DEVELOPING SPEED AND INTENSITY INDICATORS IN THE LEARNING ANALYTICS PLATFORM FOR SECONDARY SCHOOLS

BACHELOR THESIS REPORT

Author
Balaji Natarajan (B. Tech Computer Science and Engineering SASTRA University)

Thesis supervisor
Tomas Aluja Banet

Co-Supervisor
Albert Obiols Vivas

Internal Examiner
Maria Ribera Sancho Samso
Acknowledgements

First of all, I thank my thesis director, Prof. Tomas Aluja for his constant support in the development of my project. I would also thank Prof. Maria Ribera for her valuable suggestions and guidance in my TFG work as well as my TFG report. I am very grateful for Mr. Albert Obiols Vives for his kindness in helping me with learning materials and helping me overcome some of the obstacles in my TFG work. I take immense pleasure in conveying my gratitude towards my mom Dr. A. Umamakeswari for her continuous support throughout this journey. I also thank Dr. M. Sridharan for coordinating us with immense interest. I also thank my home university SASTRA for giving me this valuable experience to work in a world class University. I also thank my team members at inLab, FIB especially Jordi Cassanovas for offering me valuable help whenever I needed the most. I thank the God for my good health which helped me complete my project successfully.
Table of Contents

Acknowledgements ............................................................................................................. ii

TABLE OF FIGURES .......................................................................................................... iv

ABSTRACT .......................................................................................................................... v

CHAPTER 1 .......................................................................................................................... 1

INTRODUCTION .................................................................................................................. 1

1.1 CONTEXT ...................................................................................................................... 1

1.2 STAKEHOLDERS AND USERS OF THE SYSTEM ..................................................... 2

1.3 OBJECTIVES ............................................................................................................... 2

1.4 LEARNING ANALYTICS PROJECT AT InLab, FIB .................................................... 2

CHAPTER 2 .......................................................................................................................... 5

STATE OF THE ART ............................................................................................................. 5

2.1 STATE OF THE ART ................................................................................................... 5

CHAPTER 3 .......................................................................................................................... 7

THEORETICAL FRAMEWORK ........................................................................................... 7

3.1 DATA MINING ............................................................................................................. 7

3.2 ETL - EXTRACT TRANSFORM LOAD ....................................................................... 8

3.3 STRUCTURED QUERY LANGUAGE ............................................................................. 9

3.3.1 Data Definition Language (DDL) ......................................................................... 9

3.3.2 Data Manipulation Language (DML) ................................................................... 9

3.3.3 Joins ........................................................................................................................ 9

3.4 TOOLS USED ............................................................................................................. 10

3.5 MOTIVATION ............................................................................................................... 11

3.5.1 Components of Motivation .................................................................................. 11

3.5.2 Characteristics of Motivation ................................................................................ 11

3.5.3 Constraints in measuring Motivation .................................................................. 12

3.5.4 Importance of Indicators and measurement of motivation .................................. 12

3.6 PCA (Principal Component Analysis) ......................................................................... 13

3.7 INDICATORS PROPOSED AND DEVELOPED ........................................................ 14

CHAPTER 4 .......................................................................................................................... 17

METHODOLOGY ............................................................................................................... 17

CHAPTER 5 .......................................................................................................................... 19

SCOPE AND REQUIREMENTS ENGINEERING .............................................................. 19

5.1 SCOPE ......................................................................................................................... 19

5.2 REQUIREMENTS ENGINEERING ............................................................................... 20
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.6 SPEED DELIVERY</td>
<td>78</td>
</tr>
<tr>
<td>8.7 TRANSITION TIME</td>
<td>79</td>
</tr>
<tr>
<td>8.8 DELIVERY RATE</td>
<td>80</td>
</tr>
<tr>
<td>8.9 VALIDATION BY PCA</td>
<td>81</td>
</tr>
<tr>
<td>CHAPTER 9</td>
<td>84</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>84</td>
</tr>
<tr>
<td>9.1 SCOPE FOR FUTURE WORK</td>
<td>84</td>
</tr>
<tr>
<td>9.2 LEARNING OUTCOMES</td>
<td>84</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>86</td>
</tr>
</tbody>
</table>
# TABLE OF FIGURES

Figure 1 INITIAL ARCHITECTURE OF LEARNING ANALYTICS PROJECT AT InLab ........................................... 3  
Figure 2 CURRENT ARCHITECTURE; AUTHOR: JORDI CASANOVAS .......................................................... 4  
Figure 3 DATA MINING STEPS ....................................................................................................................... 8  
Figure 4 TYPES OF JOIN ................................................................................................................................. 10  
Figure 5 ORANGE JUICE EXAMPLE .................................................................................................................. 14  
Figure 6 waterfall vs agile approach ................................................................................................................. 17  
Figure 7 SUMMARY OF BUDGET ..................................................................................................................... 26  
Figure 8 PROJECT FLOW ................................................................................................................................. 28  
Figure 9 STARTING DATE RESULT ................................................................................................................... 31  
Figure 10 AGILITY RATE ................................................................................................................................. 35  
Figure 11 TIME SPENT OUTPUT ..................................................................................................................... 41  
Figure 12 SPEED DELIVERY ............................................................................................................................. 45  
Figure 13 TRANSITION TIME OUTPUT ............................................................................................................ 50  
Figure 14 DELIVERY RATE .............................................................................................................................. 55  
Figure 15 Engagement level .............................................................................................................................. 60  
Figure 16 COMPETITIVE LEVEL ...................................................................................................................... 65  
Figure 17 CORRELATION PLOT - PCA ............................................................................................................ 67  
Figure 18 VARIMAX ROTATION ....................................................................................................................... 68  
Figure 19 MOTIVATION PLOT FOR INFORMATICA ...................................................................................... 70  
Figure 20 HISTOGRAM OF MOTIVATION INDEX ........................................................................................... 71  
Figure 21 CORRELATION PLOT FOR VALIDATION ....................................................................................... 82  
Figure 22 VARIMAX ROTATION - VALIDATION ............................................................................................. 83
ABSTRACT

DEVELOPING SPEED AND INTENSITY INDICATORS IN THE LEARNING ANALYTICS PLATFORM FOR SECONDARY SCHOOLS

Learning analytics is a branch of data science that mainly encompasses statistical techniques to understand and optimise learning process particularly in an online learning environment like MOODLE (Agora). The main goal of Learning Analytics project at inLab is to provide information associated with students’ motivation level from data generated through logs in the Learning Management System from 1,500 secondary schools across Catalunya. This project refines, and thereby enhances the work done by members who had contributed to the previous version of the project. The final result of the project is shown with the help of a dashboard which shows various indicators associated with the motivation level of students. The project also aims at using Principal Component Analysis to compute motivation level of each student as a numerical value. The various indicators used to compute motivation level comes under four classes namely speed, intensity, persistence and choice. My TFG work consists of development of indicators such as agility rate, time spent, transition time and speed delivery (belonging to speed category) along with delivery rate, engagement level and competitive level (belonging to intensity category). New indicators like speed delivery were proposed in the above list and refinement in the existing ones were also done. Right now, the percentage of students successfully completing secondary schools in Catalunya is low and so the final product which will act as a monitoring tool for the teachers is definitely going to improve this to achieve better standards.
CHAPTER 1

INTRODUCTION

This chapter gives an introduction to the project work done: “Developing speed and intensity indicators in the learning analytics platform for secondary schools”, its context and various stakeholders associated with the project.

1.1 CONTEXT

First of all, it is important to define what learning analytics is. Learning analytics is a branch of data science that applies usage of technologies in online learning platform in order to optimize the learning process (1). In this project, the supported learning platform is the Moodle based Learning Management System (LMS) called Agora, which is one of the most used learning platforms in the secondary schools of Catalunya. Data generated from log files of students are analysed which can be used to characterize the learning pattern of students. It is expected to be in use all over Catalunya region reaching 1,500 schools and 400,000 students.

This project is intended to capture the learning pattern as well as the motivation of student(s) as a function of several indicators. My contribution has been made in the development and visualization of speed and intensity indicators.

SoLAR (Society for Learning Analytics Research) and SNOLA (Spanish Network of Learning Analytics) are some of the organisations that groups existing research groups and individuals in the field of learning analytics. It is worth mentioning that this project is within the scope of PILARES (Plataforma Inteligente de Learning Analytics para mejorar el Rendimiento en Educacion Secundaria) funded by Ministerio de Economia y Competitividad (MEC).

Areas of expertise involved in the project are:

- Data mining & big data analytics
- Data and information analysis
- Statistical methods for the measurement of intangible
- E-learning & virtual campus environments
- Data visualisation
- Software Engineering
1.2 STAKEHOLDERS AND USERS OF THE SYSTEM

Stakeholders refer to a person or a group of people who are directly or indirectly affected by the product, here the Learning Analytics tool. Thus, the various stakeholders of the systems are summarized below.

• The teachers of the secondary schools, who are the direct users of the system developed.
• The students of secondary schools in Catalunya, who are studied and thereby affected as a result of the product.
• The parents of the associated students who are indirectly affected by the system, as they are responsible for their children’s education.
• The Education department of Catalunya. The project is carried out in accordance with them.
• Directors of the secondary schools who are responsible for their school’s performance.

• inLab FIB (as organisation)
  The inLab Participating team:
  • Jordi Casanovas
  • Unai Sanchez Luque
  • Balaji Natarajan (Myself)
  • Attuluri Mohana Preethi
  • Prof. Tomas Aluja Banet (Director), Prof. Maria Ribera Sancho Samso, (Ponent) and Albert Obiols (Supervisor).

1.3 OBJECTIVES

It is very important at this point to mention the work which has been carried out by me in inLab, FIB. The following are the objectives:
A. Development of speed and intensity indicators which will be used to calculate motivation index.
B. Visualisation of those developed indicators.
C. Performing Principal Component Analysis to compute the motivation value from all the defined indicators.

1.4 LEARNING ANALYTICS PROJECT AT InLab, FIB

The Learning Analytics project at inLab, FIB (2) started on 2014 with master’s thesis of Miriam Ramirez (DMKM student) who proposed the 1st design of the project. She initiated the ETL process which has been enhanced by inLab development team. As a part of her thesis she developed 4 indicators namely,
1. Percentage of Accesses.
2. Number of Accesses.
3. Time to first access
4. Time spent on the activity.

The 1st proposal of the project was implemented by Jordi Casanovas and Pau Vila. Then, Hari Hara and Pranathi joined the project (Spring 2015) to contribute to the dashboard design and implementation of new indicators which complemented Ivan’s work on theoretical justification of motivation. Then, Joanna Sykurska as a part of her master’s thesis contributed to the project by developing more indicators (time spent, curiosity rate, forum activity and delivery rate). Now, for this spring, I and A M Preethi have joined this project to contribute to development of indicators (belonging to speed, intensity, persistence and choice classes) development and perform PCA (Principal Component Analysis) in order to perform correlation between developed indicators and to compute motivation value for each student across different subjects.

\[ \text{Figure 1 INITIAL ARCHITECTURE OF LEARNING ANALYTICS PROJECT AT InLAB} \]
The above figures (Figures 1 and 2) show the evolution of architecture of LA project at inLab, FIB. My contribution has been done on the Learning Analytics Database layer in the Figure 2, while ETL has been done by Jordi Cassanovas.
CHAPTER 2

STATE OF THE ART

This chapter discusses related work to my TFG available in the field of Learning Analytics, cites references in this field and describes how my TFG is part of inLab’s Learning Analytics project.

2.1 STATE OF THE ART

The project I have undertaken as a part of my TFG is a branch of data science. The concept of Learning Analytics is defined as “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs” in the 1st International Conference on Learning Analytics & Knowledge (LAK 2011) (3).

Learning analytics cover a wide range of analytic, which are defined as macro, meso and micro-levels. Micro-level analytics focuses on individuals involved in learning i.e. students whereas meso and macro-level analytics operates at institution and region/state/national/international levels respectively (4). The Learning Analytics project at inLab’s objective is to perform analytics at micro-level. To be more precise Learning Analytics project at inLab performs analytics at the school level, where the main goal is to monitor the performance of students.

Ji Won You presented a study, “Identifying significant indicators using LMS data to predict course achievement in online learning” in order to identify indicators of learning to predict course achievement (5). This study used data generated from 530 college students in an online course. It describes how indicators such as regular study, late submissions of assignments, number of sessions (the frequency of course logins), and proof of reading the course information packets were able to predict course achievement by correlation analyses between the indicators. These indicators do not directly match our need. The indicator late submissions of assignments gave inspiration to develop Speed Delivery indicator in this project which shows the time gap between deadline and submission time.

Paulo Blikstein in the paper, “Using learning analytics to assess students’ behaviour in open-ended programming tasks” described an automated technique to assess, analyse and visualize students’ behaviour in learning computer programming (6). He developed metrics such as compilation frequency, code size, code evolution pattern, frequency of correct/incorrect compilations, etc. The main purpose was to make way for instructors so that they could monitor students’ performance in real time, mainly when the system indicates that students are in a critical zone. We do the
same thing by monitoring the students’ performance in real time but we do for secondary schools with different set of indicators because we develop a tool that monitors performance of students in secondary schools associated with the relevant subjects. The objective is also same for both i.e. to alert the respective teacher about students who are in critical zone.

In the paper “Exploring Learning Analytics as Indicators of Study Behaviour” by Rob Phillips, Dorit Maor, Greg Preston and Wendy Cumming-Potvin, a Learning Analytics tool which was used to capture learning pattern of students and categorize them is described (7). The categories used are Conscientious, High-achieving, Good intentioned, Repentant users, Bingers, Free-timers, Cramming, One-hit wonders, Random and Non-user. The learning patterns they analysed are delay in listening to recordings and number of accesses. We do not categorize students like this instead we monitor students’ motivation in form of indicators and visualise them.

LOCO-Analyst is a learning analytics tool developed by Liaqat Ali, Marek Hatala, Dragan Gasevic and Jelena Jovanovic to provide educators with feedback on students learning activities and performance (8). The feedback elements included are individual lesson, a group of (related) lessons, a learning module as a whole, students’ performance on a quiz, student’s activities in discussion forums and chat rooms, Student’s interaction with the learning content (lessons), Student’s comprehension of the studied topics (based on his/her annotations). The main difference between LOCO analyst and Learning Analytics project at inLab, FIB is that we do not take students’ performance directly into account since it shows only the motivation of students to do well in tests.

