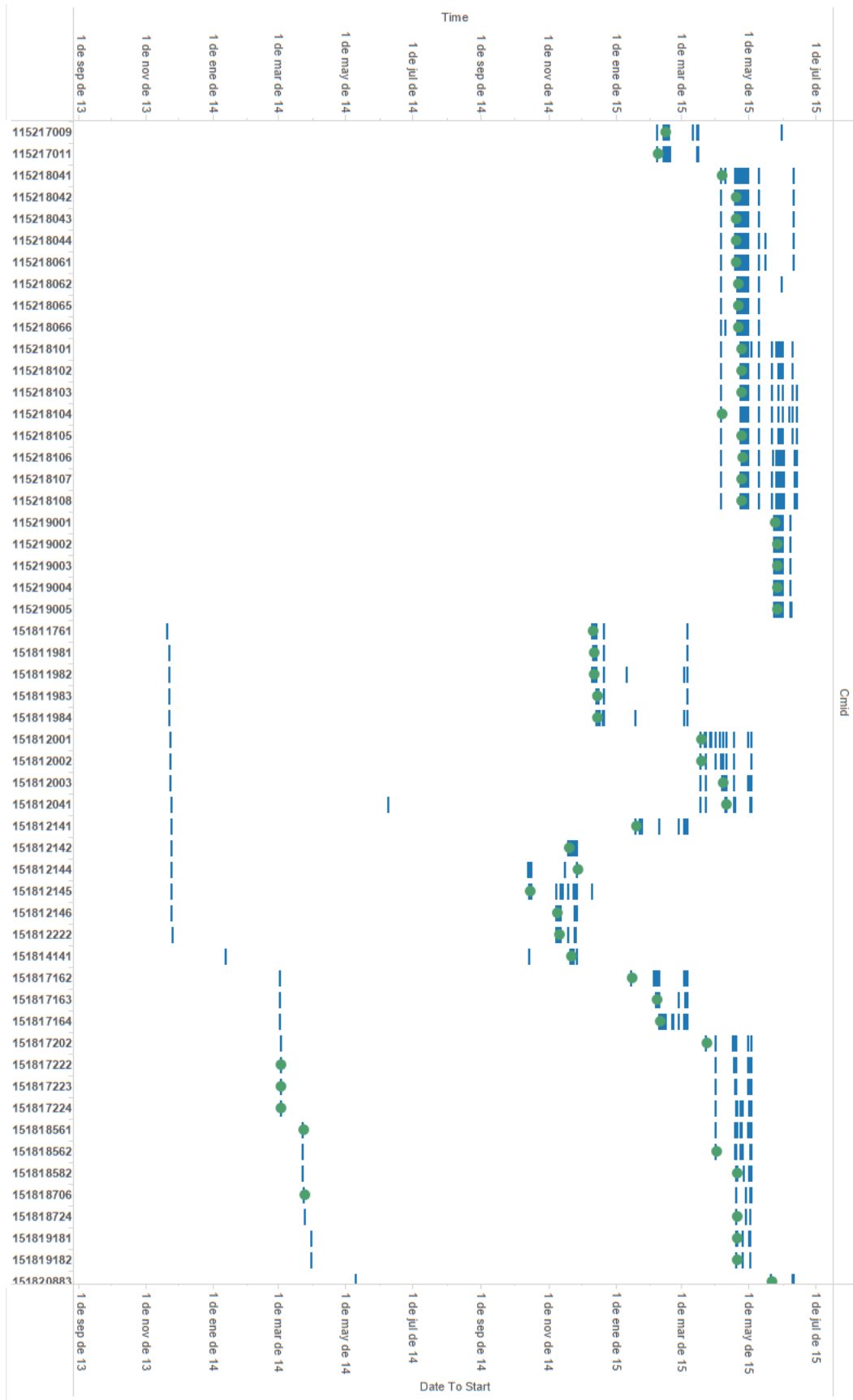


Figure 11: Output of Starting Date

7.2.2. Second approach: percentage of students

The new approach of calculating Starting Date had to base on some constant value known at the beginning of the semester. The solution is the number of students subscribed for a certain course. Usually it is around 30 people. The threshold for setting the Starting Date is given as 10% of students in the course which access the platform within 7 days.

Presented SQL code is more complicated than the previous one because I had make a nested select, to firstly take whole list of students, courses and modules and from such list to count numbers of students for each course and in the final step join it with the log table to have all entry times.

```
SELECT lo.time, lo.userid, T2.*
FROM (SELECT
DISTINCT count(*) as studentNumber, courseid , cmid
FROM ( SELECT u.id , co.id as courseid, mo.id as cmid
FROM
mdl_course_modules mo
join mdl_course co ON mo.course=co.id
join mdl_context con ON con.instanceid = co.id
join mdl_role_assignments ra ON ra.contextid = con.id
join mdl_user u ON u.id = ra.userid
join mdl_role r ON ra.roleid = r.id

WHERE
con.contextlevel = 50 AND
r.shortname = 'student'

GROUP BY co.id, mo.module, u.id ) AS T
GROUP BY courseid, cmid ) AS T2

JOIN mdl_log lo ON lo.cmid = T2.cmid
```

