DEVELOPING CHOICE AND PERSISTENCE INDICATORS IN A LEARNING ANALYTICS PLATFORM FOR SECONDARY SCHOOLS OF CATALONIA

BACHELOR THESIS REPORT

AUTHOR

ATTULURI MOHANAPREETHI (INTERNATIONAL STUDENT)

DIRECTORS

PROF.TOMAS ALUJA BANET

PROF.MARIA RIBERA SANCHO

COORDINATOR

MR.ALBERT OBIOLS

Bachelor's Degree in Computer Science and Engineering

Facula informatics de Barcelona (FIB)

Universidad Polytechnic de Catalonia (UPC) - BarcelonaTech

June, 2016

RESUME (ENGLISH)

DEVELOPING CHOICE AND PERSISTENCE INDICATORS IN A LEARNING ANALYTICS PLATFORM FOR SECONDARY SCHOOLS OF CATALONIA

Goal of Learning Analytics for Secondary Education is to provide insight Information of the behaviour of students regarding the usage of Virtual learning Environment. The learning platform used in the project is Moodle based Agora. Agora is one of the most widely used Learning platforms in secondary schools of Catalonia. In the Learning Analytics Project, data from schools in the region has been collected and populated into a local database. The data is filtered using ETL process using which indicators are developed. Inturn, motivation index of each student is calculated and visualized in the dashboard. In the existing system, very few indicators for six schools have been developed and the data has been stored in the MySQL database. To flip the situation of modest set of indicators all factors have been considered and new indicators have also been device. To supplement the existing, at present, a total of 14 indicators, classified into four categories are proposed. My part of the project contributes to the main objective by developing Choice and Persistence indicators which include Resilience level, Number of attempts to finish a task, Curiosity rate, Forum access, Forum Participation, Persistence level, Priority rate, Break time and Number of logs executed using R language. These indicators are visualized using ggplot and integrated with the existing platform. This fillips the teacher's interest by displaying personalized dashboard showing the indicators at different levels implemented as a set of filters and thus assisting in describing the motivation index of each student.

RESUMEIX (Catalan)

DESENVOLUPAMENT DE L'OPCIÓ I PERSISTÈNCIA INDICADORS EN UNA PLATAFORMA D'APRENENTATGE ANÀLISI DE SECUNDÀRIA DE CATALUNYA

Objectiu d'aprenentatge Analytics per l'Educació Secundària és perquè es conegui la informació del comportament dels estudiants en relació amb l'ús de l'entorn virtual d'aprenentatge. La plataforma d'aprenentatge utilitzat en el projecte es basa Moodle Àgora. Àgora és una de les plataformes d'aprenentatge més utilitzats a les escoles secundàries de Catalunya. En el Projecte d'Aprenentatge Analytics, les dades de les escoles de la regió s'han recollit i poblada en una base de dades local. Les dades es filtra mitjançant el procés de ETL utilitzant indicadors que es desenvolupen. Inturn, índex de motivació de cada estudiant es calcula i es visualitza al tauler d'instruments. En el sistema actual, molt pocs indicadors per a sis escoles s'han desenvolupat i les dades han estat emmagatzemades a la base de dades MySQL. Per capgirar la situació del modest conjunt d'indicadors de tots els factors han estat considerades i nous indicadors també han estat dispositiu. Com a complement a la ja existent, en l'actualitat, es proposa un total de 14 indicadors, que es classifica en quatre categories. La meva part del projecte contribueix a l'objectiu principal mitjançant el desenvolupament de Choice i persistència indicadors que inclouen nivell de resiliència, Nombre d'intents per acabar una tasca, la taxa de curiositat, l'accés Fòrum, Fòrum de Participació, el nivell de persistència, la taxa de Prioritat, trencament i del número de registres executat utilitzant el llenguatge R. Aquests indicadors es van visualitzar utilitzant ggplot i integrats amb la plataforma existent. Aquest Fillips interès del mestre mostrant tauler personalitzat que mostra els indicadors a diferents nivells implementats com un conjunt de filtres i ajudant així al descriure l'índex de la motivació de cada estudiant.