Initial Context

The final product is going to be deployed as a web portal which is a pilot version. This pilot version does Learning Analytics from data across 6 secondary schools and 15 different subjects. It can be scaled to work with data across 1500 secondary schools in Catalunya. The previously deployed version shows 4 plots Percentage of accesses, Number of accesses, Time to first Accesses and total time spent. On completion of my TFG 7 plots (Agility rate, time spent, transition time, speed delivery, delivery rate, engagement level and competitive level) are going to be added. An extensive study of the MOODLE framework and its databases was required initially to work on. The indicators are programmed using R and MySQL and hence significant knowledge in these was required.

It is worth mentioning that the product developed does not violate any governmental policies and it sticks to all the regulations and the laws imposed.
CHAPTER 3

THEORETICAL FRAMEWORK

This section gives description of the theoretical framework involved in the project. This is more important because only with strong basics it is possible to make robust contribution. In this section the various concepts such as the ETL, SQL queries, indicators and motivation are discussed.

3.1 DATA MINING

Data mining is the process of discovering significant patterns and knowledge from large amounts of data (9). It consists of the following steps:

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining
6. Pattern evaluation
7. Knowledge presentation

Most the above steps are followed in the Learning Analytics project at inLab. Data cleaning, integration, selection and transformation is carried out by means of ETL process. All the indicators that run every day in form of R scripts do the job of data mining. Pattern evaluation is currently not carried out, but there are indicators developed (time spent, transition time, etc.) that can be used to do this. Knowledge presentation is done by visualising the output in form of a dashboard.
The above figure (Figure 3) shows how data mining steps take place to get the final result. i.e. the knowledge from data.

### 3.2 ETL - EXTRACT TRANSFORM LOAD

In computing, **Extract, Transform and Load (ETL)** (10) refers to a process in database usage and especially in data warehousing that performs:

- **Data extraction** — extracts data from homogeneous or heterogeneous data sources
- **Data transformation** — transforms the data for storing it in the proper format or structure for the purposes of querying and analysis
- **Data loading** — loads it into the final target (database, more specifically, operational data store, data mart, or data warehouse)

This is the first step done which runs every 24 hours in order to populate the database (Westeros Agora). This is done by my team member Jordi Casanovas Muñoz using Perl scripts. The tool which is used to perform ETL is Pentaho Data Integration kettle.
3.3 STRUCTURED QUERY LANGUAGE

In this section, the theoretical foundations of the database technology is described. The Database model used in the project is MySQL. This section describes the various types of queries associated with my TFG.

SQL can be broadly categorized into the DDL (Data Definition Language), DML (Data Manipulation Language) and Database Control Commands.

3.3.1 Data Definition Language (DDL)

The Data Definition Language (DDL) is used to define new tables or alter the existing ones. Some of the commonly used queries as a part of the development of the project:

CREATE TABLE - This command is used to create a table to the specified database.
ALTER TABLE - This command is used to add, delete, or modify columns in an existing table in the database.
DROP TABLE - This command is used to delete a table from an existing database which cannot be reverted.

3.3.2 Data Manipulation Language (DML)

The Data manipulation language (DML) of the SQL is used to perform operations on the data stored in the database. They were very important in the development of all the indicators. The most commonly used DML commands are:

SELECT- This is used to select a subset of rows from a table based on specified condition in the WHERE clause.
UPDATE- This is used to update an existing table just by modifying the values of necessary columns.
INSERT- This is used to insert new records into the table.

3.3.3 Joins

There are 4 types of join (12):

INNER JOIN: The INNER JOIN keyword selects all rows from both tables as long as there is a match between the columns in both tables.
LEFT OUTER JOIN: The LEFT OUTER JOIN keyword returns all rows from the left table (table1), with the matching rows in the right table (table2). The result is NULL in the right side when there is no match.
RIGHT OUTER JOIN: The RIGHT JOIN keyword returns all rows from the right table (table2), with the matching rows in the left table (table1). The result is NULL in the left side when there is no match.
**FULL JOIN.** The FULL OUTER JOIN keyword returns all rows from the left table (table1) and from the right table (table2).
The FULL OUTER JOIN keyword combines the result of both LEFT and RIGHT joins. This is not available in MySQL and hence results of LEFT and RIGHT OUTER JOINS needs to be combined by union operation.

![Diagram of JOIN types](http://tech.queryhome.com/58353/what-are-the-different-types-of-joins-in-sql-and-mysql)

The above figure (Figure 4) visualizes all the four different types of join in Venn diagrams.

### 3.4 TOOLS USED

**R STUDIO**

It is an open source IDE (Integrated Development Environment) for R language (13). R is a statistical programming language mainly used for data analysis of large data sets developed by Ross Ihaka and Robert Gentleman in the 90’s. I mainly used R for data analysis part and visualisation using ggplot2, an R library used to create publication quality plots.

**DATABASE BROWSER**

It is an open source application software which I used mainly to run MySQL queries and validation of indicators which is stored in form of tables (14). The main advantage is that it is easy to work with. On the other hand, it has the problem of crashing when heavy computation is running.
3.5 MOTIVATION

Motivation is defined as “the process that initiates, guides, and maintains goal-oriented behaviours” (15). Motivation is what makes us act from fetching water for quenching thirst to reading a book for gaining knowledge.

3.5.1 Components of Motivation

There are three major components to motivation: activation, persistence and intensity.

Activation involves the decision to initiate a behaviour, which makes a person to start working. Woody Allen once said “80% of success is showing up”. Consider a boy who joins a gym. Initially, he may hate getting pain and the thought of getting ready for it may be stress provoking than the actual physical exercises. Only his decision to initiate can help him accomplish his goal.

Persistence is the continued effort toward a goal in spite of existence of obstacles. It shows how bad you want to accomplish the goal. “Nothing in the world can take the place of Persistence. Talent will not; nothing is more common than unsuccessful men with talent. Genius will not; unrewarded genius is almost a proverb. Education will not; the world is full of educated derelicts. Persistence and determination alone are omnipotent. The slogan “Press On” has solved and always will solve the problems of the human race.” – Calvin Coolidge (16)

For example, you may become a lot more enthusiastic about dieting and exercising once you’ve lost those first 10 pounds and feel your clothes fitting more loosely (16).

Intensity can be seen in the amount of effort a person makes in order to accomplish a goal. For example, one student might progress easily without much effort, while another student studies regularly, participate in discussions and take advantage of all the opportunities even outside of the class. The first student lacks intensity, while the second moves forward with much greater intensity (15).

3.5.2 Characteristics of Motivation

Complex phenomenon: Complex structure, complex interconnections with the other phenomena. For example, a motivated boy performing consistently in exams cannot perform well suddenly when he gets sick.

Intangible: Motivation cannot be observed directly. We don't actually observe a motive; rather, we infer that one exists based on the behaviour we observe Nevid (2013) (17).
**Dynamic**: It changes over time, and those changes can be extreme. E.g. a sportsman suffering from a bad form can get motivated all of a sudden after a few minutes of positive outcome.

**Personal**: It is an internal feeling which is different for each person. E.g. although all students in a class are taught by the same teacher each performs differently from one another in various platforms like studies, sports, etc.

### 3.5.3 Constraints in measuring Motivation

#### Motivation has to be measured indirectly

As motivation is intangible psychological quantity one has to use indicators to estimate its level. This means that indicators are measured directly and motivation is estimated using those indicators. Learning results are used as an indicator of learning motivation. However, Romainville (1994), Bessant (1997) and Chen (2004) found out that there is a correlation between learning strategy and learning results (18). According to the theory of self-regulated learning and research of Wang et al (2008), both learning motivation and learning strategy have direct effects on learning results (19). Therefore, using only learning results as an indicator of learning motivation is wrong.

#### Motivation is measured in relative terms

Motivation is always compared to something else. It can be compared to its own previous or subsequent levels, to motivation in a different goal state, to motivation of different people etc.

#### Motivation has to be measured constantly

It is a quantity which is very dynamic and hence it has to be measured constantly.

#### Motivation has to be externally measured

Self-reported measures of motivation is an approach where people are asked to rate their motivation level. However, as psychologists David C. McClelland and John W. Atkinson argued, although one can be motivated, he or she need not have to be conscious of their own motivation level (20). In fact, one does not necessarily need to have conscious understanding of his or her own psychological state. Thus, this approach can capture only the conscious part of motivation while neglecting a large part of it.

### 3.5.4 Importance of Indicators and measurement of motivation

Before explaining the indicators of motivation developed as a part of my TFG, it is more important to explain the necessity of developing indicators. Motivation is an
intangible quantity so to measure it indirectly we are developing indicators which can be used to compute motivation of students by Principal component Analysis. In my TFG I have proposed one indicator called Speed Delivery which was also developed by me. This project is an enhancement of work already done in order to improve its robustness (2). The validation of the developed indicators is done mainly by correlation analysis using a matrix. Motivation accounts only for 40% of the learning outcomes but absence of it cannot let learning happen.

3.6 PCA (Principal Component Analysis)

Principal Component Analysis (21) is a statistical technique that reduces the amount of variance in the data in orthogonal dimensions. It is used to reduce dimensions in data.

The main goals (22) of principal component analysis is:

- to identify hidden pattern in a data set
- to reduce the dimensionality of the data by removing the noise and redundancy in the data
- to identify correlated variables

Correlation of variables means dependency between two variables. If two variables are positively correlated, then increase in value of one increases value of another and vice-versa. The opposite thing happens when two variables are negatively correlated.

To obtain the orthogonal dimensions we perform the diagonalization (Eigen decomposition) of the correlation matrix. Eigen value and Eigen vector (23) occur in pairs. Eigen vector gives the direction of maximum variance whereas Eigen value is a number which tells how much variance is there in a particular direction.
The above figure (Figure 5) shows correlation analysis of orange (24) juice data set done using FactoMineR. Dim 1 having 67.77% indicates that the dimension can explain 67.77% of variation in data with additional of 19.05% explained by second component. The first component correlates almost perfectly with the variable Odor Typicality, while the second component correlates very highly with Odor Intensity and Pulpiness. It is evident that Bitterness is the opposite of Sweetness and much more information can be derived by looking at the correlation of variables from the above plot. This is used in my TFG to understand the correlation between the indicators I developed and there by validating them.

3.7 INDICATORS PROPOSED AND DEVELOPED

SPEED INDICATORS: measures how fast an individual start the task after its recognition, how fast an individual completes the task and how fast an individual moves from one task to the next one.

INTENSITY INDICATORS: It is widely accepted that motivation positively influences intensity of the effort done to achieve the goal; thus, intensity indicators can be used as an indirect measure of motivation.
<table>
<thead>
<tr>
<th>S. NO.</th>
<th>NAME OF THE INDICATOR</th>
<th>DEFINITION</th>
<th>ADDITIONAL INFORMATION</th>
<th>CLASS OF THE INDICATOR</th>
<th>STATISTICAL DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Agility Rate</td>
<td>This indicator shows how much time a student takes to access a learning object (like quiz, forum, etc.) for the first time.</td>
<td>Calculated for all the tasks (both mandatory and non-mandatory).</td>
<td>Speed</td>
<td>(Starting Date of the task) – (First Access to that task)</td>
</tr>
<tr>
<td>2</td>
<td>Time Spent</td>
<td>It is the total time spent on the Virtual Learning Environment (VLE) in order to complete a task.</td>
<td>Calculated for all mandatory tasks (Quiz, Assignment and Hotpot).</td>
<td>Speed</td>
<td>(Continue attempt1 – Start attempt) + (Continue attempt2 - Continue attempt1) + …. + (Submit attempt – Continue attemptN)</td>
</tr>
<tr>
<td>3</td>
<td>Transition Time</td>
<td>It is computed as the mean time of time spent between consecutive tasks across different subjects.</td>
<td>Calculated every day for each subject taking all the tasks into account.</td>
<td>Speed</td>
<td>MEAN(Starting time of access to the task – Closing time of access to the previous task)</td>
</tr>
<tr>
<td>4</td>
<td>Speed Delivery</td>
<td>It shows the time gap between deadline of a task and submission time of that task for each student.</td>
<td>Calculated for all mandatory tasks.</td>
<td>Speed</td>
<td>(Deadline) – (Submission time)</td>
</tr>
<tr>
<td>5</td>
<td>Delivery rate</td>
<td>This indicator reflects percentage of pending obligatory tasks a student has completed during the past 7 days.</td>
<td>Calculated every day for all mandatory tasks across each subject.</td>
<td>Intensity</td>
<td># fulfilled obligatory tasks in 7 days # active assigned obligatory tasks in 7 days</td>
</tr>
<tr>
<td></td>
<td>Engagement level</td>
<td>This indicator reflects how active a particular student is on a given day, in comparison to his best performance in the last 14 days.</td>
<td>Calculated every day for each subject.</td>
<td>Intensity</td>
<td># all activities in 1 day</td>
</tr>
<tr>
<td>---</td>
<td>------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------</td>
<td>-----------</td>
<td>--------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td># best of last 14 days</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Competitive level</td>
<td>This indicator reflects how active a student is on a given day, compared to the most active student in the past 7 days.</td>
<td>Calculated every day for each subject.</td>
<td>Intensity</td>
<td># all activities in 1 day</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td># most active student in the past 7 days</td>
</tr>
</tbody>
</table>
CHAPTER 4

METHODOLOGY

As a methodology for this project, the inLab Learning Analytics team follows agile software development methodology. Agile software development consists of a set of principles for software development, where requirements and hence solutions evolve over time (25). It promotes adaptive planning, evolutionary development, early delivery, and continuous improvement, and it encourages necessary response to change (26). From the above figure (Figure 6), it can be seen that quality, time and cost are variable in waterfall model since features are fixed (not accepting changes). Whereas in the agile model, the opposite thing happens cost, time and quality are fixed but features are variable.

![Figure 6 Waterfall vs Agile Approach](http://www.intechnic.com/blog/agile-vs-waterfall-website-project-management-methodologies/)

**Agile principles**

The Agile Manifesto is based on twelve principles (27):

- Customer satisfaction by early and continuous delivery of valuable software
- Welcome changing requirements, even in late development
- Working software is delivered frequently (weeks rather than months)
- Close, daily cooperation between business people and developers
- Projects are built around motivated individuals, who should be trusted
- Face-to-face conversation is the best form of communication (co-location)
• Working software is the principal measure of progress
• Sustainable development, able to maintain a constant pace
• Continuous attention to technical excellence and good design
• Simplicity—the art of maximizing the amount of work not done—is essential
• Best architectures, requirements, and designs emerge from self-organizing teams
• Regularly, the team reflects on how to become more effective, and adjusts accordingly

There are many agile methods. Among those Agile Unified Process (AUP) method is apt for this project.