Source Code 4: SQL code for final approach of Starting Date

In the R code I firstly calculated the 10% of a given students number, and then given a certain date, which can be real date of today. I mocked it to obtain expected results because I was working on data from the last year. As the next step I filtered all rows which indicate previous date earlier than seven days before given date. As a next step I counted the number of modules and started iteration over them to find their real Starting Date. The condition which has to be fulfilled is that number of rows for a given module is greater or equal to calculated 10% of students. If this condition is satisfied it means that Starting Date exists. If it is not the case, there is no Starting Date for this module. Also then 10% of students haven't yet entered the platform, the Starting Date is not calculated. As a final Starting Date we take the initially considered date minus 7 days and 7 hours.

```

miraLogs <-

function (i_logs) {
  i<- 1
  temp <- i + i_logs[i,"percent"];
  rnum<-nrow(i_logs);

  if(rnum>=i_logs[1,"percent"])
    newrow<- c(i_logs[i,"time"],i_logs[i,"cmid"],i_logs[i,"courseid"])
  else newrow<- c(i_logs[i,1],0,0)
  return (newrow)
}

## Add a Percent column with 10% value

# currentData<- as.numeric(as.POSIXct(sys.date())) # take current date and
conver it into timestamp
givenData <- 1412035200
logs <- input[input$time>=(givenData-604800), ]
logs$percent <- ceiling (logs$studentNumber * 0.1)
df<-data.frame(startTime = numeric(0),cmid= numeric(0) )
args<-unique(logs$cmid)
if(length(args) != 0) {
  len <-length (args)
  for(k in 1:len)
  df[k,] <-miraLogs(logs[logs$cmid==args[[k]],])
}
df$startTime <- (df$startTime-25200-604800) *1000
final <- data.frame(final)
final

```

Source Code 5: R code for final approach of Starting Date

In the given example of the output (figure below) on one axis are the modules for which Starting Date has been calculated and on the other axis is a timeline where entrees and calculated Starting Date is presented. In the Figure 12 the initial date was the 30th September 2014 and visible are only the courses for which Starting Date has been calculated. The final Starting Date is set to 23rd of September, so 7 days before the initial date within which 10% of students have accesses the platform. Most probably for some of the given cases, the Starting Date could be calculated earlier since there is big amount of entrees also before the Starting Date. Such cases will be omitted in the real running of the system, because it will be run every 24 hours (like the arrow shows) and Starting Dates will be calculated every day only for those modules which didn't have their Starting Date calculated earlier.