RESUME (Spanish)

DESARROLLO DE LA OPCIÓN Y PERSISTENCIA INDICADORES EN UNA PLATAFORMA DE APRENDIZAJE ANÁLISIS DE SECUNDARIA DE CATALUÑA

Objetivo de aprendizaje Analytics para la Educación Secundaria es para que se conozca la información del comportamiento de los estudiantes en relación con el uso del entorno virtual de aprendizaje. La plataforma de aprendizaje utilizado en el proyecto se basa Moodle Ágora. Ágora es una de las plataformas de aprendizaje más utilizados en las escuelas secundarias de Cataluña. En el Proyecto de Aprendizaje Analytics, los datos de las escuelas de la región se han recogido y poblada en una base de datos local. Los datos se filtra mediante el proceso de ETL usando indicadores que se desarrollan. Inturn, índice de motivación de cada estudiante se calcula y se visualiza en el tablero de instrumentos. En el sistema actual, muy pocos indicadores para seis escuelas se han desarrollado y los datos han sido almacenados en la base de datos MySQL. Para voltear la situación del modesto conjunto de indicadores de todos los factores han sido consideradas y nuevos indicadores también han sido dispositivo. Como complemento a la ya existente, en la actualidad, se propone un total de 14 indicadores, que se clasifica en cuatro categorías. Mi parte del proyecto contribuye al objetivo principal mediante el desarrollo de Choice y persistencia indicadores que incluyen nivel de resiliencia, Número de intentos para terminar una tarea, la tasa de curiosidad, el acceso Forum, Foro de Participación, el nivel de persistencia, la tasa de Prioridad, rotura y del Número de registros ejecutado utilizando el lenguaje R. Estos indicadores se visualizaron utilizando ggplot e integrados con la plataforma existente. Este Fillips interés del maestro mostrando tablero personalizado que muestra los indicadores a diferentes niveles implementados como un conjunto de filtros y ayudando así al describir el índice de la motivación de cada estudiante.

ACKNOWLEDGEMENT

Foremost, I am very thankful to the almighty for his love and blessings for what I am today. I would like to express my heartfelt gratitude to my home institute, SASTRA University, India, for having given us an opportunity to do our Internship in UPC, Barcelona.

At the same time, I am thankful to UPC for accepting us and giving us exposure. I am really grateful for this opportunity provided by our beloved Dean, Dr.P.Swaminathan and Assosciate Dean, Dr.A.Umamakeswari . We feel privileged to have studied in the department of Computer Science which had talented teaching staff and we would like to thank each of the teaching and non-teaching staff of Department of CSE.

I am very thankful to my guide Prof. Tomas Aluja for all his guidance and support throughout the duration of the project. His advices and wide knowledge in my project domain lead us to be done with a successful project. Prof. Maria Ribera was the one behind the successful completion of my project thesis. She let us to do things properly and in organized way by reviewing my work periodically and giving valuable suggestions. I am very thankful to her. I would like to thank Mr. Albert Obviols for his constant support and encouragement in inLab. He helped a lot to progress further in the project. He was so kind and understanding.

I also thank my team in inLab namely Jordi Casinovas, Balaji Natarajan, Daniel Gilbert for all the help they did, whenever needed in the course of the project. Jordi was so helpful for me and Balaji. He explained our doubts with patience and made everything clear for us. I wish toextend my thanks to Prof.Jasmina Berbegal for her suggestions and feedback in completing my GEP course.

Last but not the least, I would like to convey lots of thanks and loads of love to my family and friends without whom this would have not happened. Thank you each and every one.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The title of my TFG is “Developing Choice and Persistence Indicators in a Learning Analytics platform for all Secondary schools in Catalonia”. Now a days, E-Learning has been emerged and has been extensively used by many of the learning institutions in almost all parts of the world. This leads to the emergence of many Virtual Learning Environments which necessarily requires proper handling of Data. Here comes Learning Analytics.