The Agile Unified Process (AUP) (27) adopts a “serial in the large” and “iterative in the small” philosophy for building computer-based systems. It is a simplified version of RUP (Rational Unified Process). It involves usage of agile techniques while still being similar to RUP. The UP phased activities: inception, elaboration, construction, and transition. AUP provides a serial overlay which enables a team to visualize the process flow for any software project. Each AUP iteration addresses the following activities:

• Modelling. Models of the problem domains are created. However, to stay agile, these models should be “just barely good enough” to allow the team to proceed.
• Implementation. Models are translated into source code.
• Testing. The team designs and executes a series of tests to uncover errors and ensure that the source code meets its requirements.
• Deployment. Focuses on the delivery of a software increment and the acquisition of feedback from end users.
• Configuration and project management. Project management tracks and controls the progress of the team and coordinates team activities.

AUP is more suitable because requirements change dynamically. Periodical technical reviews have been conducted every week where the software developed in that period is shown along with the necessary changes made in the past.

Change in the methodology was not necessary because each task involved in the project addresses all the activities involved in AUP. First, the problem was identified and viable solutions were discussed and then it was implemented in R. After that, testing of the developed code took place, followed by deployment in the repository. And, the most important of all, it always focused on project management to control the progress and coordinate activities. Following this methodology, the development has been done within the specified time frame and cost, with all the necessary changes incorporated.
CHAPTER 5

SCOPE AND REQUIREMENTS ENGINEERING

5.1 SCOPE

This project aims at developing new indicators and enhancing the existing ones in the Learning Analytics platform. Data generated from logs of students in the Learning Management System is used to populate the MOODLE database which is utilised to find significant information from the data.

This project involves:

Data Mining which is the process of discovering interesting patterns and knowledge from large amount of data (9). This process is used to compute indicator values of motivation of students to identify their learning pattern.

Principal Component Analysis (PCA) which is the process of finding a direction that is orthogonal having maximum variance (21). This is used to find motivation index of all the students from several indicators developed.

Exploratory Data Analysis (EDA) which is used to get summary information from data sets in the form of visual output Invalid source specified.. This process is used validate the correctness of data analysis and present the results.

The scope of the project is summarized as follows

• **JUSTIFICATION**: Once the final product is ready, it can be used to study the learning style of students from schools all over Catalunya. The project also visualizes the results in informative graphs and plots which helps the teachers and the analysts to draw inferences. My TFG work encompasses half of the indicators (Speed and Intensity) developed and visualises them.

• **PRODUCT SCOPE**: The outcome of the Learning Analytics project is a product. This product is a platform that enables school teachers and headmasters to track the learning behaviour of students. The final result is a complete product which will be used to perform learning analytics on data obtained from 6 schools in Catalunya. My TFG comes under the scope of Learning Analytics project at inLab, FIB mainly in data mining part.

• **CONSTRAINTS**: For the product to work properly, both integrity and availability of data has to be ensured as the product completely depends on data from log files to produce desired output. Another constraint is timeframe as the project has to be completed on time within the deadline. These constraints are same for my TFG as well as Learning Analytics project at inLab, FIB.
5.2 REQUIREMENTS ENGINEERING

This part describes both functional and non-functional requirements associated with this TFG. The functional requirements is associated with behaviour of the system whereas, non-functional is associated with performance characteristics.

5.2.1 Functional requirements

i. **Day to start on (Starting date)**, which is used to find starting date of all the learning materials in LMS for which the teacher has not specified starting date.

ii. **Agility rate**, which reflects how fast a student accesses a task uploaded in LMS.

iii. **Time spent**, which shows the total amount of time a student takes to submit a compulsory task.

iv. **Transition time**, which reflects how fast a student starts another task after doing the previous one across different subjects.

v. **Speed delivery**, which represents the time gap between the time of submission of a compulsory task and the deadline associated with that compulsory task.

vi. **Delivery rate**, which reflects the % of pending task (compulsory) a student completes in a period of 7 days.

vii. **Engagement level**, which shows how active a student is on a given day, in comparison to his best performance in the past 14 days.

viii. **Competitive level**, which shows how active a student is on a given day, with respect to the most active student in the past 7 days.

5.2.2 Non-functional requirements

i. **Time efficiency of algorithms**: Each indicator developed uses one or more algorithms. Efficient algorithms will take less amount of time to run. The algorithms developed are used every day and takes few seconds to run.

ii. **Space efficiency of algorithms**: Resource utilization while using the algorithms is also important. The indicators developed as a part of this TFG are stored in form of tables. The indicators that have values are only stored and all other records which have NA (Not Available) values are not stored.

iii. **Handling Large Amounts of data**: This project deals with a lot of data associated with 535 students across 6 schools and 15 different subjects. Every day all the tables grow so the developed programs should be capable of handling a large amount of data.
CHAPTER 6

PROJECT MANAGEMENT

This chapter deals with project management aspects of this project. This section explains the scope, project planning, budget planning and sustainability of this project.

6.1 PLANNING

This part represents initial and final Gantt chart and explains how and why they differ from one another.

<table>
<thead>
<tr>
<th>MONTH</th>
<th>FEB</th>
<th>MAR</th>
<th>APR</th>
<th>MAY</th>
<th>JUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>TASK / WEEKS</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>INITIAL SETUP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R ONLINE COURSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEP COURSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLANNING</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAY TO START ON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGILITY RATE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIME SPENT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPEED DELIVERY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRANSITION TIME</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DELIVERY RATE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6.1.1 Task description and reasons for deviation from the original schedule

The above Gantt chart (Table 1) shows the activities till mid of April only. Based on the priorities and obstacles few changes can be seen in the final Gantt chart (Table 2) below. The meetings with the team members had been held on weekly basis for discussing necessary changes and verification of the presented module had also been done.

Before beginning the project it is necessary to decide which tools and software products are going to be used. It is also equally important to get the basics right. In the initial setup stage, I did go through some of the documentation of previously done work, by the team members of Learning Analytics project at inLab, FIB.

To develop the project R, MySQL, and ggplot2 (for visualization) were used. The installation of the corresponding software products (RStudio and Database Browser) was also done at this stage. This step did not suffer any deviation from the initial plan.

Then I did an online course “Introduction to R” offered by Microsoft through edx in order to learn R programming which I had to use to develop indicators. In the meantime, I also had to work for my GEP deliverables and presentation. Regarding GEP work, I had to wait for a week to get inscription of my project done before beginning my GEP work. Following this, planning was done in consultation with my thesis director on definition of indicators which was going to be developed. This was done in the last week of March after Easter holidays.

After that I started to work on day to start on (starting date) module which I developed for data of secondary school students across 6 schools and 15 different subjects for the previous academic year. Then, new data was obtained and because of MOODLE update new tables were added. Hence, I had to rework on day to start on
module which took one week additional to complete. Similarly, I calculated agility rate once I computed day to start on and then rework had to be done. I was able to complete till agility rate step at the end of April.

For other indicators, from time spent module till competitive level, I spent the exact amount of time I intended to spend but I had to begin this work one week later because of the problem stated in the previous paragraph.

After all those modules were completed I had limited amount of time so I started to work on Principal Component Analysis (PCA) and dashboard design was removed from my project scope and was assigned to another team member who had joined. Although, PCA is a part of validation, individual modules had to be tested for its robustness. The documentation work was started previously by GEP but for TFG report I had to dedicate more time. The documentation work I started at the beginning of June and completed in the middle of the same month.
To make sure that this project gets completed on time, every week on Tuesday I have a weekly meeting with all the team members. The weekly workload to develop this TFG is 25 hours. And cost is also taken into account, as most of the software I have used are open source. Another important thing that might have delayed the progress of the project is the failure of hardware resources. Spare hardware resources were available to handle such situations.

### 6.3 Budget Planning

Budget planning is one among the most important phase of the project management. The main aim of this step is to provide an optimized budget for this project. Various expenses like software costs, hardware costs and human resource costs are studied and analysed to give a single value.

#### 6.3.1 Budget estimation

The overall expenses are divided among hardware, software and human resources. For the purpose of calculating amortized cost, the life time of the resources is considered. Since, the project has been completed in 5 months, the amortized cost is calculated by dividing the total cost by a special factor. This special factor is determined by dividing the useful life of the resource (in years) by 0.42 (5/12).
6.3.1.1 Hardware Budget

The hardware budget gives information about various hardware components used in the course of this project, its actual and amortized cost.

<table>
<thead>
<tr>
<th>S NO</th>
<th>HARDWARE COMPONENT</th>
<th>USEFUL LIFE (in years)</th>
<th>TOTAL COST (in €)</th>
<th>AMORTIZED COST (in €)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Personal Computer</td>
<td>4</td>
<td>1.000,00</td>
<td>105,00</td>
</tr>
<tr>
<td>2</td>
<td>Servers and disk space</td>
<td>8</td>
<td>2.500,00</td>
<td>131,23</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td></td>
<td>3.500,00</td>
<td>236,23</td>
</tr>
</tbody>
</table>

**Table 3 Hardware Budget**

6.3.1.2 Software Budget

The below software budget shows an estimate for the various software used in the project along with their estimated costs.

<table>
<thead>
<tr>
<th>S. NO.</th>
<th>SOFTWARE COMPONENT</th>
<th>USEFUL LIFE (in years)</th>
<th>TOTAL COST (in €)</th>
<th>AMORTIZED COST (in €)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Windows 8.1</td>
<td>1</td>
<td>150,00</td>
<td>63,00</td>
</tr>
<tr>
<td>2</td>
<td>Microsoft Word</td>
<td>2</td>
<td>70,00</td>
<td>14,70</td>
</tr>
<tr>
<td>3</td>
<td>Rstudio</td>
<td>1</td>
<td>NA (open source)</td>
<td>0,00</td>
</tr>
<tr>
<td>4</td>
<td>Database Browser</td>
<td>2</td>
<td>NA (open source)</td>
<td>0,00</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td></td>
<td>220,00</td>
<td>77,70</td>
</tr>
</tbody>
</table>

**Table 4 Software Budget**

6.3.1.3 Human Resources

The table below contains estimate of value of my work and value of two professors involved in the project.

<table>
<thead>
<tr>
<th>ROLE</th>
<th>COST / HOUR (in €)</th>
<th>HOURS / WEEK</th>
<th>TOTAL WEEKS</th>
<th>COST (in €)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scientist</td>
<td>20,00</td>
<td>25</td>
<td>17</td>
<td>8.500,00</td>
</tr>
<tr>
<td>Professors (2)</td>
<td>40,00</td>
<td>1</td>
<td>17</td>
<td>680,00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>60,00</td>
<td></td>
<td></td>
<td>9.180,00</td>
</tr>
</tbody>
</table>

**Table 5 Human Resources**
From the above pie chart (Figure 7), it can be concluded that most (97%) of the budget is allocated to human resources. Hardware and software budget accounts for 2% and 1% of total budget respectively.

6.4 SUSTAINABILITY

Sustainability is an important element in any project. This project is evaluated by means of three factors: economic, social and environmental.

6.4.1 Economic sustainability

Economic sustainability deals with using various strategies for employing existing resources optimally to ensure maximum benefit. From the estimation, it can be said that this has been the maximum bound on the budget for the project. The project involved mostly open source software and low-cost hardware equipment. Time was allocated for each task based on priority. Cost can be saved if this work done by me is reused for the upcoming version. The estimated cost exactly matched with the actual cost spent.

6.4.2 Social sustainability

This project aims at developing web portal for teachers across secondary schools in Catalunya. The final dashboard will give information about the learning pattern of the students so as a result the outcome of the project is going to improve
the standard of learning across secondary schools in Catalunya. Students who need extra help will be indicated to the teachers which is going to improve their learning.

6.4.3 Environmental sustainability

In the development of the project, a computer along with servers are going to be running. Computer, servers and paper to print the documentation are the resources going to be used.

If we assume that the energy utilized by the computer I have been using as 250W and for the 425 hrs. of work the energy expenditure will be 125KW which amounts to 48.125 kg of CO₂. This is a high amount although it is within permissible level. This can be reduced by reusing the existing code. Some algorithms developed has been reused and improved which has reduced development time and hence the cost.

6.4.4 Sustainability matrix

A sustainability matrix is prepared in accordance with guidelines provided by Christian Felber and scores are given for: Economic, Social and Environmental for the planning phase as shown in the table. The scores are assigned based on sustainability in each area. This reveals that the total score is 47 out of 60. Thus, the project is highly sustainable.

<table>
<thead>
<tr>
<th>SUSTAINABLE?</th>
<th>ECONOMIC</th>
<th>SOCIAL</th>
<th>ENVIRONMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLANNING</td>
<td>Economic viability</td>
<td>Improved quality of life</td>
<td>Resource analysis</td>
</tr>
<tr>
<td>assessment</td>
<td>9</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>OUTCOMES</td>
<td>Final costs vs forecast</td>
<td>Impact on social environment</td>
<td>Resource consumption</td>
</tr>
<tr>
<td>assessment</td>
<td>9</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>RISKS</td>
<td>Adapting to changes of scenery</td>
<td>Social damage</td>
<td>Environmental damage</td>
</tr>
<tr>
<td>assessment</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TOTAL</td>
<td>47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 6 SUSTAINABILITY MATRIX**
CHAPTER 7

DESIGN AND IMPLEMENTATION

This is the most important chapter which discusses the work done by me at inLab, FIB in detail. My TFG completely focuses on the data mining part of Learning Analytics project at inLab, FIB. Starting Date (Day to start on) along with 4 speed indicators (agility rate, time spent, transition time and speed delivery) and 3 intensity indicators (delivery rate, engagement level and competitive level) were developed by me. Other class of indicators (persistence and choice) were developed by A M Preethi as a part of her TFG. This section also contains visualisation of developed indicators.

7.1 PROJECT FLOW

The above figure (Figure 8) represents the overall flow of my TFG work. The elements coloured in blue represent a result whereas those coloured in yellow, red and black represent process, validation and database respectively. In MOODLE there are around 400 tables among those the necessary ones (34 tables) are loaded into Westeros Agora database. The indicators developed as a part of my TFG are mostly based on work done by Ivan Vukić as a part of his Masters’ thesis “Measurement of motivation of high school students for real-time tracking from the Virtual learning
environment” defended in June 2014 at UPC and Joanna Sykurska as a part of her Masters’ thesis “Data Mining in Learning Analytics” defended in February 2016 at UPC.

7.2 DYNAMIC UPDATES

All the modules developed are designed to perform dynamic updates every day. So, each indicator implemented in form of R script, will run every day from the beginning till the end of the academic year. This is more important because, motivation is very dynamic and has to be captured regularly.