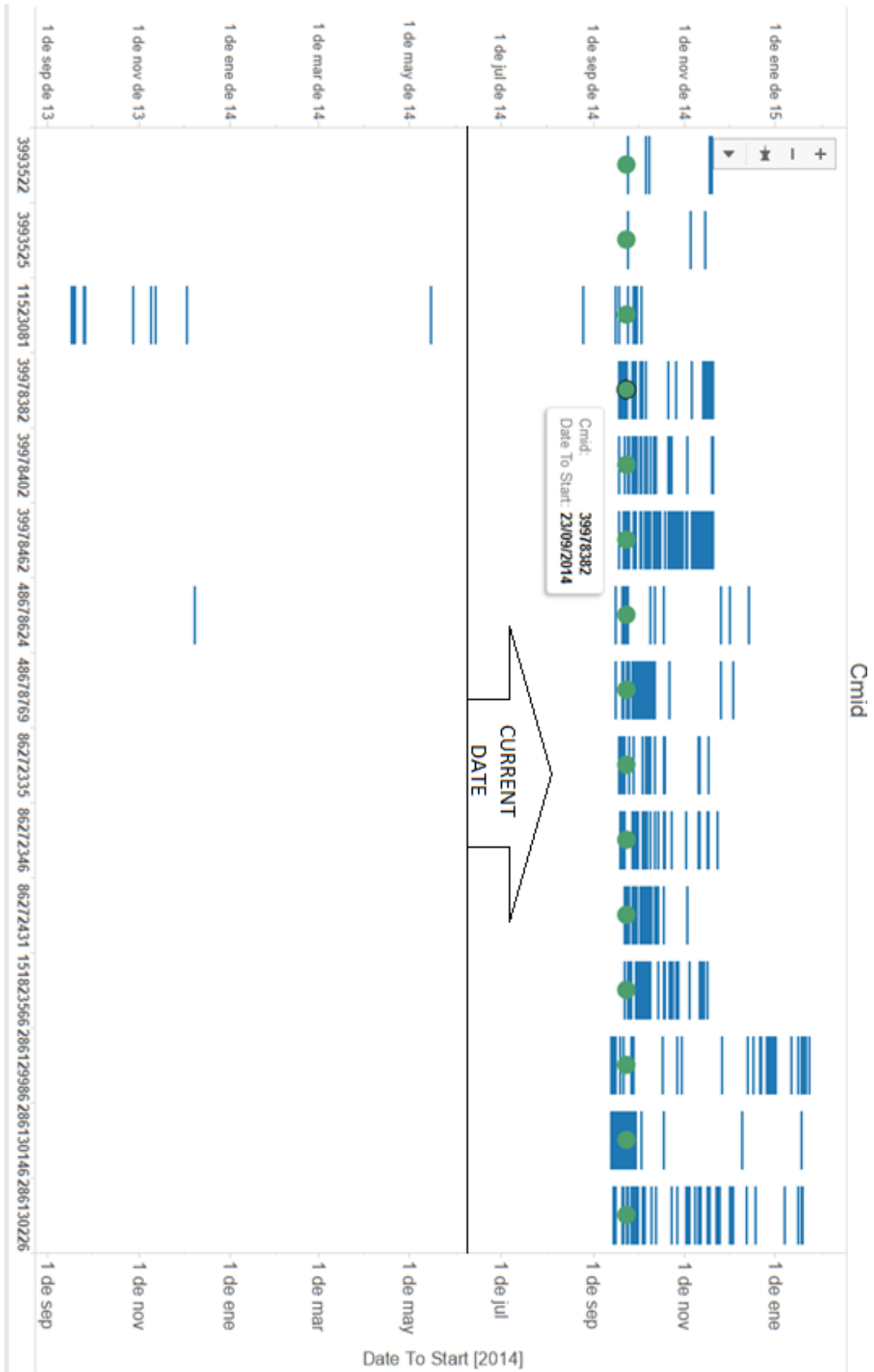


Figure 12: Final output of Starting Date

In the figure below I have presented another output but with initial date 18th of September. As you can see in the plot, some modules which are satisfying the condition have been found the Starting Date is given as 11th of September. The modules marked in a box are the same ones as in the previous figure. Normally in the running application, those modules which have their Starting Date found earlier wouldn't be taken into consideration while calculating the Starting Date for the following days.

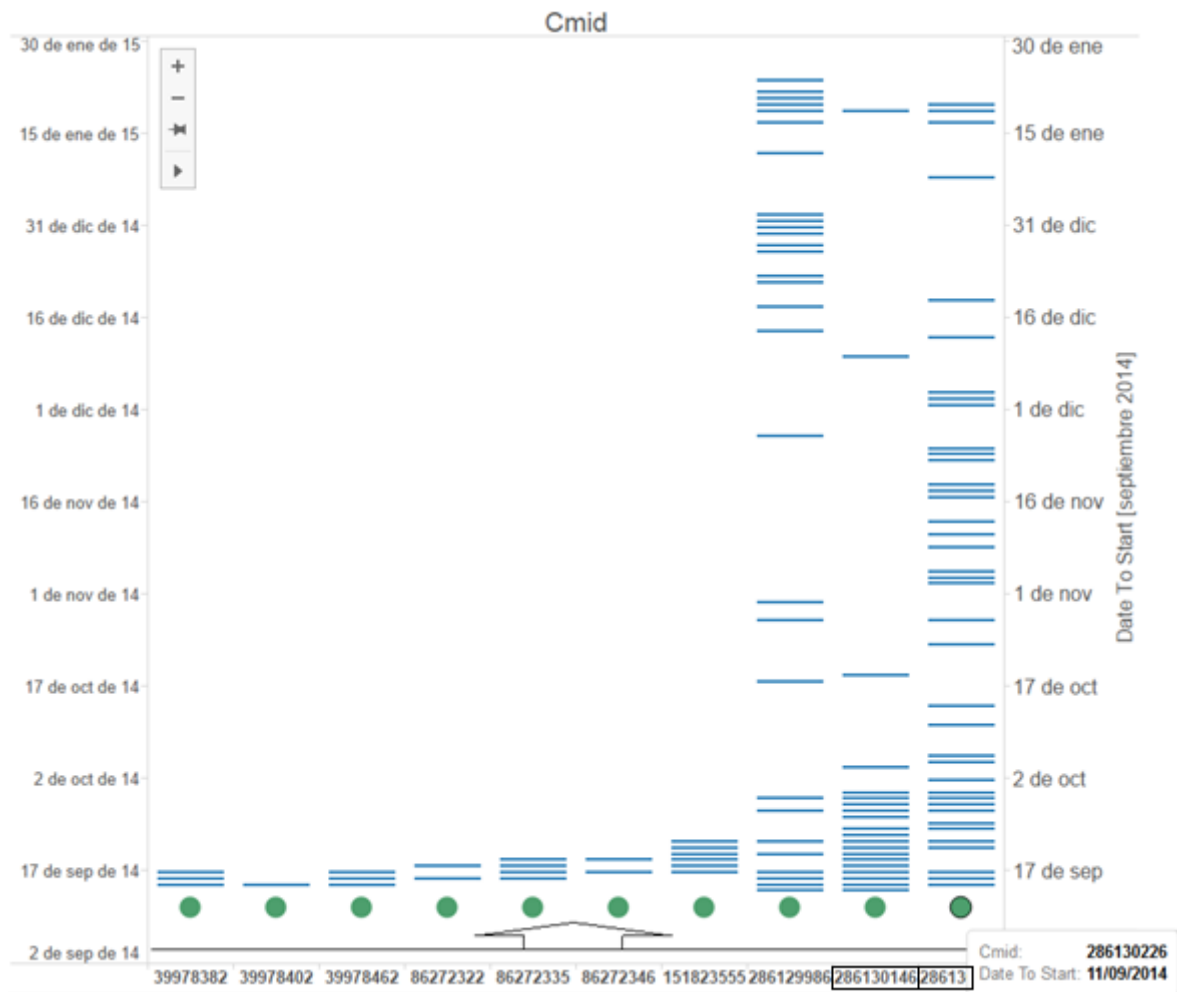


Figure 13: Final output of Starting Date

7.2.3. Justification of selected approach

The second approach is much more reliable as the first one. Also the second approach fulfills the idea of measuring motivation which should be continuous and assessed within most possibly short time. With the second approach we are able to obtain the Starting Date even after one day, only when corresponding amount of students have accessed the

module. Basing on the constant number of students in the group is much better solution than analyzing constantly changing registers of entrees.