1.1 Context

The Learning Analytics Management system is used to help the teachers of all secondary schools in Catalonia by providing their students’ behaviour in Learning Analytics platform. Before going in detail, we shall see what Learning Analytics mean.

Definition: Learning analytics is 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.

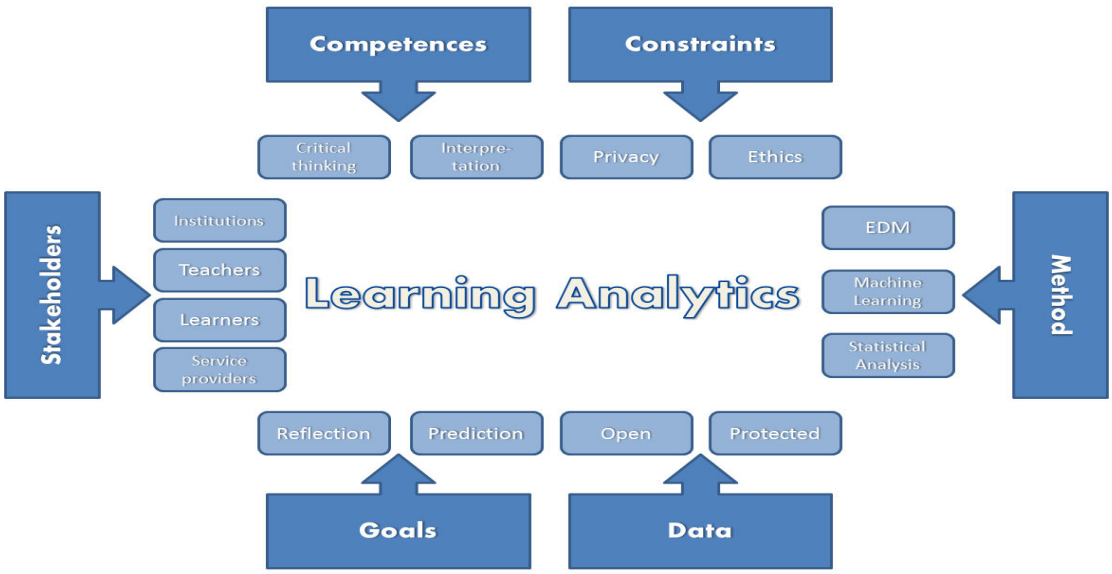


Figure 1.1 Learning Analytics overview

Learning Analytics is a branch of Data Analytics that deals with the study of learning behaviour of students from a learning platform. Learning analytics assumes the application of statistics, data mining, machine learning, operations research, network analysis, information visualization and other methods to extract patterns and knowledge from digital information in order to support decision making. For this project the supporting Learning Platform is the MOODLE based Agora.

The Agora is one of the most widely used Learning platforms in secondary schools of Catalonia. The main idea of the Learning analytics project is to collect the data from different schools, i.e. from Moodle database and is stored in local database. This data is processed and filtered using several Data mining techniques. The different indicators are developed for every student in every course. The Motivation Index is calculated using the developed indicators. Finally, the visualization techniques are used and the personalized dashboards are designed for the teachers. My part is developing indicators, visualising them and calculating motivation index.

This project is done with the collaboration of the Educational department of the Generalist de Catalonia. In Lab FIB is an innovation and research lab based in the Barcelona School of Informatics, Universidad Polytechnic de Catalonia that integrates academic personnel from different UPC departments and its own technical staff to provide solutions to a wide range of demands that involve several areas of expertise. This project is aimed to do collaborative research and open innovation in the field of learning analytics.

In the next section the various Objectives and stakeholders of the project are described and their various roles are discussed.

1.2 Main Objectives

The main objectives of my TFG are designing and implementing the set of indicators for the data taken from the Moodle platform and calculating the motivation index of the students. This is the first and foremost objective of my TFG. However, being a part of the big Learning analytics project, some part of the data wrangling and exploratory data analysis is also been the objectives of my project. Including all considerations, the objectives in detail are explained below.