7.3 STARTING DATE (DAY TO START ON)

This starting date is used for computation of the motivation indicators (like agility rate). Normally, teachers upload learning lessons and home work (quiz, assignment and hotpot) way before the scheduled time. Some of the time even before beginning of the academic year. Sometimes, they reuse the materials uploaded in previous academic year(s). So, as a result the starting date of the materials mostly contains irrelevant data. However, starting date of the learning materials are needed to perform data analysis and hence an approach to compute starting date of all the available learning materials is necessary.

The approach is to use the accesses of students to infer the actual day to start on. The idea is to detect that value when significant bulk of students accesses a given task in a short period of time. It is implemented in such a way that, for all the learning materials in a window of 7 days when 10% of the students enrolled in the respective courses accesses, then the first access time in that window is taken as starting date of that learning material.

```
SELECT mdl_logstore_standard_log.userid, timecreated as time, Number_of_students,
mdl_logstore_standard_log.courseid, contextinstanceid as cmid from
`mdl_logstore_standard_log` JOIN `mdl_role_assignments` ON
`mdl_role_assignments`.userid = `mdl_logstore_standard_log`.userid JOIN ENROLLED ON
ENROLLED.courseid = mdl_logstore_standard_log.courseid where (timecreated BETWEEN
1441576800 AND 1442181600 ) AND contextlevel = 70 AND roleid= 5 order by time;
```

**SOURCE CODE 1: MySQL query to extract log data from the Database**

The above source code (source code 1) extracts log data of 1 week period from September 7th 2015 till September 14th 2015 in order to compute starting date of learning materials. The above query selects log data of students (roleid of 5) who belong to modules context (contextlevel of 70).
**startingDate** <- function (logs)
{
  #PROCESSING DTSO FOR EACH LEARNING MATERIAL
  num_of_students_accessed <- length(unique(logs$userid))
  total_class <- logs[1,"Number_of_students"]
  if(ceiling((num_of_students_accessed / total_class) * 100) >= 10)
  {
    data <- min(logs,"time")
    startDate <- c(data, logs[1,"cmid"])
  }
  else
  {
    startDate <- c(NA, logs[1,"cmid"])
  }
  return startDate
}

SOURCE CODE 2: R function to compute starting date of each learning materials

The above R function (source code 2) takes log data in a window of 7 days of a particular learning material belonging to a particular subject and returns starting date depending on whether 10% of students had accessed in that window.

**NOTE**: For some of the learning materials starting date is provided by the teacher(s). Those values are stored in the database. For those which do not have starting date only the above source code (source code 2) is used to compute starting date.

The figure (Figure 9) below, shows the visualisation of starting date of learning materials plotted against learning modules (materials). The red dots represent first accesses of students and the indigo triangles represent starting date of the learning modules. It can be seen from the below figure, that few of the modules do not have starting dates because 10% of students have not accessed them in any window of 7 days. In some cases, it is computed a bit late than expected, as 10% of students have accessed them in a window of 7 days later only.
Figure 9: Starting Date Result for S4-Informàtica 2015/2016

- **DATE**
  - Jan
  - Apr
  - Oct

- **MODULES**

- **Legend**
  - Factor (cond)
    - **FIRST ACCESSES**
    - **START DATE**

- **Legend Details**
  - 0.3
  - First accesses
  - Start date

**Figure 9** Starting Date Result
SUMMARY OF THE MODULE DEVELOPED:

NAME: Day to start on (Starting date)

DESCRIPTION: Day to start on of a learning material is the point of time, from where a student can have access to the respective learning material for the given academic year.

FEATURES

CALCULATION DONE: From September 15th, 2015 till June 14th, 2016

FREQUENCY OF CALCULATION: Everyday

EXTRACTION CRITERIA: Students who have made access to the learning materials.

FORM OF REPRESENTATION: timestamp

EXAMPLE

Assume that there are 30 students in a class and there is a quiz uploaded by the teacher for which 3 students have managed to access it in a period of 7 days.

So, \((3/30)*100 = 10\) which implies that 10% of the students have accessed that quiz. So, the starting date will be the timestamp of first access of students in the period of 7 days.

NOTE: If 10% of the students do not access the learning materials in a window of 7 days then starting date for that learning material is NA (Not Available).

7.4 SPEED INDICATORS

7.4.1 Agility rate

Agility rate gives information of how fast a student accesses a learning material after it has been uploaded by the teacher. It is obvious that a student who accesses learning materials as early as possible is motivated to learn. This reflects speed of learning of a student. It is computed as the time difference between first access to the learning material by a student and the starting date of the corresponding learning material. Lesser the agility rate value, more is the student motivated.

There are 3 possible results of agility rate:

- Positive value, which means that the student has accessed the learning material after the starting date of the module is calculated.
• Zero, which means that the student has accessed the learning material concurrently when the teacher has uploaded the corresponding learning material. Practically, this will never happen as the students do not know when the teacher uploads.

• Negative value, which means that the student has accessed the learning material before the starting date of the module. This is possible because, starting date of a module is calculated only if 10% of the students make accesses in a window of 7 days. Most of the time, some students make access before the above active period. So, for them agility rate is negative.

```r
agilityRate <- function(date_info)
{
    date <- date_info
    rs = dbSendQuery(mydb, paste("SELECT mdl_logstore_standard_log.userid, MIN(timecreated) AS FIRSTACCESS, courseid, contextinstanceid as cmid FROM `mdl_logstore_standard_log` JOIN `mdl_role_assignments` ON `mdl_role_assignments`.userid = mdl_logstore_standard_log.userid where contextlevel = 70 AND roleid = 5 AND timecreated BETWEEN ", date," AND ", date + 86400," GROUP BY cmid, userid ORDER BY timecreated;", sep = ")")
    firstAccess = fetch(rs, n=-1)
    cmids <- unique(firstAccess$cmid)
    users <- unique(firstAccess$userid)
    rs = dbSendQuery(mydb, 'SELECT * FROM Balaji_FIRSTACCESS;')
    fa_table = fetch(rs, n=-1)
    i <- 1
    FA_update <- data.frame(userid = NA, FIRSTACCESS = NA, courseid = NA, cmid= NA)
    while(i <= length(users))
    {
        if(firstAccess[i,4] %in% fa_table[fa_table$userid==firstAccess[i,1],4])
        {
            #do nothing
        }
        else
        {
            FA_update <- rbind(FA_update, firstAccess[i,])
        }
        i <- i + 1
    }
    FA_update <- FA_update[-1,]
}
```
rs = dbWriteTable(mydb, name='Balaji_FIRSTACCESS', value = FA_update, append = TRUE, row.names=FALSE)

rs = dbSendQuery(mydb, 'SELECT * FROM DTSO;')
startdate = fetch(rs, n=-1)
i <- 1
while(i <= nrow(FA_update))
{
stime <- startdate[startdate$cmid==FA_update[i,4],1]
if(length(startdate[startdate$cmid==FA_update[i,4],1])>0)
  FA_update[i,2] <- (FA_update[i,2] - stime)/3600
else
  FA_update[i,2] <- NA
i <- i + 1
}
names(FA_update) <- c('userid', 'AGILITYRATE', 'courseid', 'cmid')
if(nrow(FA_update)>0)
{
  FA_update$Date <- date
}
rs = dbWriteTable(mydb, name='Balaji_AGILITYRATE', value = FA_update, append = TRUE, row.names=FALSE)

**SOURCE CODE 3: R function to compute agility rate for each student corresponding to each learning module**

The above R function (source code 3) takes timestamp of the date for which agility rate has to be computed. First, the first access to a leaning material made by the students are tracked every day and updated in a table. Whenever first access is made agility rate is calculated (the time difference between first access to the learning material by a student and the starting date of the corresponding learning material). In case, if starting date is not available then agility rate will not be calculated and updated in the table. Agility rate can be negative because during the development, day to start on is computed from beginning till the end of the academic year and then agility rate is computed for the same period. When deployed, both will run every day concurrently and there won’t be negative values. For this pilot version, negative value gives more information than NAs which is crucial for PCA (Principal Component Analysis).
Figure 10: Agility Rate for Socials 4t

Graph showing agility rate in days across different modules with various factors and types indicated.
NOTE: THE red line represents the median while the blue line represents a particular student (student id: 1152321).

The above figure (Figure 10) represents the output of agility rate for the subject Socials 4t. Various modules uploaded by the teacher ordered by starting date is plotted on X-axis and the agility rate for students are plotted on Y-axis. So, the above scatter plot is agility rate of students enrolled in the subject Socials 4t for each learning module. It can be found from the above figure that, the student marked do not have agility rate for all the modules. This is because of the fact that he/she has not even made even a single access to the corresponding module. The triangle shaped points the plot represent outlier points. Points which have agility rate value more (10*3rd quartile) value are considered as positive outliers and the points which have agility rate value less than (-10*1st quartile) value are considered as negative outliers.

SUMMARY OF THE MODULE DEVELOPED:

NAME: Agility rate

DESCRIPTION: Agility rate of a learning material is the amount of time a student takes to make first access to a learning module (material).

FEATURES

CALCULATION DONE: From September 15\textsuperscript{th}, 2015 till June 14\textsuperscript{th}, 2016

FREQUENCY OF CALCULATION: Everyday

EXTRACTION CRITERIA: Students who have made first access to the learning materials.

FORM OF REPRESENTATION: days

EXAMPLE

Assume that there is a student Jim who access a URL uploaded by his teacher.

Upload time: 15\textsuperscript{th} September 2015 8:00 AM (1442296800)

Access time of Jim: 18\textsuperscript{th} September 2015 8:00 PM (1442599200)

Agility rate = (1442599200 - 1442296800) / (3600*24) = 3.5 days.

This means that Jim has taken exactly 3 and a half days to access the URL for the 1\textsuperscript{st} time.
7.4.2 Time spent

Time spent measures the total amount of time a student spends on VLE (Virtual Learning Environment) to complete mandatory tasks which are quiz, assignment and hotpots. In the data available, it is not possible to differentiate whether a submission is successful or not. Most of the time students perform more than one submission. So, all the time a student spends on mandatory tasks is taken for computing time spent indicator.

Sometimes, there can be a huge time gap between two accesses of a student so the time difference between the last attempt and the first attempt cannot give the correct value. So, a different approach is followed. If a student has a time gap between two consecutive attempts as more than 30 minutes then the timeout is set to 30 minutes, else the difference value is used to compute time spent. Finally, the time spent will be the sum of all the time spent attempts.

This indicator is computed for each course module only (updated every day). But, for the purpose of performing PCA it needs to be calculated for each day. In order to incorporate reusability, every day time spent data is aggregated for each student belonging to each course module.

```r
timespent <- function(date_info)
{
    date <- date_info
    rs = dbSendQuery(mydb, paste("SELECT userid, 
        timecreated AS time,
        action,
        'mdl_logstore_standard_log' . courseid,
        contextinstanceid AS cmid,
        mo.name AS module_type,
        CONCAT('mdl_',mo.name) AS `table`
    FROM `mdl_logstore_standard_log`
    INNER JOIN `mdl_course_modules` cm ON cm.id = `mdl_logstore_standard_log`.contextinstanceid
    INNER JOIN mdl_modules mo ON cm.module = mo.id
    
    # Function to calculate time spent
    time_spent <- function(r, date)
    {
        time_diff <- r[time] - r[timecreated]
        if (time_diff > 30) {
            timeout <- 30
        } else {
            timeout <- time_diff
        }
        return(timeout)
    }
    
    # Apply function to each row
    time_spent_list <- lapply(rs, time_spent, date)
    
    # Sum up the time spent
    total_time_spent <- sum(time_spent_list)
    
    return(total_time_spent)
}
```
WHERE (`mdl_logstore_standard_log`.eventname LIKE '%hotpot%' OR `mdl_logstore_standard_log`.eventname LIKE '%quiz%' OR `mdl_logstore_standard_log`.eventname LIKE '%assign%') AND timecreated BETWEEN ",date ," AND ",date + 86400," ORDER BY userid, cmid, time, action;", sep=""))
values = fetch(rs, n=-1)
mdat <- data.frame(course=NA,cmid=NA,userid=NA,timespent=NA)
colnames(mdat) <- c("course", "cmid", "userid", "timespent")
i <- 1
while (i <= nrow(values)-1)
{
  subd <- values[values$cmid == values$cmid[i] & values$userid == values$userid[i],]
  ii <- 1
tmp <- 0.0
  for(ii in 1:nrow(subd))
  {
    if(nrow(subd)>1)
    {
      if(subd$time[ii+1] == subd$time[ii])
      {
        tmp <- tmp + 0.0
      }
      if(abs(subd$time[ii+1] - subd$time[ii]) > 1800)
      {
        tmp <- tmp + 1800
      }
    }
else
{
    tmp <- abs(subd$time[ii+1] - subd$time[ii]) + tmp
}
if(ii == nrow(subd)-1)
    break
}

ttt <-
data.frame(course=subd$course[1],cmid=subd$cmid[1],userid=subd$userid[1],timespent=tmp)
mdat <- rbind(mdat, ttt)
tmp <- 0.0
i <- i + nrow(subd)

if(nrow(values)-1 > 0)
{
    mdat <- mdat[-1,]
    mdat$Date <- date
    rs = dbWriteTable(mydb, name='Balaji_TIME_SPENT_PCA', value = mdat, append = TRUE, row.names=FALSE)
}

SOURCE CODE 4: R function to compute time spent for each student (for PCA)

The above R function (source code 4) takes date (timestamp) for which time spent has to be calculated. It gives how much time a student spends on a course module each day. For performing PCA, the amount of time a student spends on mandatory tasks is required. Only if the logs stating that a student makes attempt is present, then only time spent is calculated else it is not.
SELECT course, cmid, userid, timespent FROM `Balaji_TIME_SPENT_PCA` GROUP BY userid, cmid;

The above query aggregates time spent so that every student will have the total time spent for each course module associated.
TIME SPENT FOR 400 Informàtica 4t ESO. Projecte mSchools. (català)

Figure 11 TIME SPENT OUTPUT
The above figure (Figure 11) represents the output of time spent for the subject S4 Informatica. The mandatory task modules are represented by numbers which is the order in which it is assigned by the teacher. That forms the X-axis and Y-axis which represents time spent in mandatory tasks by students. The red line represents the median time spent whereas the blue line represents time spent of a particular student.