7.3. Curiosity rate

Curiosity rate give us information about students' choices concerning their individual work. For example, teacher Tom is putting a lot of materials on Moodle, many of them are not obligatory but he hopes his students are interested in his subject and they broaden they knowledge using the materials he uploaded. Tom would like to know whether students are using the resources, who are these students, what type of additional modules do they like the most and what could he do to improve his teaching methods.

Curiosity was one of the first indicators I have calculated, because it is all obtained only with Pentaho steps. With SQL code I accessed log table and names of modules with matching actions which for these modules are significant. The main task for this indicator was the transformation of date format from UNIX date stamp to real date. Firstly I had to add a constant, in the next step perform the multiplication of UNIX datestamp by this constant. It was essential to make conversion for next step which works only for milliseconds. As following I removed time from date because it wasn't relevant in the output and in the final step I removed intermediate steps which aren't needed for the final result. In the end I saved obtained table in the database.

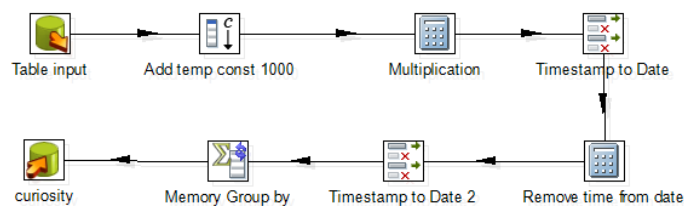


Figure 14: Curiosity scheme

SQL query used for extracting data is similar to the previous ones. Needed is access to the log table for not obligatory resources. The important is only the information if resource has been entered, thus module names are also matched with the action names which mean that they have been opened, for example: assign-submit, book-view chapter, etc.

```

lo.course, lo.cmid, lo.userid, lo.time, lo.action, mo.name as 'moduleName'
FROM
mdl_log lo
INNER JOIN mdl_course_modules cm ON lo.cmid = cm.id
INNER JOIN mdl_modules mo ON cm.module = mo.id
WHERE (mo.name="assign" and lo.action="submit") OR (mo.name="book" and
lo.action="view chapter")
OR (mo.name="scorm" and lo.action="launch") OR (mo.name="geogebra" and
lo.action="result") OR (mo.name="jcllic" and lo.action="view")

```

Source Code 6: SQL code for Curiosity indicator

The output for curiosity is presented in the table. We can see when certain module has been accessed. In the Table 5 we have information about course, type of module and time when it has been accessed.

course	userid	time	moduleName
15181242	15181462	28/01/2015	assign
486922	4861602	04/03/2015	assign
1152204	115221	08/01/2014	scorm
862544	862304	03/12/2014	assign
15181242	15181386	20/10/2014	assign
2863023	286540	18/11/2014	jcllic
15181242	15181412	05/02/2015	assign
1152204	1152437	14/01/2015	scorm
862544	862336	06/05/2015	assign
862404	862863	16/10/2014	assign
862544	862261	28/11/2014	book
1152204	1152466	28/10/2014	scorm
1152204	1152375	19/10/2014	scorm
862544	862296	29/09/2014	assign
1152204	1152463	09/12/2014	scorm
1152204	1152437	29/04/2015	scorm
15181242	15183134	19/11/2014	assign

Table 5: Curiosity table

In the figure below (Figure 16) I have presented the quantity of entrees because it is different type of representation, giving valuable information but also different than already presented forum activity. Of course settings of visualization do not influence the output data. On one axis are information about course and type of module and on the other axis the amount of entrees. Possible are also other types of representing the data, it could be presented on a timeline but similar way of presenting data will be given in the next chapter.