Developing Indicators and Data visualization: Deals with developing Persistence and Choice Indicators to measure the Motivation Index of the Students and Graphic design and Human-Computer Interaction (HCI). Graphic design helps in representing and refining the data.

Data Handling: Since the data is taken from the huge databases, the data is to be handled properly without any losing and without getting incorrect data. After the computations also, the data has to be properly stored into the database.

Exploratory data analysis: This is associated with data sciences and statistics. It involves filtering and mining of data, applying statistical and mathematical theories to discover insights. The principle component analysis is to be done for the developed indicators to find the correlation between the indicators and thus to obtain the index of motivation.

1.3 Stakeholders and users of the system

A stakeholder in the architecture of a system is an individual, team, organization or classes there of having an interest in the realisation of the system. They are directly or indirectly affected by the project.

The various stakeholders in this system are being listed below.

1. The teachers of all the secondary schools.

The teachers of the secondary schools in Catalonia are the main and direct end-users of the system developed. The system itself is to help the teachers to see the motivation of their students. Thus they can easily analyse their students and help them to get improved. For Example, a student of particular course is highly motivated in all the days except one or two days, then it can be expected that the student is in trouble and the teacher may try to help him solve the problem and continue with his range of interest. Thus we can say that the teachers are the most prior stakeholders of my TFG.

2. The students of Catalonia, whose Motivation Index is calculated.

The students are considered as the stakeholders with next higher priority. Their work itself is the input for my project and the output is theirs motivation. So obviously there are the largely affected group of stakeholders. They too can see

their motivation value and can self-realise how they are working. If more efforts are to be kept, they can come to know and can rectify it in the future work. Also they can analyse themselves like in which course, they have more interest, what is their position among the peer students etc.

3. The Educational department of Spain.

The project is carried out in accordance with the Board of Education of the Catalonia. According to the statistics [2] provided by The Ministry of Education, the number of students in all the Elementary, medium and higher grade schools is very low in the academic year 2013/14. It is calculated that the overall literacy rate of Spain is 55.6% which is considered to be low when compared with other European Countries. Thus letting the Government know the statistics of the students may help them to implement strategies to improve the literacy rate by indulging interest provoking activities and courses in schools. And hence the Education board of Catalonia as well as entire Spain can make use of this Project. This may be extended to the other parts of Europe. But it may take some time and other technologies to handle huge amounts of data.

4. Prof. Tomas Abuja Banet, Prof. Maria Ribera, our project directors and Albert Obiols, our project co-director at in Lab FIB.

All the indicators' definitions are framed under the guidance of Prof. Tomas Aluja Banet and Prof. Maria Ribera and working with all the tools and platforms are mentored by Mr. Albert Obiols.

5. The inLab participating team (Developers)

Jordi Casanovas

Daniel Gibert

Attuluri MohanaPreethi (myself)

Balaji Natarajan

Uyaan

Each one in our group has the different modules as our TFGs. The integration of all the small projects is the final result of the Learning Analytics Research Project.

CHAPTER 2

STATE OF THE ART

This section aims at studying the efforts that are already done in the field of learning analytics. I have included some of the studies made in Data analysis in Academic context, data mining and Students' progress in Online working platform etc.,

Marquez, with Michael Brostock [3] and Murray Scott [4] presents a study done with 15-year old students in Mexico to identify relevant factors for students' academic failure. They conducted a survey about personal and family situation, a socioeconomic study and the final scores provided by School Management. Dekker presented an analysis [5] of dropout rate of freshmen based on early grades and personal situation.

Romero [6] presents a summary of different applications of Data mining techniques in Learning Analytics. He also presents his results using clusters to group students according to their activity level and C4.5 algorithm [7] to characterize students who passed or failed the course.

The aim of Bogarin et al. [8] is to find models for each niche of students which turn out to be more precise than a single model for all the students at once. This considers the time between uploading the resources the student accessing it for the first time as an indicator.