**SUMMARY OF THE MODULE DEVELOPED:**

**NAME:** Time spent

**DESCRIPTION:** Time spent measures how much time a student spends in completing mandatory tasks.

**FEATURES**

**CALCULATION DONE:** From September 15\textsuperscript{th}, 2015 till June 14\textsuperscript{th}, 2016

**FREQUENCY OF CALCULATION:** Everyday

**EXTRACTION CRITERIA:** Students who have made attempts to a mandatory task.

**FORM OF REPRESENTATION:** minutes

**EXAMPLE**

Assume that there is a student Cooper who spends time to complete an assignment as follows.

1\textsuperscript{st} attempt – time spent: 5 min

2\textsuperscript{nd} attempt – time spent: 10 min

3\textsuperscript{rd} attempt – time spent: 3 minutes

Time spent = 5 + 10 + 3 = 18 minutes.

This means that Cooper has spent 18 minutes for this assignment.

7.4.3 Speed delivery

Speed delivery shows the time gap been deadline of a mandatory task and submission time of the student corresponding to that task. So, if a student makes submission long before the deadline he/she will have high value of speed delivery. On
the other hand, if the student makes submission after the deadline period then he/she will have very low value which will be negative. This indicator reflects speed because a student can submit a task only if he/she acts fast. If a student consistently submits mandatory tasks after the deadline then it will be apparent to the teacher via visualisation in the dashboard and corrective measure can be taken. This indicator was proposed by myself. Initially, I found out that there are no indicators proposed or defined that captures the vital information whether a student submits on time or not. So, I proposed to capture this information by indicator called consistency which measures the percentage of mandatory tasks a student submits before the deadline. Later it was modified to become speed delivery.

For computing this indicator, deadline of the mandatory tasks was needed. There were 3 mandatory tasks: quiz, assignment and hotpot. For assignments and hotpots, deadline information was not available. For very few quizzes, deadline information was available. This issue was discussed and a decision was made. The decision is that for all the mandatory tasks that do not have deadline, a time period of 7 days from the starting date of the respective modules needs to be taken as deadline.

```r
speed_delivery <- function(date_info) {
  date <- date_info
  rs = dbSendQuery(mydb, paste("SELECT timecreated as SPEED_DELIVERY, "
  "mdl_logstore_standard_log".userid, courseid, contextinstanceid as cmid FROM
  "mdl_logstore_standard_log" join "mdl_role_assignments" ON
  "mdl_role_assignments".userid = "mdl_logstore_standard_log".userid where roleid = 5 AND
  contextlevel = 70 AND (eventname like '%hotpot%' OR eventname like '%quiz%' OR
  eventname like '%assign%') AND action = 'submitted' and timecreated BETWEEN ", date, "
  AND ", date + 86400, " ORDER BY SPEED_DELIVERY;", sep=""))

  submissionData = fetch(rs, n=-1)
  if(nrow(submissionData)>0) {
    rs = dbSendQuery(mydb, "select * from \'Balaji_deadLine\';")
    deadLineData = fetch(rs, n=-1)

    i <- 1
    while(i <= nrow(submissionData)) {
      deadline <- deadLineData[deadLineData$cmid==submissionData[i,4], 1]
      if(length(deadline) > 0) {
      }
  ```

  43
submissionData[i,1] <- {(deadline - submissionData[i,1])/3600)
} else {
    print('RARE CONDITION WHERE DEADLINE IS NA BECAUSE STARTING DATE IS NA')
    submissionData[i,1] <- NA
} i <- i + 1
}

submissionData$Date <- date
rs = dbWriteTable(mydb, name='Balaji_SPEED_DELIVERY', value = submissionData,
append = TRUE, row.names=FALSE)

SOURCE CODE 5: R function to compute speed delivery for each student

The above R function (source code 5) computes the speed delivery of students. The above function computes speed delivery only when a submission is made by any of the student(s). The function takes the date information in form of timestamp for which the speed delivery is calculated. Deadline is computed most of the time in the program with reference to the problem mentioned in this section previously. Even then sometimes deadline cannot be calculated as it is computed from the starting date of the modules. If starting date is not available then speed delivery will be NA.

The figure below (Figure 12) represents the output of speed delivery for the subject S4 Informatica. The mandatory task modules are represented by numbers which is the order in which it is assigned by the teacher. That forms the X-axis and Y-axis which represents speed delivery of students. The red line represents the median speed delivery whereas the blue line represents speed delivery of a particular student. It can be observed the most of the students submit after the deadline as the median is negative in some cases. This is because the deadline information is not provided and a decision had to be made to compute this indicator.
NOTE: THE red line represents the median while the blue line represents a particular student.

SUMMARY OF THE MODULE DEVELOPED:

NAME: Speed delivery

DESCRIPTION: Speed delivery measures how fast a student submits a task (mandatory) which can be before or after the deadline.

FEATURES

CALCULATION DONE: From September 15\textsuperscript{th}, 2015 till June 14\textsuperscript{th}, 2016

FREQUENCY OF CALCULATION: Everyday

EXTRACTION CRITERIA: Students who have made submission to a mandatory task.

FORM OF REPRESENTATION: hours

EXAMPLE

Assume that there is a student Tom who submits an assignment uploaded by his teacher.

Deadline time: 15\textsuperscript{th} October 2015 9:30 AM (1444894200)

Submission time of Tom: 7\textsuperscript{th} October 2015 11:00 PM (1444251600)

Speed delivery = (1444894200 - 1444251600) / (3600) = 178.50 hours.

This means that Tom has submitted the assignment 178.50 hours prior to the deadline.

7.4.4 Transition time

Transition time gives the information of how much time a student takes to make transition from one task to another. It is computed for each task of a subject and later aggregated by taking mean of all the tasks that belong to a particular subject. If a student makes transition from any of the task(s) that belong to any of the subject(s) he/she is enrolled in, then the time difference is assigned to the next task he/she accesses. This indicator may not be able to measure motivation of the students but it can definitely be used to understand the learning style of the students. This reflects speed of learning of the students.
transitionTime <- function(tsv) # tsv => timestamp of the date of calculation
{
  rs = dbSendQuery(mydb, paste("SELECT ASD.time AS time, ASD.userid, contextinstanceid as cmid, ASD.courseid, ASD.action FROM(SELECT timecreated as time, userid, action, contextinstanceid, courseid FROM `mdl_logstore_standard_log` WHERE timecreated between ", tsv, " AND ", tsv + 86400," AND contextlevel = 70 ORDER BY userid, time, contextinstanceid)AS ASD JOIN `mdl_role_assignments` ON `mdl_role_assignments`.userid = ASD.userid WHERE roleid = 5 ORDER BY userid, time, cmid;",sep=""))
  trans_time = fetch(rs, n=-1)
  date <- as.Date(as.POSIXct(trans_time[1,1], origin="1970-01-01"))
  users <- unique(trans_time$userid)
  upto <- length(users)
  j <- 1
  trant <- data.frame(transitionTime = NA, userid = NA, cmid = NA, courseid = NA)
  while(j <= upto)
  {
    activity <- trans_time[trans_time$userid==users[j],]
    i <- 1
    if(nrow(activity)==1)
    {
      tmp <- data.frame(transitionTime = NA, userid = NA, cmid = NA, courseid = NA)
    }
    while(i < nrow(activity))
    {
      if(activity$cmid[i] == activity$cmid[i+1])
      {
        #Do nothing
      }
      else
      {
      
    
```
SOURCE CODE 6: R function to compute transition time for each student and for each subject every day

The above R function (source code 6) is used to compute transition time every day. It computes value only if a student makes transition from one activity to another. The transition can happen within the same subject as well as from some other subject. Finally, the values are aggregated in such a way that on a given date a student will have transition time for that particular course alone (by means of average) he/she is enrolled in.
The figure below (Figure 13) shows the visualisation of transition time of students enrolled in Socials 4t. Transition time for a given day cannot be more than 24 hrs. Points whose values are greater than \((10 \times 3^{rd\text{ quartile}})\) value of the indicator are filtered out and fixed firmly outside the limit with a different shape as outliers. It can be found out that the student highlighted takes less time to make transition when compared with the median. It is not necessary that students should have transition time every day, as moving to another task / subject after doing some work is a matter of choice and convenience.
FIGURE 13 TRANSITION TIME OUTPUT
SUMMARY OF THE MODULE DEVELOPED:

NAME: Transition time

DESCRIPTION: Transition time measures how quickly a student moves to another task / an activity after doing a task / an activity in a single day.

FEATURES

CALCULATION DONE: From September 15th, 2015 till June 14th, 2016

FREQUENCY OF CALCULATION: Everyday

EXTRACTION CRITERIA: Students who have made transition from one task to another.

FORM OF REPRESENTATION: minutes

EXAMPLE

Assume that there is a student Joseph who accesses a URL, then submits a quiz and finally submits a questionnaire all belonging to the same subject say Mathematics.

URL access time: 8:30 AM
Quiz access time: 11:30 AM
Questionnaire access time: 1:30 PM

Transition time = (3 (from 8:30 AM to 11:30 AM) + 2 (from 11:30 AM to 1:30 PM)) / (3)
= 5/3 = 1.67 hours = 100 minutes

This means that transition time of Joseph is 100 minutes.

7.5 INTENSITY INDICATORS

7.5.1 Delivery rate

Delivery rate gives the information of whether a student has submitted the mandatory task which is assigned to him. It is computed in a window of 7 days. It will show what proportion of the assigned mandatory tasks a student has completed in a period of 7 days. This reflects intensity of effort of the student(s). It is computed as the ratio between the number of mandatory tasks a student has completed in a period of 7 days to the number of mandatory tasks assigned (active ones for which deadline is not over) in until the end of 7 days.
Possible cases:

- A student submits a mandatory task for which deadline is not over. For this case, delivery rate is computed normally as per the definition.
- A student submits a mandatory task for which starting date is not known and hence deadline may not be available. So for this case, student will have delivery rate greater than one if he is assigned at least 1 other task. Otherwise, it will be like the student submitted one task where he is assigned nothing (1/0) which is NaN (Not a Number).
- A student submits a mandatory task after the deadline. For this case, the delivery rate of the student will be NA.
- A student does not submit a mandatory task which is active (deadline is not over). For this case, delivery rate of the student will be zero.

def del_rate <- function(d) {
  date <- d
  rs = dbSendQuery(mydb, paste("SELECT userid, courseid, cmid, time FROM(SELECT FROM_UNIXTIME(MIN(timecreated)) as time, mdl_logstore_standard_log.userid, eventname, action, mdl_logstore_standard_log.courseid, contextinstanceid as cmid from `mdl_logstore_standard_log` join mdl_role_assignments on mdl_role_assignments.userid = mdl_logstore_standard_log.userid where contextlevel=70 AND roleid=5 AND (eventname like '%assign%' OR eventname like '%quiz%' OR eventname like '%hotpot%') AND action = 'submitted' AND timecreated between ( ",(date - 604800 )," ) AND (", date + 86400 ," ) group by cmid, userid order by userid, courseid;"", sep=""'))
  TEMP_NUMER = fetch(rs, n=-1)
  rs = dbWriteTable(mydb, name='Balaji_numertmp', value = TEMP_NUMER, overwrite = TRUE, row.names= F)
  rs = dbSendQuery(mydb, "SELECT userid, courseid, cmid, count(*) as numerator FROM (SELECT userid, Balaji_numertmp.courseid, time, Balaji_numertmp.cmid, FROM_unixtime(deadLine), moduletype FROM `Balaji_numertmp` JOIN `Balaji_deadLine` ON `Balaji_deadLine`.cmid = Balaji_numertmp.cmid WHERE UNIX_TIMESTAMP(time) < DEADLINE)AS ASD group by ASD.userid, ASD.courseid")
  numerator = fetch(rs, n=-1)
  
  rs = dbSendQuery(mydb, paste("SELECT userid, courseid, count(*) as denom FROM(SELECT userid, FROM_UNIXTIME(unix_timestamp(startTime)) as starttime, deadLine, moduleName, type,/courseName, ASD.cmid, ASD.courseid FROM (SELECT Balaji_course_enroll_info1.USERID, Balaji_course_enroll_info1.COURSEID,","

52
FROM_UNIXTIME(deadLine) as deadLine, moduleName, cmid, moduleType FROM
`Balaji_course_enroll_info1` JOIN `Balaji_deadLine` ON
`Balaji_course_enroll_info1`.courseid = `Balaji_deadLine`.courseID)AS ASD JOIN
`Balaji_startDate` on `Balaji_startDate`.cmid = ASD.cmid WHERE unix_timestamp(deadLine) > ",(date - 604800)," AND unix_timestamp(startTime) < " , date + 86400," ORDER BY
STARTTIME)as asd1 group by asd1.userid, asd1.courseid;(;sep=""))

denom = fetch(rs, n=-1)
if(nrow(denom)>0)
{
    denom$delRate <- NA
    denom$numer <- NA
}
i <- 1
upto <- nrow(denom)
while(i <= upto)
{
    if(denom[i,$]$denom > 0)
    {
        match <- numerator[numerator$userid == denom[i,$$USERID &
numerator$courseid==denom[i,$$COURSEID,]
        match <- na.omit(match)
        if(nrow(match)>1)
            print(match)

        if(nrow(match)>0)
            denom[i,$$delRate <- match[1,$$numerator / denom[i,$$denom
            denom[i,$$numer <- match[1,$$numerator
    }
    i <- i + 1
}
}

if(nrow(denom)>0)
{
    denom$Date <- date
    rs = dbWriteTable(mydb, name='Balaji_d_rate', value = denom, append = TRUE,
    row.names= F)
}

SOURCE CODE 7: R function to compute delivery rate for each student enrolled in each subject
The above R function (source code 7) takes the date information in timestamp for which delivery rate has to be computed. First, for each student the number of submissions he/she has made before the deadline in each subject(s) is calculated which is the numerator of the final computation. Then, for each student he/she is enrolled in, the number of task(s) for which deadline period is not over in the window of 7 days is computed which is the denominator. And the final result will be, the numerator divided by the denominator value for each student enrolled in each course.

The figure below (Figure 14) represents the final visualisation output of delivery rate for the students enrolled in the subject Informatica. The median delivery rate and delivery rate of a particular student throughout the year is represented by the red and blue points as well as lines. It gives information of how much part of the tasks a student has completed among those assigned to him. If there is no point associated for a date, then it means that no task is active during that period. Delivery rate of 1 means that the student has completed all the task(s) assigned to him.
Figure 14: Delivery Rate for 400 Informàtica (català)
SUMMARY OF THE MODULE DEVELOPED:

**NAME:** Delivery rate

**DESCRIPTION:** Delivery rate measures what part of the assigned mandatory task(s) a student has completed among those assigned to him/her.

**FEATURES**

**CALCULATION DONE:** From September 15\(^{th}\), 2015 till June 14\(^{th}\), 2016

**FREQUENCY OF CALCULATION:** Everyday (in a window of 7 days)

**EXTRACTION CRITERIA:** Students who are assigned a mandatory task.

**FORM OF REPRESENTATION:** No unit (just a number)

**EXAMPLE**

Assume that there is a student Mark who submits 3 quizzes uploaded by his teacher.

Deadline time: 5\(^{th}\) December 2015 9:00 AM

Submission time of Tom: 27\(^{th}\) November 2015 10:00 AM

Number of quizzes submitted my Mark: 2

Delivery rate = 2/3 = 0.66 (for a period of 7 days until 5\(^{th}\) December provided he did not submit the 3\(^{rd}\) quiz and no new mandatory task is assigned to him).

This means that Mark has completed 66.67% of the tasks assigned to him.

7.5.2 Engagement level

Engagement level shows how active a student is on a given day with respect to his best activeness in the window of 14 days (Past 13 days + day of computation). It reflects the intensity of effort of the student. It is computed every day for each student enrolled in each subject irrespective of whether the tasks are mandatory or not. It is computed as the ratio of total number of activities performed by a student in a given day to the maximum of number of activities performed by the student in the window of 14 days.