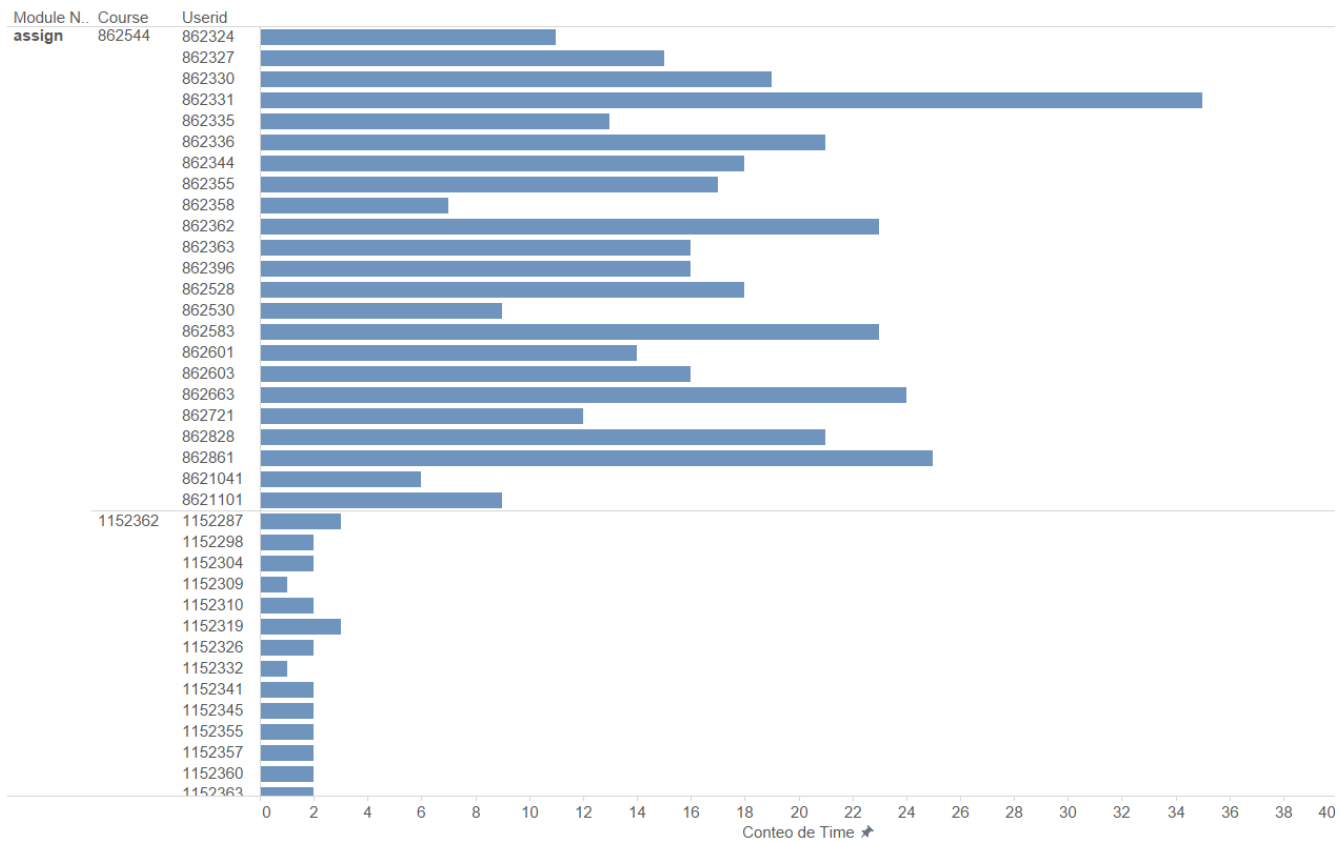


Figure 15: Curiosity output

7.4. Forum activity

Moodle platform gives possibility to exchange opinions on forum. We can imagine teacher Tom who is willing to know whether his students are discussing given problems, do they take active part in exchanging ideas or only observe moves of their colleagues. Maybe some students are too shy to present their doubts in public during classes and prefer to ask somebody by writing a post on forum.

As a first step in the transformation after accessing the database I made a date transformation from timestamp to real time just as in the previous indicator of curiosity. I added constant of 1000, multiplied timestamp by it and used transformation step provided by Pentaho which converts timestamp in milliseconds to real date. I also removed time from the date because it wasn't necessary and would complicate further calculation and grouping according to date. Later there is a set of filters where I collect logs from forums where users have either viewed the discussion or have taken part in it. Finally I saved obtained two tables in the database as shown in Figure 17.