. Bailey presents a study [9] on University of Derby students drop out whose aim is the early detection of students at risk in order to offer help for concluding their major or for changing to a more suitable one. Despite they felt the need of a dashboard of indicators, the outcome of the case study was to scope out requirements to service design analysis, not to develop a tool for university staff and students themselves, which is considered for later, starting from a database able to integrate information from different IT systems.

Based on 1534 students of 34 online courses from a Greek university, Kazanidis determines that the time dedicated to learning is the main predictor of good grades. They use association rule mining, but also clustering, classification and regression trying to measure Visits per Session, Visits per Duration, Course

Utilization and user Perception and Average File as metrics departing from the following measures: the total number of sessions per course viewed by all users, the total number of visits per course by all users, and duration.

Dawson proposes an analysis at different levels: enterprise, faculty and teacher [10] through the visualization of students' pattern of overall engagement from a big Canadian University. Comparing the activity of professors and students, he concludes that those students who had a set of goals and guidelines of forum use were the most active ones, despite the frequency of the reply from faculty.

In this set of publications, we notice the main focus is on undergraduate students and MOOCs, leaving almost empty the field of Secondary School students, those from 13 to 16 years old with face to-face courses supported by a Learning Management System. As our group of interest is lacking attention, it is a valid reason to start studying it.

So we propose to develop a learning platform which includes the validation of various indicators against secondary school students' data and offers customized tools for students' education which may foster, through the supervision of student on-line activity.

CHAPTER 3

METHODOLOGY AND THEORETICAL FRAMEWORK

Chapter 3 mainly discusses what methodology is used for the development of my part of the project and for the development of the project as a whole and the theoretical framework for advancing into the project.

3.1 Methodology

Methodology is the heart of the Project. In my project, as discussed earlier, the data is collected from different schools and is updated every 24 hours once. The data collected should be integrated properly with the already existing data. Certain measures should be taken such that the size of the data should not be a problem. During the initial stage of the academic year, there is less amount of data and can be handled easily. As the time proceeds, the data heaps up and at some point, its size becomes vast and difficult to handle. Sometimes the data may not be available to be populated into the local database. For example, during the weekends, most of the schools do not work and thus there is no data available on those days. If the server checks for the data on those days, it is only waste of server processing time. There are also possibilities of some situations which are not expected to occur, which may affect the efficiency of the project. Examples include the above discussed excess size of the data, unavailability of data and other factors.

While developing the indicators, the requirements and specifications of each module may change. According to the dynamic rules and thesis, the strategies and logic are to be modified. Thus, the algorithms may change as the development progresses.

Considering all these circumstances, we landed up in choosing a dynamic and flexible methodology rather a rigid and static one. The methodology that best suits our project development is AGILE Methodology.

3.1.1 Agile Methodology [11]

Agile software development is a set of principles for software development in which requirements and solutions evolve through collaboration between self-organizing, cross-functional teams. It promotes adaptive planning, evolutionary development, early delivery, and continuous improvement, and it encourages rapid and flexible response to change. Agile itself has never defined any specific methods to achieve this, but many have grown up as a result and have been recognized as being 'Agile'.

Agile approaches help teams respond to unpredictability through incremental, iterative work cadences, known as sprints. Agile methodologies are an alternative to waterfall, or traditional sequential development which is considered to be better that the later.

The Agile movement is not anti-methodology, in fact many of us want to restore credibility to the word methodology. We want to restore a balance. We embrace modelling, but not in order to file some diagram in a dusty corporate repository. We embrace documentation, but not hundreds of pages of never-maintained and rarely-used tomes. We plan, but recognize the limits of planning in a turbulent environment. Those who would brand proponents of XP or SCRUM or any of the other Agile Methodologies as "hackers" are ignorant of both the methodologies and the original definition of the term hacker.