```r
eng_level <- function(da)
{
  date <- da
```
rs = dbSendQuery(mydb, paste("SELECT * FROM(SELECT DATE(time) AS DATE, userid, courseid, count(*) as Access FROM(SELECT from_unixtime(timecreated) as time, mdl_logstore_standard_log.userid, eventname, component, action, courseid, contextinstanceid as cmid FROM `mdl_logstore_standard_log` join `mdl_role_assignments` ON `mdl_role_assignments`.userid = mdl_logstore_standard_log.userid WHERE contextlevel = 70 AND roleid = 5 AND timecreated between ", date," AND ",date + 86400," ORDER BY timecreated, courseid, userid)as asd GROUP BY courseid, userid order by USERID, COURSEID)as fir LEFT OUTER JOIN `Balaji_course_enroll_info` ON (`Balaji_course_enroll_info`.userid = fir.userid AND `Balaji_course_enroll_info`.courseid = fir.courseid) UNION SELECT * FROM(SELECT DATE(time) AS DATE, userid, courseid, count(*) as Access FROM(SELECT from_unixtime(timecreated) as time, mdl_logstore_standard_log.userid, eventname, component, action, courseid, contextinstanceid as cmid FROM `mdl_logstore_standard_log` join `mdl_role_assignments` ON `mdl_role_assignments`.userid = mdl_logstore_standard_log.userid WHERE contextlevel = 70 AND roleid = 5 AND timecreated between ", date," AND ",date + 86400," ORDER BY timecreated, courseid, userid)as asd GROUP BY courseid, userid order by USERID, COURSEID)as fir RIGHT OUTER JOIN `Balaji_course_enroll_info` ON (`Balaji_course_enroll_info`.userid = fir.userid AND `Balaji_course_enroll_info`.courseid = fir.courseid);", sep=""))

engage = fetch(rs, n=-1)

e <- engage[,c(2,3)]

e$DATE <- e[1,1]
e[is.na(e)] <- 0
e <- e[order(e$courseid, e$userid),]

personal_best <- data.frame(DATE=NA, Access=NA, userid=NA, courseid=NA)
personal_best <- rbind(personal_best, e)
personal_best <- personal_best[-1,]

i <- 1
d <- date - 86400
while(i <= 13)
{
  rs = dbSendQuery(mydb, paste("SELECT * FROM(SELECT DATE(time) AS DATE, userid, courseid, count(*) as Access FROM(SELECT from_unixtime(timecreated) as time, mdl_logstore_standard_log.userid, eventname, component, action, courseid, contextinstanceid as cmid FROM `mdl_logstore_standard_log` join `mdl_role_assignments` ON `mdl_role_assignments`.userid = mdl_logstore_standard_log.userid WHERE contextlevel = 70 AND roleid = 5 AND timecreated between ", date," AND ",date + 86400," ORDER BY timecreated, courseid, userid)as asd GROUP BY courseid, userid order by USERID, COURSEID)as fir LEFT OUTER JOIN `Balaji_course_enroll_info` ON (`Balaji_course_enroll_info`.userid = fir.userid AND `Balaji_course_enroll_info`.courseid = fir.courseid) UNION SELECT * FROM(SELECT DATE(time) AS DATE, userid, courseid, count(*) as Access FROM(SELECT from_unixtime(timecreated) as time, mdl_logstore_standard_log.userid, eventname, component, action, courseid, contextinstanceid as cmid FROM `mdl_logstore_standard_log` join `mdl_role_assignments` ON `mdl_role_assignments`.userid = mdl_logstore_standard_log.userid WHERE contextlevel = 70 AND roleid = 5 AND timecreated between ", date," AND ",date + 86400," ORDER BY timecreated, courseid, userid)as asd GROUP BY courseid, userid order by USERID, COURSEID)as fir RIGHT OUTER JOIN `Balaji_course_enroll_info` ON (`Balaji_course_enroll_info`.userid = fir.userid AND `Balaji_course_enroll_info`.courseid = fir.courseid);", sep=""))

engage = fetch(rs, n=-1)

e <- engage[,c(2,3)]

e$DATE <- e[1,1]
e[is.na(e)] <- 0
e <- e[order(e$courseid, e$userid),]

personal_best <- data.frame(DATE=NA, Access=NA, userid=NA, courseid=NA)
personal_best <- rbind(personal_best, e)
personal_best <- personal_best[-1,]

i <- 1
d <- date - 86400
while(i <= 13)
{...
= 70 AND roleid = 5 AND timecreated between ", d," AND ",d + 86400," ORDER BY
timecreated, courseid, userid)as asd GROUP BY courseid, userid order by USERID,
COURSEID)as fir LEFT OUTER JOIN `Balaji_course_enroll_info` ON
(`Balaji_course_enroll_info`.userid = fir.userid AND `Balaji_course_enroll_info`.courseid =
firm.courseid) UNION
SELECT * FROM(SELECT DATE(time) AS DATE, userid, courseid, count(*) as Access
FROM(SELECT from_unixtime(timecreated) as time, mdl_logstore_standard_log.userid,
eventname, component, action, courseid, contextinstanceid as cmid FROM
`mdl_logstore_standard_log` join `mdl_role_assignments` ON
`mdl_role_assignments`.userid = mdl_logstore_standard_log.userid WHERE contextlevel =
70 AND roleid = 5 AND timecreated between ", d," AND ",d + 86400," ORDER BY
timecreated, courseid, userid)as asd GROUP BY courseid, userid order by USERID,
COURSEID)as fir RIGHT OUTER JOIN `Balaji_course_enroll_info` ON
(`Balaji_course_enroll_info`.userid = fir.userid AND `Balaji_course_enroll_info`.courseid =
firm.courseid);", sep="")
engage_tmp = fetch(rs, n=-1)
et <- engage_tmp[-c(2,3)]
date_info <- as.POSIXct(d, origin="1970-01-01")
et$DATE <- date_info
et[is.na(et)] <- 0
et <- et[order(et$courseid, et$userid),]

personal_best <- rbind(personal_best,et)

d <- d - 86400
i <- i + 1

upto <- nrow(e)
j <- 1
e$engagementLevel <- NA

while(j <= upto)
{
  uid <- e[j,3]
cid <- e[j,4]
m <- max(personal_best[personal_best$userid==uid &
personal_best$courseid==cid]$Access)
e[j]$engagementLevel <- e[j]$Access / m
j <- j + 1
}
e <- replace(e, is.na(e), 0)
rs = dbWriteTable(mydb, name='Balaji_engagementLevel1', value = e, append = TRUE, row.names= F)
}

**SOURCE CODE 8: R function to compute engagement level for each student enrolled in each subject**

The above R function (source code 8) computes engagement level for the date (timestamp) supplied as input for each student enrolled in each course(s). First, the total number of activities performed by a student in a given subject on the given day is computed. Finally, the number of activities performed on that day divided by the maximum value in the 14 days (past 13 days + day of computation) period gives the engagement level.

**NOTE:** If a student has neither performed any activity in a day nor on the past 13 days then 0 / 0 gives NaN (Not a Number). For this case, NaN is replaced by 0 because the student has done nothing in the 14 days which means he/she is not engaged at all.

The figure below (Figure 15) represents engagement level of students enrolled in the subject Socials 4t. It can be concluded from the figure that most of the students (median) perform activities only few days (may be because of deadline). The student highlighted in the graph has performed activities only on 2 days in the whole academic year. The purpose of the Learning Analytics project at inLab, is to identify such students and alert the respective teacher as early as possible. If a student performs activities for the first time in the 14 days period then engagement level will be maximum (1) on that day alone. It will vary from then onwards depending on the number of activities he/she performs in the subsequent days.
FIGURE 15 ENGAGEMENT LEVEL
SUMMARY OF THE MODULE DEVELOPED:

NAME: Engagement level

DESCRIPTION: Engagement level measures how engaged (active) a student is on a given subject in comparison to his best engagement in the period of 14 days.

FEATURES

CALCULATION DONE: From September 15th, 2015 till June 14th, 2016

FREQUENCY OF CALCULATION: Everyday (in a window of 14 days)

EXTRACTION CRITERIA: Students who have performed any type of activities on any of the tasks.

FORM OF REPRESENTATION: No unit (just a number)

EXAMPLE

Assume that there is a student John who performed 30 activities on a given day.

Date of engagement: 15th December 2015

Best engagement in the period of 14 days (from 8th December 2015 till 15th December 2015): 40

Engagement level = 30/40 = 0.75

This means that John’s engagement level is 75% on 15th December 2015.

7.5.3 Competitive level

Competitive level captures similar information to that captured by engagement level. The reason is that it shows how active a student is on a given day with respect to the most active student in a window of 7 day (past 6 days + day of computation). This means that every day there need not be a student with competitive level of 100%. It also reflects the intensity of effort of the student.

```r
clLevel <- function(da)
{
  d <- da
  competitive_level <- data.frame(DATE=NA, Access=NA, userid=NA, courseid=NA)
  rs = dbSendQuery(mydb, paste("SELECT * FROM(SELECT DATE(time) AS DATE, userid, courseid, count(*) as Access FROM(SELECT from_unixtime(timecreated) as time, *
```

```r
FROM(SELECT from_unixtime(timecreated) AS time, Access, userid, courseid FROM Access)
)
) AS DATE, Access, userid, courseid) AS competitive_level
)
)
```
mdl_logstore_standard_log.userid, eventname, component, action, courseid,
contextinstanceid as cmid FROM `mdl_logstore_standard_log` join `mdl_role_assignments`
ON `mdl_role_assignments`.userid = mdl_logstore_standard_log.userid WHERE contextlevel = 70 AND roleid = 5 AND timecreated between ", d," AND ",d + 86400," ORDER BY
timecreated, courseid, userid)as asd GROUP BY courseid, userid order by USERID,
COURSEID)as fir LEFT OUTER JOIN `Balaji_course_enroll_info` ON
(`Balaji_course_enroll_info`.userid = fir.userid AND `Balaji_course_enroll_info`.courseid =
fir.courseid) UNION
SELECT * FROM(SELECT from_unixtime(timecreated) as time,
mdl_logstore_standard_log.userid, eventname, component, action, courseid,
contextinstanceid as cmid FROM `mdl_logstore_standard_log` join `mdl_role_assignments`
ON `mdl_role_assignments`.userid = mdl_logstore_standard_log.userid WHERE contextlevel = 70 AND roleid = 5 AND timecreated between ", d," AND ",d + 86400," ORDER BY
timecreated, courseid, userid)as asd GROUP BY courseid, userid order by USERID,
COURSEID)as fir RIGHT OUTER JOIN `Balaji_course_enroll_info` ON
(`Balaji_course_enroll_info`.userid = fir.userid AND `Balaji_course_enroll_info`.courseid =
fir.courseid);", sep="")

compet_tmp = fetch(rs, n=-1)
et <- compet_tmp[,c(2,3)]
date_info <- as.POSIXct(d, origin="1970-01-01")
et$DATE <- date_info
et[is.na(et)] <- 0
et <- et[order(et$courseid, et$userid),]
competitive_level <- rbind(competitive_level,et)
competitive_level <- competitive_level[-1,]
cl=data.frame(DATE = NA, Access = NA, userid=NA, courseid=NA)
i <- 1
upto <- nrow(competitive_level[competitive_level$DATE==competitive_level[1,]$DATE,])
while(i <= upto)
{
  s <- sum(competitive_level[competitive_level$userid==competitive_level[i,3] &
  competitive_level$courseid==competitive_level[i,4],]$Access)
```r
tmp <- data.frame(DATE = date, Access = s, userid = competitive_level[i,3], courseid = competitive_level[i,4])
cl <- rbind(cl,tmp)
i <- i + 1
}
cl <- cl[-1,]
competitive_level <- cl
rm(cl)
courses <- unique(competitive_level$courseid)
competitor <- data.frame(courseid = NA, Accesses = NA)
i <- 1
while(i <= length(courses))
{
cid <- courses[i]
maxv <- max(competitive_level[competitive_level$courseid==cid]$Access)
tmp1 <- data.frame(couid=cid, acc=maxv)
names(tmp1) <- c('courseid', 'Accesses')
competitor <- rbind(competitor,tmp1)
tmp1 <- NA
i <- i + 1
}
competitor <- competitor[-1,]
j <- 1
competitive_level$competeLevel <- NA
competitive_level$classBest <- NA
while(j <= nrow(competitive_level))
{
cid <- competitive_level[j,]$courseid
```
den <- competitor[competitor$courseid==cid,]$Accesses

competitive_level[j,]$competLevel <- competitive_level[j,]$Access / den
competitive_level[j,]$classBest <- den

j <- j + 1

rs = dbWriteTable(mydb, name='Balaji_competitiveLevel1', value =
competitive_level[competitive_level$classBest>0,], append = TRUE, row.names= F)

SOURCE CODE 9: R function to compute competitive level for each student enrolled in each subject

The above R function (source code 9) like the other indicators takes date as input and computes competitive level for that particular date alone. The numerator computation is same as for engagement level while the denominator is the maximum of total number of activities performed by the students in a period of 7 days in a given subject.

The below figure (Figure 16) represents the visualisation output of competitive level of students enrolled in Socials 4t. The same student who is highlighted in engagement level is highlighted here also. Both plots look alike as they capture similar information.
COMPETITIVE LEVEL FOR Socials 4t

factor(Type)
- Median
- Student 11521101

FiguRe 16 COMPETITIVE LEVEL
SUMMARY OF THE MODULE DEVELOPED:

NAME: Competitive level

DESCRIPTION: Competitive level measures how engaged (active) a student is on a given subject in comparison to the best student in the same subject in a period of 7 days.

FEATURES

CALCULATION DONE: From September 15th, 2015 till June 14th, 2016

FREQUENCY OF CALCULATION: Everyday (in a window of 7 days)

EXTRACTION CRITERIA: Students who have performed any type of activities on any of the tasks.

FORM OF REPRESENTATION: No unit (just a number)

EXAMPLE

Assume that there is a student Stuart who performed only 10 activities on a given day.
Date of calculation: 25th January 2016

Most active student’s engagement in the period of 14 days (from 19th January 2016 till 25th January 2016): 25

Competitive level = 10/25 = 0.40

This means that Stuart is only 40% competitive on 25th January 2016.

7.6 PRINCIPAL COMPONENT ANALYSIS AND MOTIVATION MEASUREMENT

Here, the PCA result of indicators developed by myself and A M Preethi are documented and explained. PCA results belongs to data across 6 schools and 15 different subjects. The figure (Figure 17) below gives the correlation between indicators developed in this academic semester (Spring 2016).

NOTE: For the purpose of PCA alone Agility rate is taken as Starting date – First access of student(s) because, for all other indicators higher the value higher is the student motivated.
The figure below (Figure 18) shows the varimax rotation. It shows which indicator has more impact on each component.
It can be inferred from the above figures (Figure 17 and 18) that the

1st factor is due to Engagement level and Competitive level
2nd factor is due to Speed Delivery
3rd factor is due to Time spent
4th factor is due to Persistence
5th factor is due to Break time
6th factor is due to Resilience
7th factor is due to Transition time
8th factor is due to Delivery rate
9th factor is due to Agility Rate
10th factor is due to Forum Participation
The variable factor plot gives information of which indicators can be used to measure motivation. They are: Agility rate, speed delivery, delivery rate, mean of engagement and competitive level, resilience, persistence and forum participation. Motivation index can be measured by scaling all the indicators mentioned above and taking the mean among them. Engagement and Competitive level are combined into one indicator since both measure almost the same thing.