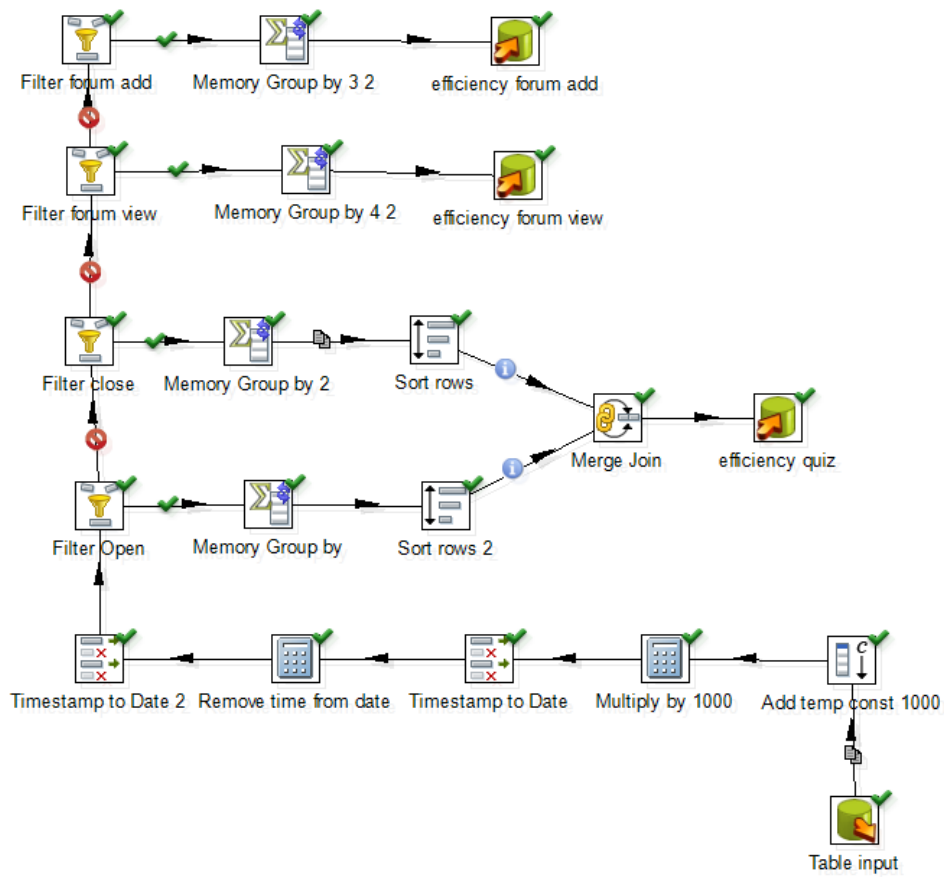


Figure 16: Efficiency scheme

SQL query used for this indicator is similar to the ones used previously. The module types we are now interested in are quizzes and hotpots (type of quiz). I only extracted log rows with activity names relevant for this indicator, such as: attempt, close attempt, submit (for hotpot) and add and view discussion (for forum).

```

SELECT distinct
lo.*, quiz.timeopen, quiz.timeclose, quiz.name
FROM
mdl_log lo
INNER JOIN mdl_course_modules cm ON lo.cmid = cm.id
INNER JOIN mdl_modules mo ON cm.module = mo.id
LEFT JOIN mdl_quiz quiz ON cm.course=quiz.course
where
( (mo.name='quiz' OR mo.name='hotpot' ) AND (lo.action='close attempt' OR
lo.action='submit' OR lo.action='attempt') AND quiz.timeopen!=0) OR
(mo.name='forum' AND (lo.action='view discussion' OR lo.action='add' ))

ORDER BY lo.course, lo.userid ASC

```

Source Code 7: SQL code for forum activity

course	userid	date	forum_add
1152362	115250	18/07/2014	2
2863023	28668	14/09/2014	1
2863023	28668	15/09/2014	1
3993422	39913442	15/09/2014	1
2863023	28668	22/09/2014	1
2863023	28668	25/09/2014	1
2863023	28668	13/10/2014	1
862544	862109	27/10/2014	1
2863023	28668	01/12/2014	1
2863023	28668	09/12/2014	1
2863023	28668	21/12/2014	1
1152362	115250	30/12/2014	1
2864003	28650	19/01/2015	1
2864003	28650	01/02/2015	1
2863023	28668	03/02/2015	1
2863023	28668	09/03/2015	1
2864003	28650	09/03/2015	1
2863023	28668	18/03/2015	2

Table 6: Added posts on forum

course	userid	date	forum_view
2863023	28668	14/09/2014	3
2863023	28668	15/09/2014	1
2863023	286462	15/09/2014	1
2863023	286481	15/09/2014	1
2863023	286529	15/09/2014	1
2863023	286532	15/09/2014	2
2863023	286533	15/09/2014	1
2863023	286540	15/09/2014	1
2863023	286543	15/09/2014	1
2863023	286548	15/09/2014	1
486922	48641	16/09/2014	1
486922	4864441	16/09/2014	1
2863023	28668	16/09/2014	2
2863023	286465	16/09/2014	1
2863023	286475	16/09/2014	1
2863023	286477	16/09/2014	2
2863023	286485	16/09/2014	1
2863023	286533	16/09/2014	1

Table 7: Views on forum

The output of this transformation is presented in two tables (6 and 7) presented above and also in the Figure 18 below. The visualization is presented in the time scale on one axis and modules and users on the other axis. We can see the entrees of students into forum, which are depicted as blue circles on the right part of the figure. On the left side are presented numbers of posts on the forum as red circles. In the given example we can see that actually only one person was posting on the forum and did it very frequently. Other users, specially these in the bottom part of the visualization have frequently checked the discussion.