```r
big_table <- big_table[, -c(6,7,12,14,15)]
#REMOVING INDICATORS THAT DO NOT MEASURE MOTIVATION
comb <- big_table[, c(8,9)]
t <- apply(comb, 1, mean, na.rm = T)
#COMBINING ENGAGEMENT AND COMPETITIVE LEVEL AS ONE INDICATOR
big_table$ENGplusCOMPET <- t
big_table <- big_table[, -c(8,9)]
#REMOVING ENGAGEMENT AND COMPETITIVE LEVELS INDIVIDUALLY
active_v <- big_table[, -c(1,2,3,4)]
new_df <- scale(active_v)
#SCALING INDICATOR VALUES
tt <- apply(new_df, 1, mean, na.rm = T)
#COMPUTING MEAN OF THE SCALED INDICATORS
big_table$MOTIVATION_INDEX <- tt
MOTIVATION <- cbind(big_table[,c(1,2,3,4)], big_table$MOTIVATION_INDEX)
```

**SOURCE CODE 10: R script to motivation index for each student enrolled in each subject**

The above R script (Source code 10), measures motivation index from the integrated big table, which contains all the indicators developed by myself and A M Preethi for each student enrolled in each subject from September 15th 2015 till June 12th 2016.

The figure below (Figure 19), represents Motivation plot for the subject Informatica. It can be inferred from the above figure that students enrolled in that subject were more motivated in 2015 than in 2016. And, the student highlighted in blue line is also behaving in the same way.
Figure 19 Motivation plot for Informatica
The above figure (Figure 20) shows the frequency of students having respective motivation indices in the form of histogram. The negative value of motivation index represents demotivation, whose frequency is very high.
CHAPTER 8

VALIDATION

This section validates all the algorithms developed as a part of my TFG from day to start on till competitive level by testing all possible cases (including extreme cases).

8.1 DAY TO START ON

Here, day to start on module is evaluated by test cases. All the possible test cases are: No logs available, less than 10% of the students accessing the course module, Day to start on already available in the table stored in the data base and more than 10% of the students accessing the learning module.

TEST CASE 1: No logs available

TEST CASE INPUT:

<table>
<thead>
<tr>
<th>[1] time</th>
<th>userid</th>
<th>Number_of_students</th>
<th>courseid</th>
<th>cmid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<0 rows> (or 0-length row.names)

EXPECTED RESULT: Output data frame with no data

OBSERVED RESULT:

<table>
<thead>
<tr>
<th>[1] startTime</th>
<th>cmid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<0 rows> (or 0-length row.names)

TEST CASE STATUS: PASS

TEST CASE 2: Less than 10% of the students accesses the course module.

TEST CASE INPUT:

<table>
<thead>
<tr>
<th>Time</th>
<th>userid</th>
<th>Number_of_students</th>
<th>courseid</th>
<th>cmid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1442038190</td>
<td>28668</td>
<td>35</td>
<td>2863023</td>
<td>286129986</td>
</tr>
<tr>
<td>1442038195</td>
<td>28668</td>
<td>35</td>
<td>2863023</td>
<td>286129986</td>
</tr>
<tr>
<td>1442051412</td>
<td>286388</td>
<td>35</td>
<td>2863023</td>
<td>286130146</td>
</tr>
</tbody>
</table>

EXPECTED RESULT: Output data frame for the course module with startTime as NA

OBSERVED RESULT:

<table>
<thead>
<tr>
<th>[1] startTime</th>
<th>cmid</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>286129986</td>
</tr>
</tbody>
</table>

TEST CASE STATUS: PASS
**TEST CASE 3:** Day to start on is already available in the table stored in the data base.

**TEST CASE INPUT:**

<table>
<thead>
<tr>
<th>Time</th>
<th>userid</th>
<th>Number_of_students</th>
<th>courseid</th>
<th>cmid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1452039199</td>
<td>28668</td>
<td>35</td>
<td>2863023</td>
<td>286129986</td>
</tr>
</tbody>
</table>

**EXPECTED RESULT:** The algorithm (R function) not getting invoked

**OBSERVED RESULT:** The function startingDate is not getting invoked

**TEST CASE STATUS:** PASS

**TEST CASE 4:** More than 10% of the students accesses the learning module.

**TEST CASE INPUT:**

<table>
<thead>
<tr>
<th>time</th>
<th>userid</th>
<th>Number_of_students</th>
<th>courseid</th>
<th>cmid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1442258896</td>
<td>862113</td>
<td>49</td>
<td>862544 86272321</td>
<td></td>
</tr>
<tr>
<td>1442258901</td>
<td>862113</td>
<td>49</td>
<td>862544 86272322</td>
<td></td>
</tr>
<tr>
<td>1442387353</td>
<td>8621809</td>
<td>49</td>
<td>862544 86272335</td>
<td></td>
</tr>
<tr>
<td>1442387414</td>
<td>8621814</td>
<td>49</td>
<td>862544 86272335</td>
<td></td>
</tr>
</tbody>
</table>

**EXPECTED RESULT:** A dataframe with startTime as 1442258896 and cmid 86272321.

**OBSERVED RESULT:**

<table>
<thead>
<tr>
<th>startTime</th>
<th>cmid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1442258896</td>
<td>286129986</td>
</tr>
</tbody>
</table>

**TEST CASE STATUS:** PASS

**8.2 AGILITY RATE**

Here, agility rate (number of days to perform first access) module is evaluated by test cases. All the possible test cases are: First access after day to start on, First access before day to start on and No first access is made.

**TEST CASE 1:** First access after day to start on.

**TEST CASE INPUT:**

<table>
<thead>
<tr>
<th>userid</th>
<th>courseid</th>
<th>cmid</th>
<th>FIRST_ACCESS</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>8621809</td>
<td>862544</td>
<td>86272335</td>
<td>1442354400</td>
<td>1442354400</td>
</tr>
</tbody>
</table>

Day to start on: 1442298000

**EXPECTED RESULT:** 15.67 hours

**OBSERVED RESULT:**

<table>
<thead>
<tr>
<th>userid</th>
<th>courseid</th>
<th>cmid</th>
<th>AGILITY RATE</th>
<th>FIRST_ACCESS</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>8621809</td>
<td>862544</td>
<td>86272335</td>
<td>15.666667</td>
<td>1442354400</td>
<td>1442354400</td>
</tr>
</tbody>
</table>
TEST CASE STATUS: PASS

TEST CASE 2: First access before day to start on.
TEST CASE INPUT:
<table>
<thead>
<tr>
<th>userid</th>
<th>courseid</th>
<th>cmid</th>
<th>FIRST_ACCESS</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>86217866</td>
<td>862544</td>
<td>86272334</td>
<td>1442354400</td>
<td>1442354400</td>
</tr>
</tbody>
</table>
Day to start on: 1442562000
EXPECTED RESULT: -57.67 hours
OBSERVED RESULT:
<table>
<thead>
<tr>
<th>userid</th>
<th>courseid</th>
<th>cmid</th>
<th>AGILITY RATE</th>
<th>FIRST_ACCESS</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>86217866</td>
<td>862544</td>
<td>86272334</td>
<td>-57.666667</td>
<td>1442354400</td>
<td>1442354400</td>
</tr>
</tbody>
</table>
TEST CASE STATUS: PASS

TEST CASE 3: No first access is made.
TEST CASE INPUT:
<0 rows> (or 0-length row.names)
Day to start on: 1445060400
EXPECTED RESULT: No output
OBSERVED RESULT:
<0 rows> (or 0-length row.names)
TEST CASE STATUS: PASS

8.3 ENGAGEMENT LEVEL
Here, engagement level module is evaluated by test cases. All the possible test cases are: Student engaging today only and no engagement in the past 13 days, student engaged already in the past 13 days and student engaged neither today nor in the past 3 days.

TEST CASE 1: Student engages today only, no engagement in the past 13 days.
TEST CASE INPUT:

<table>
<thead>
<tr>
<th>DATE</th>
<th>Access</th>
<th>userid</th>
<th>courseid</th>
</tr>
</thead>
<tbody>
<tr>
<td>08-11-2015</td>
<td>13</td>
<td>3991601</td>
<td>399502</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>7</td>
<td>39914901</td>
<td>399502</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>5</td>
<td>39914902</td>
<td>399502</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>7</td>
<td>39914903</td>
<td>399502</td>
</tr>
</tbody>
</table>
08-11-2015  22  39914905  399502

**EXPECTED RESULT:** Engagement level of all the students being 1.

**OBSERVED RESULT:**

<table>
<thead>
<tr>
<th>DATE</th>
<th>Access</th>
<th>userid</th>
<th>courseid</th>
<th>engagementLevel</th>
</tr>
</thead>
<tbody>
<tr>
<td>08-11-2015</td>
<td>13</td>
<td>3991601</td>
<td>399502</td>
<td>1</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>7</td>
<td>39914901</td>
<td>399502</td>
<td>1</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>5</td>
<td>39914902</td>
<td>399502</td>
<td>1</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>7</td>
<td>39914903</td>
<td>399502</td>
<td>1</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>22</td>
<td>39914905</td>
<td>399502</td>
<td>1</td>
</tr>
</tbody>
</table>

**TEST CASE STATUS:** PASS

**TEST CASE 2:** Student engaged already in the past 13 days.

**TEST CASE INPUT:**

<table>
<thead>
<tr>
<th>DATE</th>
<th>Access</th>
<th>userid</th>
<th>courseid</th>
</tr>
</thead>
<tbody>
<tr>
<td>08-11-2015</td>
<td>13</td>
<td>3991601</td>
<td>399502</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>7</td>
<td>39914901</td>
<td>399502</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>5</td>
<td>39914902</td>
<td>399502</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>7</td>
<td>39914903</td>
<td>399502</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>22</td>
<td>39914905</td>
<td>399502</td>
</tr>
</tbody>
</table>

**EXPECTED RESULT:** Engagement level not 1 for any of the test inputs

**OBSERVED RESULT:**

<table>
<thead>
<tr>
<th>DATE</th>
<th>Access</th>
<th>userid</th>
<th>courseid</th>
<th>engagementLevel</th>
</tr>
</thead>
<tbody>
<tr>
<td>08-11-2015</td>
<td>13</td>
<td>3991601</td>
<td>399502</td>
<td>0.33</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>7</td>
<td>39914901</td>
<td>399502</td>
<td>0.19</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>5</td>
<td>39914902</td>
<td>399502</td>
<td>0.18</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>7</td>
<td>39914903</td>
<td>399502</td>
<td>0.41</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>22</td>
<td>39914905</td>
<td>399502</td>
<td>0.82</td>
</tr>
</tbody>
</table>

**TEST CASE STATUS:** PASS

**TEST CASE 3:** Engagement neither today nor in the past 13 days.

**TEST CASE INPUT:**

<table>
<thead>
<tr>
<th>DATE</th>
<th>Access</th>
<th>userid</th>
<th>courseid</th>
</tr>
</thead>
<tbody>
<tr>
<td>08-11-2015</td>
<td>0</td>
<td>3991601</td>
<td>399502</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>0</td>
<td>39914901</td>
<td>399502</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>0</td>
<td>39914902</td>
<td>399502</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>0</td>
<td>39914903</td>
<td>399502</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>0</td>
<td>39914905</td>
<td>399502</td>
</tr>
</tbody>
</table>

**EXPECTED RESULT:** 0
OBSERVED RESULT:

<table>
<thead>
<tr>
<th>DATE</th>
<th>Access</th>
<th>userid</th>
<th>courseid</th>
<th>engagementLevel</th>
</tr>
</thead>
<tbody>
<tr>
<td>08-11-2015</td>
<td>0</td>
<td>3991601</td>
<td>399502</td>
<td>0</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>0</td>
<td>39914901</td>
<td>399502</td>
<td>0</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>0</td>
<td>39914902</td>
<td>399502</td>
<td>0</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>0</td>
<td>39914903</td>
<td>399502</td>
<td>0</td>
</tr>
<tr>
<td>08-11-2015</td>
<td>0</td>
<td>39914905</td>
<td>399502</td>
<td>0</td>
</tr>
</tbody>
</table>

TEST CASE STATUS: PASS

8.4 COMPETITIVE LEVEL

Here, competitive level module is evaluated by test cases. All the possible test cases are: One student accessing the most and the others accessing less and none accessing in a period of 7 days.

TEST CASE 1: One student accesses most and the others accesses less.

TEST CASE INPUT:

<table>
<thead>
<tr>
<th>DATE</th>
<th>userid</th>
<th>courseid</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-05-27</td>
<td>486924</td>
<td>486922</td>
<td>5</td>
</tr>
<tr>
<td>2016-05-27</td>
<td>4862541</td>
<td>486922</td>
<td>7</td>
</tr>
<tr>
<td>2016-05-27</td>
<td>4864221</td>
<td>486922</td>
<td>2</td>
</tr>
<tr>
<td>2016-05-27</td>
<td>4867041</td>
<td>486922</td>
<td>3</td>
</tr>
</tbody>
</table>

EXPECTED RESULT: Competitive level 1 for the one who access the most and appropriate values for students with other accesses.

OBSERVED RESULT:

<table>
<thead>
<tr>
<th>DATE</th>
<th>userid</th>
<th>courseid</th>
<th>Access</th>
<th>COMPETITIVE_LEVEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-05-27</td>
<td>486924</td>
<td>486922</td>
<td>5</td>
<td>0.71</td>
</tr>
<tr>
<td>2016-05-27</td>
<td>4862541</td>
<td>486922</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>2016-05-27</td>
<td>4864221</td>
<td>486922</td>
<td>2</td>
<td>0.29</td>
</tr>
<tr>
<td>2016-05-27</td>
<td>4867041</td>
<td>486922</td>
<td>3</td>
<td>0.43</td>
</tr>
</tbody>