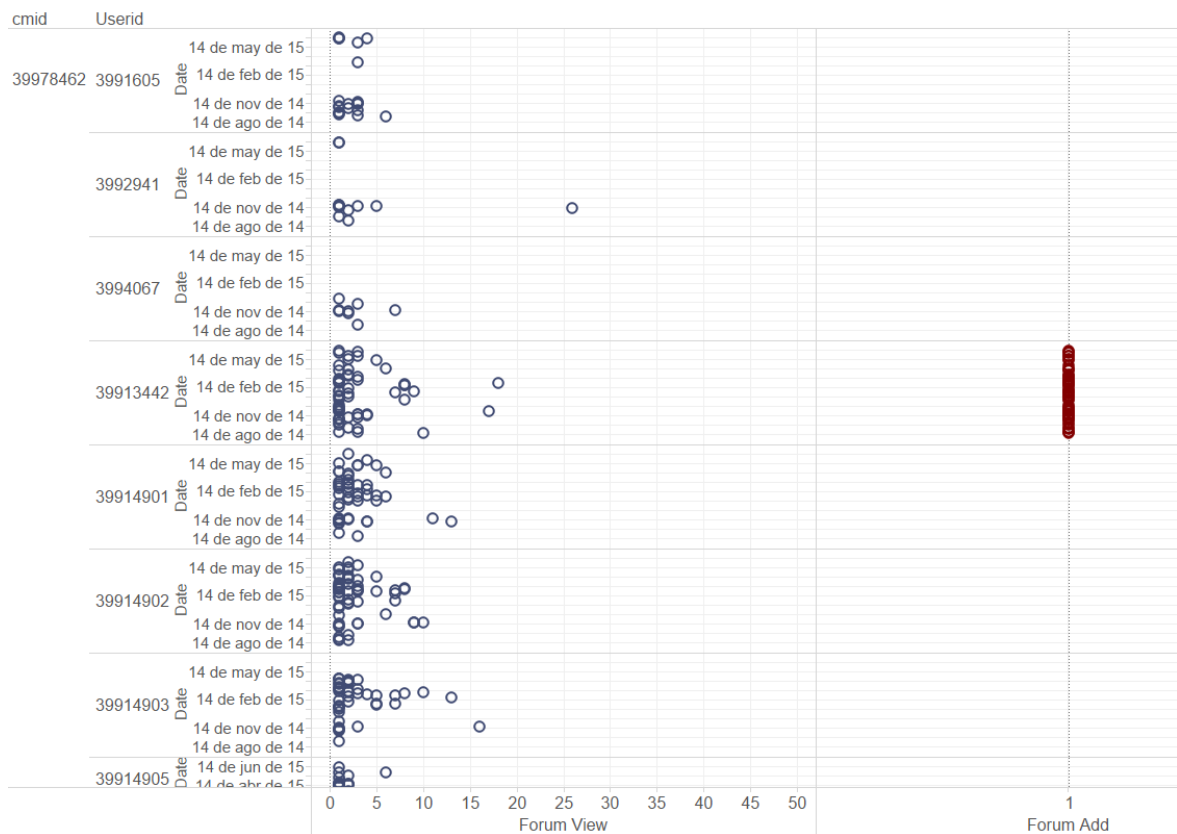


Figure 17: Number of forum views and posts

7.5. Delivery Rate

One of the simplest way to measure students' efficiency of work is by number of solved tasks. This value, but not only this one are measured by Delivery Rate. Let's imagine teacher Tom who would like to see the number of tasks done by his students. However, he doesn't only like to see the results of finished modules, he is willing to know which part of all started modules student has finished. Having this information he can see if student wanted to do more but didn't succeed or has done everything what the/she planned to do.

Delivery rate is calculated in the same transformation as forum efficiency. Thus the transformation is depicted in the same Figure 17. After access to the data and transformation of date format, I filtered quizzes depending on their module. In the first and second branch I was looking for quizzes. Moreover, in the first branch I looked for an open activity and for closing in the second one. I obtained two tables in two branches which I later ordered and merged. As a final step I summed these numbers up. In the given output in Figure 19 you can see the number of opened(circle) and closed(cross) modules. In one axis are given

modules and users and on the perpendicular axis number of open and closed modules. Alternative way of presenting the results of this indicator would be the percentage of closed quizzes. I decided however to give this visualization, since in the data I had, obtained number aren't too big and percentage could be calculated in head. Moreover, the visualisation given below has also important information about exact number of opened and closed quizzes. Teacher can decide how to assess performance of a student who has started and finished 2 tasks and other one who finished only one but tried to solve 4.

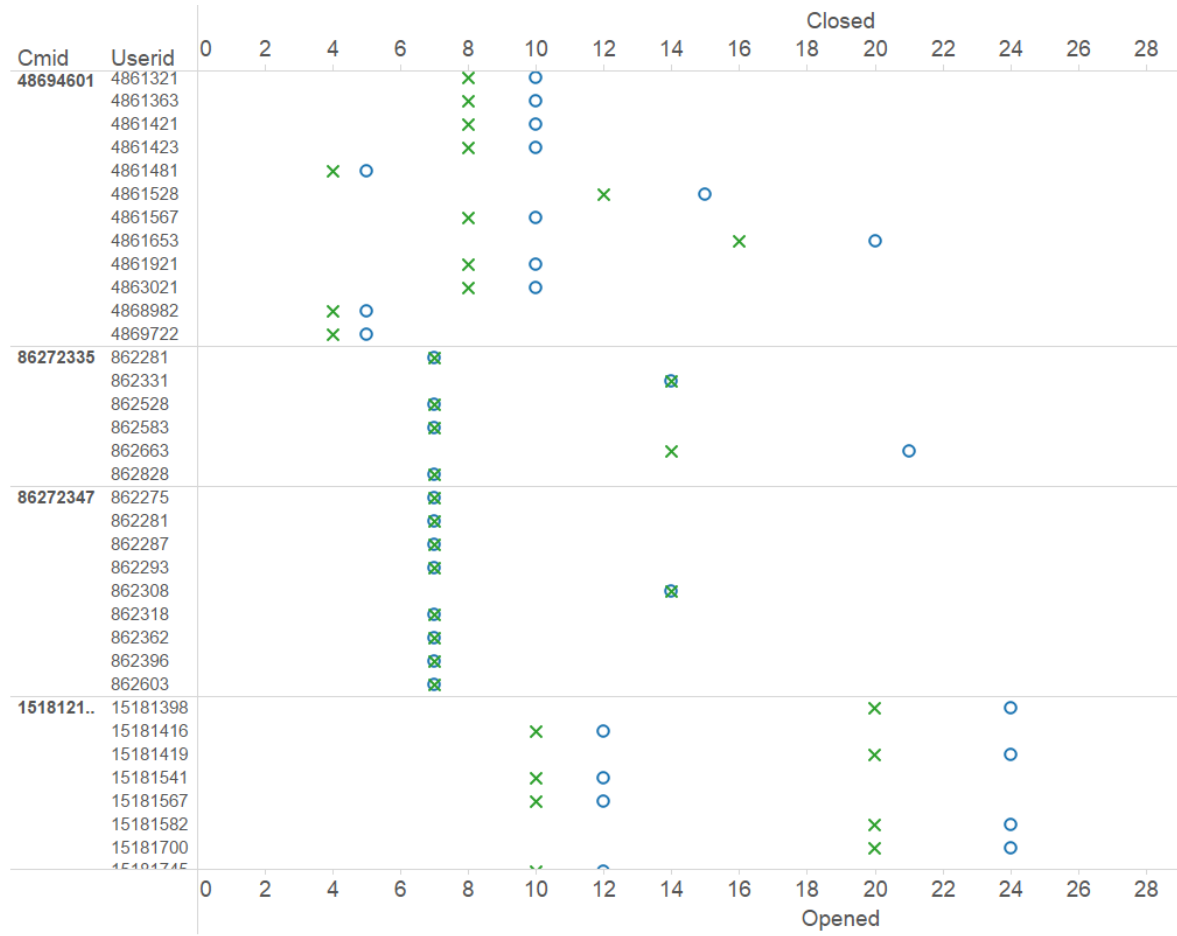


Figure 18: Delivery output

Above presented result applies not only to quizzes but also to other modules, where ratio of finished and started tasks can be calculated, or ratio of accessed to all existing tasks. The way of calculating this indicator will be the same, only names of modules and activity names should be changed.

7.6. Validation

Validation of all calculated indicators was conducted in several ways. For some indicators I could check with an SQL query if results fit into specific range. For such indicator like Delivery Rate I checked if number of opened tasks is bigger or equal to the closed ones. In other cases I checked the correctness by inspection of output plots from Tableau. I was working on samples of data taken from Moodle from around one year only for several courses. Thus checking the usefulness and accuracy in learning process was at this stage impossible. In case of incorrect results I firstly checked if my code was doing what I intended, in Pentaho I analyzed output after each step, in RStudio I was debugging the R code. If the calculations themselves were correct, it meant that the principles were not correct and I had to come back to the beginning and repeat the problem analysis, according to CRISP methodology.

8. Conclusions

With my Master's thesis project *Data Mining in Learning Analytics* I have contributed to the project *Learning Analytics for Secondary Schools* developed at inLab at Faculty of Computer Science at Technical University of Catalonia. I have continued the work left by previous students. I either implemented previously proposed indicators and verified them or proposed my own solutions.

During the work on the thesis I have worked with SQL, R and Pentaho. Each of these tools characterises some advantages and disadvantages. On one hand Pentaho was a great tool at the beginning because without any debugging I could see the result after each step. On the other hand, some simple operations on data required many steps. Pentaho has executional plug-ins for R and SQL, however fixing code in Pentaho was very challenging due to lack of detailed information about errors. Thus, for development of R code I used an IDE RStudio which enabled me to write correct and working code and then put it into Pentaho transformation. The same situation was with SQL, I was firstly checking the correctness of my query in Database Browser and then copying it into Pentaho.

There is possible further development of indicators, which will give more detailed and more complex information about students' motivation. As a part of testing this application,

teachers from schools which are stakeholders of this project should see and comment obtained results. As a next step of developing LA4S project I would suggest connecting it to real time data and set the application to run every 24 hours to see the constant update of data and indicators.

9. Summary

Learning Analytics for Secondary Schools is a very interesting and promising project. The growing popularity of Data Analysis in education give great opportunity to improve learning process. Data Mining in this project is crucial and after implementation of indicators presented in this thesis we can already see possible advantage which these results can bring to teachers who are willing to understand and help their students at school. For me it was a great experience to work with inLab team at Technical University of Catalonia. It gave me an opportunity to gain a lot of knowledge in the topic of Data Mining. In the future I wish to follow this path and make use of what I have learnt during work on project.

References

- [1] C. Haythornthwaite, M. de Laat, D. Shane and D. Suthers, "Introduction to Learning Analytics & Networked Learning Minitrack," in *46th Hawaii International Conference on System Sciences*, 2013.
- [2] P. Long and G. Siemens, "Penetrating the Fog Analytics in Learning and Education," *Educause review*, Sep/Oct 2011.
- [3] Wikibon Blog, "A Comprehensive List of Big Data Statistics," 1 Aug 2012. [Online]. Available: <http://wikibon.org/blog/big-data-statistics/>. [Accessed 10 Jan 2016].
- [4] K. Hill, "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did," *Forbes*, 16 Feb 2012. [Online]. Available: <http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#2715e4857a0bb79af0a34c62>. [Accessed 13 Jan 2016].
- [5] B. Marr, "The Awesome Ways Big Data Is Used Today To Change Our World," 13 Nov 2013. [Online]. Available: <https://www.linkedin.com/pulse/20131113065157-64875646-the-awesome-ways-big-data-is-used-today-to-change-our-world>. [Accessed 09 Jan 2016].
- [6] inLab, "PILARES," [Online]. Available: <https://inlab.fib.upc.edu/en/pilares>. [Accessed 13 Jan 2016].
- [7] G. Piatetsky, "CRISP-DM, still the top methodology for analytics, data mining, or data science projects," *KDnuggets*, 28 Oct 2014. [Online]. Available: <http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>. [Accessed 20 Jan 2016].
- [8] Digg Data, "Project management for Data Science projects," [Online]. Available: <http://digdata.in/post/129903266636/project-management-for-data-science-projects>. [Accessed 14 Jan 2016].
- [9] Pentaho Corporation, "About Us," [Online]. Available: <http://www.pentaho.com/about>. [Accessed 15 Jan 2016].
- [10] The R Foundation, "What is R?," [Online]. Available: <https://www.r->

project.org/about.html. [Accessed 15 Jan 2016].

[11] T. Khabaza, "Nine Laws of Data Mining," KD Nuggets, [Online]. Available: <http://www.kdnuggets.com/2015/06/nine-laws-data-mining-part-1.html>. [Accessed 14 Jan 2016].

[12] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data".

[13] J. P. Campbell and D. G. Oblinger, "Academic Analytics," *Educause*, Oct 2007.