</table>

TEST CASE STATUS: PASS

TEST CASE 2: None has any access in 7 days.

TEST CASE INPUT:

<table>
<thead>
<tr>
<th>DATE</th>
<th>userid</th>
<th>courseid</th>
<th>Access</th>
<th>Competitor</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-05-27</td>
<td>486924</td>
<td>486922</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2016-05-27</td>
<td>4862541</td>
<td>486922</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
**EXPECTED RESULT:** NaN

**OBSERVED RESULT:**

<table>
<thead>
<tr>
<th>DATE</th>
<th>userid</th>
<th>courseid</th>
<th>Access COMPETITIVE_LEVEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-05-27</td>
<td>486924</td>
<td>486922</td>
<td>NaN</td>
</tr>
<tr>
<td>2016-05-27</td>
<td>4862541</td>
<td>486922</td>
<td>NaN</td>
</tr>
<tr>
<td>2016-05-27</td>
<td>4864221</td>
<td>486922</td>
<td>NaN</td>
</tr>
<tr>
<td>2016-05-27</td>
<td>4867041</td>
<td>486922</td>
<td>NaN</td>
</tr>
</tbody>
</table>

**TEST CASE STATUS:** PASS

8.5 TIME SPENT

Here, time spent module is evaluated by test cases. All the possible test cases are: No attempts and presence of attempts.

**TEST CASE 1:** No logs available that shows a student has made attempts.

**TEST CASE INPUT:**

userid | time    | action | courseid | cmid | module_type | table
---|---------|--------|----------|------|-------------|------
<0 rows> (or 0-length row.names)

**EXPECTED RESULT:** The output data frame with no rows.

**OBSERVED RESULT:**

course cmid userid timespent
NA    NA    NA    NA

**TEST CASE STATUS:** PASS

**TEST CASE 2:** Student making attempts on a given day.

**TEST CASE INPUT:**

| userid  | time               | action | courseid | cmid | module_type | table
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>486861</td>
<td>1462617914</td>
<td>viewed</td>
<td>486922</td>
<td>48679411</td>
<td>quiz mdl_quiz</td>
<td></td>
</tr>
<tr>
<td>486861</td>
<td>1462617927</td>
<td>started</td>
<td>486922</td>
<td>48679411</td>
<td>quiz mdl_quiz</td>
<td></td>
</tr>
<tr>
<td>486861</td>
<td>1462617928</td>
<td>viewed</td>
<td>486922</td>
<td>48679411</td>
<td>quiz mdl_quiz</td>
<td></td>
</tr>
<tr>
<td>486861</td>
<td>1462618395</td>
<td>viewed</td>
<td>486922</td>
<td>48679411</td>
<td>quiz mdl_quiz</td>
<td></td>
</tr>
<tr>
<td>486861</td>
<td>1462618565</td>
<td>viewed</td>
<td>486922</td>
<td>48679411</td>
<td>quiz mdl_quiz</td>
<td></td>
</tr>
<tr>
<td>486861</td>
<td>1462618572</td>
<td>submitted</td>
<td>486922</td>
<td>48679411</td>
<td>quiz mdl_quiz</td>
<td></td>
</tr>
<tr>
<td>486861</td>
<td>1462618574</td>
<td>reviewed</td>
<td>486922</td>
<td>48679411</td>
<td>quiz mdl_quiz</td>
<td></td>
</tr>
</tbody>
</table>

**EXPECTED RESULT:** 660 seconds
8.6 SPEED DELIVERY

Here, speed delivery module is evaluated by test cases. All the possible test cases are: No submission, submission after the deadline and submission before the deadline.

TEST CASE 1: No submission on a given day.

TEST CASE INPUT:

timeOfSubmission .userid  courseid  cmid

<0 rows> (or 0-length row.names)

EXPECTED RESULT: Output data frame with no rows.

OBSERVED RESULT:

SPEED_DELIVERY  userid  courseid  cmid

<0 rows> (or 0-length row.names)

TEST CASE STATUS: PASS

TEST CASE 2: Submission after the deadline.

TEST CASE INPUT:

DEADLINE: Wed, 13 Apr 2016 08:36:15 GMT

TIME OF SUBMISSION

20/05/2016, 09:28:06 GMT+2:00 DST

20/05/2016, 09:30:39 GMT+2:00 DST

EXPECTED RESULT: Negative values of speed delivery.

OBSERVED RESULT:

SPEED_DELIVERY  userid  courseid  cmid

-886.8642 4866381 486922 48679454

-886.9067 4864221 486922 48679454

TEST CASE STATUS: PASS
TEST CASE 3: Submission before the deadline.

TEST CASE INPUT:

TIME OF SUBMISSION

12/04/2016, 22:19:56 GMT+2:00 DST

12/04/2016, 22:29:51 GMT+2:00 DST

<table>
<thead>
<tr>
<th>timeOfSubmission</th>
<th>userid</th>
<th>courseid</th>
<th>cmid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1460492396</td>
<td>15183130</td>
<td>1518184</td>
<td>151817222</td>
</tr>
<tr>
<td>1460492991</td>
<td>15183130</td>
<td>1518184</td>
<td>151817202</td>
</tr>
</tbody>
</table>

Deadline: 15/04/2016, 02:18:06 GMT+2:00 DST

EXPECTED RESULT: Positive values of speed delivery.

OBSERVED RESULT:

<table>
<thead>
<tr>
<th>SPEED_DELIVERY</th>
<th>userid</th>
<th>courseid</th>
<th>cmid</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>51.96944</td>
<td>15183130</td>
<td>1518184</td>
<td>151817222</td>
<td>1460412000</td>
</tr>
<tr>
<td>51.73278</td>
<td>15183130</td>
<td>1518184</td>
<td>151817202</td>
<td>1460412000</td>
</tr>
</tbody>
</table>

TEST CASE STATUS: PASS

8.7 TRANSITION TIME

Here, transition time module is evaluated by test cases. All the possible test cases are: No logs available, student making transition between activities and student making no transition.

TEST CASE 1: No logs available.

TEST CASE INPUT:

<table>
<thead>
<tr>
<th>time</th>
<th>userid</th>
<th>cmid</th>
<th>courseid</th>
<th>action</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0 rows&gt; (or 0-length row.names)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

EXPECTED RESULT: Output data frame with no rows.

OBSERVED RESULT:

<table>
<thead>
<tr>
<th>transitionTime</th>
<th>userid</th>
<th>cmid</th>
<th>courseid</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0 rows&gt; (or 0-length row.names)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TEST CASE STATUS: PASS

TEST CASE 2: Any student makes transition between activities.

TEST CASE INPUT:
**EXPECTED RESULT:** Transition time of 46 seconds.

**OBSERVED RESULT:**

<table>
<thead>
<tr>
<th>transitionTime</th>
<th>userid</th>
<th>cmid</th>
<th>courseid</th>
<th>action</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
<td>3991010</td>
<td>399158904</td>
<td>3994702</td>
<td></td>
</tr>
</tbody>
</table>

**TEST CASE STATUS:** PASS

**TEST CASE 3:** A student makes no transition.

**TEST CASE INPUT:**

<table>
<thead>
<tr>
<th>time</th>
<th>userid</th>
<th>cmid</th>
<th>courseid</th>
<th>action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1455546900</td>
<td>3991010</td>
<td>399162101</td>
<td>3994702</td>
<td>viewed</td>
</tr>
</tbody>
</table>

**EXPECTED RESULT:** Transition time is NA.

**OBSERVED RESULT:**

<table>
<thead>
<tr>
<th>transitionTime</th>
<th>userid</th>
<th>cmid</th>
<th>courseid</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>3991010</td>
<td>399158904</td>
<td>3994702</td>
</tr>
</tbody>
</table>

**TEST CASE STATUS:** PASS

**8.8 DELIVERY RATE**

Here, delivery rate module is evaluated by test cases. All the possible test cases are: Submission when there is/are active pending task(s), submission when there is no active pending task(s) and no submission when there is active pending task(s).

**TEST CASE 1:** Submission when there is/are active pending task(s).

**TEST CASE INPUT:**

<table>
<thead>
<tr>
<th>userid</th>
<th>courseid</th>
<th>cmid</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1152283</td>
<td>1152204</td>
<td>115218104</td>
<td>2016-06-09 22:00:44</td>
</tr>
<tr>
<td>1152283</td>
<td>1152204</td>
<td>115218108</td>
<td>2016-06-09 22:06:23</td>
</tr>
<tr>
<td>1152283</td>
<td>1152204</td>
<td>115218102</td>
<td>2016-06-09 22:14:55</td>
</tr>
<tr>
<td>1152283</td>
<td>1152204</td>
<td>115218101</td>
<td>2016-06-09 21:34:48</td>
</tr>
</tbody>
</table>

**EXPECTED RESULT:** Delivery rate of 1 since he/she has submitted all the pending ones.

**OBSERVED RESULT:**

| USERID COURSEID pending delRate submissions Date |
|-----------------------------------------------|------------------------------------------------|
|                                              |                                                |
TEST CASE 2: Submission when there is no active pending task(s).

TEST CASE INPUT:
<table>
<thead>
<tr>
<th>userid</th>
<th>courseid</th>
<th>cmid</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>286353</td>
<td>2863023</td>
<td>286163566</td>
<td>2015-09-15 16:46:12</td>
</tr>
</tbody>
</table>

EXPECTED RESULT: NaN as delivery rate.

OBSERVED RESULT:
<table>
<thead>
<tr>
<th>USERID</th>
<th>COURSEID</th>
<th>pending</th>
<th>delRate</th>
<th>submissions</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>286353</td>
<td>2863023</td>
<td>0</td>
<td>NaN</td>
<td>1</td>
<td>1442268000</td>
</tr>
</tbody>
</table>

TEST CASE STATUS: PASS

TEST CASE 3: No submission when there is active pending task(s).

TEST CASE INPUT:
<table>
<thead>
<tr>
<th>userid</th>
<th>courseid</th>
<th>cmid</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0 rows&gt; (or 0-length row.names)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

EXPECTED RESULT: Delivery rate of 0.

OBSERVED RESULT:
<table>
<thead>
<tr>
<th>USERID</th>
<th>COURSEID</th>
<th>pending</th>
<th>delRate</th>
<th>submissions</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1152296</td>
<td>2863023</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1442268000</td>
</tr>
</tbody>
</table>

TEST CASE STATUS: PASS

8.9 VALIDATION BY PCA

Correlation of all the indicators developed by me can complement validation step. The figure below (Figure 20) shows the correlation between various indicators developed by me at inLab, FIB. It can be found that engagement and competitive levels are much correlated (positively) as both measure the same thing with reference to different quantities. And also, agility rate and speed delivery are correlated in the same way. If a student is quick enough to access learning materials, then it is likely that he/she will complete mandatory tasks before the deadline. This makes sense.
Variables factor map (PCA)

Figure 21 Correlation plot for validation
The above figure (Figure 21) shows the varimax rotation of indicators developed by me. The loadings gives clear information of which variables are correlated. It appears that time spent and transition time cannot be used to measure motivation. Instead, it can be used to find out the learning pattern of the students.
CHAPTER 9

CONCLUSION

This section describes the point up to which our goals are achieved, scope for future work and learning outcomes of mine from this project. We measured motivation index of each student from the subset of indicators obtained from correlation analysis of PCA. Totally, 16 indicators were developed which will definitely help teachers in secondary schools of Catalunya to identify at risk students at earliest possible time. Some indicators that were developed by previous members who worked in the learning analytics project at inLab were redeveloped in order to improve the robustness of what they measure. Some of the developed indicators, are correlated well and the outcome will be very useful.

An important thing that needs to be understood is that if the required data like starting date and deadline are provided by the teachers of secondary schools, then the indicators developed can more accurately represent the results. The existing dashboard needs to be modified to incorporate development done in form of indicators. Once it is done, the developed work can be integrated with the existing platform of LA project at inLab.

9.1 SCOPE FOR FUTURE WORK

- The motivation values and the indicators developed needs to be tested in real-time by means of teachers who are handling the associated subjects in the secondary schools.
- The time and space complexity of the algorithms developed can be improved.
- The information of student’s performance in test(s) and exam(s) can be used to validate the developed indicators.
- A predicator model can be developed which can tell how motivated a student can be tomorrow.
- The development work can be extended to work with 1500 schools across Catalunya by migrating to big data architecture.
- Sentimental analysis can be done on data regarding student participation in forums.

9.2 LEARNING OUTCOMES

Before beginning the development work I completed an online course “Introduction to R” offered by Microsoft via edx to get the foundation for programming in R. I have also completed “SQL fundamentals” course offered by SoloLearn to improve the efficiency of my queries by getting the basics right. This being
my first internship experience, I got to know how things work in a top educational organisation and how discipline is maintained. Apart from this, I worked up to my potential for the first time in my life to a significant extent.
REFERENCES


9. Jiawei Han, Micheline Kamber and Jian Pei. Data Mining Concepts and Techniques. s.l. : Morgan Kaufmann. ISBN 978-0-12-381479-1.