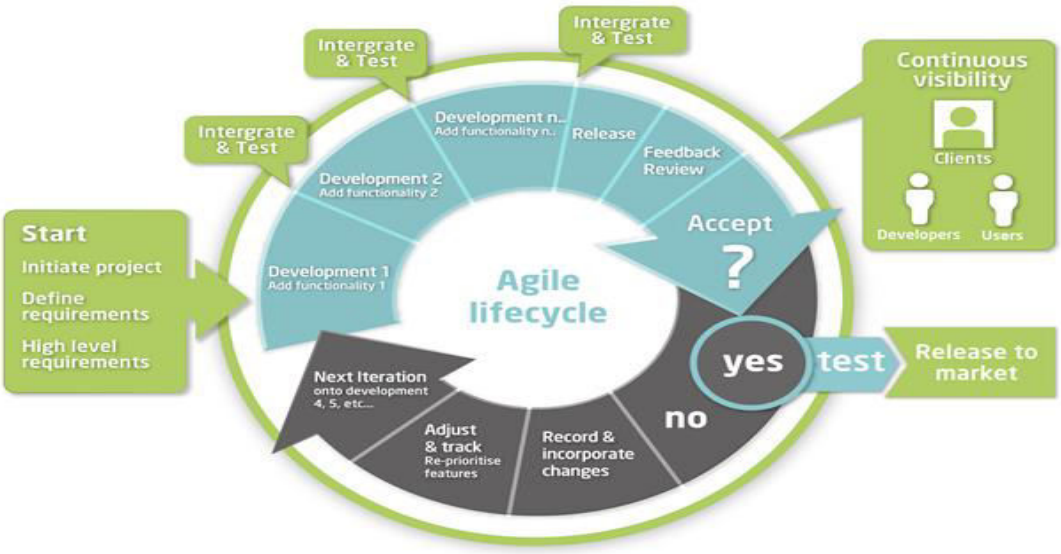


Fig 3.1

Agile Life cycle in Project development

3.1.1.1 Agile methodology in my project

As shown in the figure, as per the Agile life cycle, my part of the learning analytics Project initially started with defining requirements, giving the definitions for all the indicators and deciding what the final motivation index of the students should be. As the development of indicators progresses, the definition of the indicators are remodified to suit the available data and to improve the efficiency. All the developed indicators are checked for consistency with each other and with the progress of the academic year. Any inconsistencies that occur are rectified then and there. Proper validation and justification for the output is done not only at final release of the product but also during the course of development. The calculated motivation index is deployed in the real environment and validated. If it is not feasible and not acceptable by end users of the educational department of Catalonia, the development process is again iterated and tried to obtain the better results. This is how agile methodology is deployed for this Project.

Below are some of the Agile Principles that are followed to develop the Project of learning analytics.

3.1.1.2 Agile principles

The Agile Manifesto is based on twelve principles:

1. Customer satisfaction by early and continuous delivery of valuable software
2. Welcome changing requirements, even in late development
3. Working software is delivered frequently (weeks rather than months)
4. Close, daily cooperation between business people and developers
5. Projects are built around motivated individuals, who should be trusted
6. Face-to-face conversation is the best form of communication (co-location)
7. Working software is the principal measure of progress
8. Sustainable development, able to maintain a constant pace
9. Continuous attention to technical excellence and good design
10. Simplicity—the art of maximizing the amount of work not done—is essential
11. Best architectures, requirements, and designs emerge from self-organizing teams

12. Regularly, the team reflects on how to become more effective, and adjusts accordingly

Agile methods have been extensively used for development of software products and some of them use certain characteristics of software, such as object technologies. So as the case here.

Agile methodologies can be inefficient in large organizations and certain types of projects. Agile methods seem best for developmental and non-sequential projects. Many organizations believe that agile methodologies are too extreme and adopt a hybrid approach that mixes elements of agile and plan-driven approaches. However, DSDM is an agile methodology that in fact mixes elements of agile and plan-driven approaches in a disciplined way, without sacrificing the fundamental principles that make agile work.

3.2 Theoretical Framework

This section clearly explains the theoretical framework bolstering the project as a whole. This section is necessary to have the basic knowledge of the tools required, relative platforms to work, context and environment of the working domain. Having known about all the above mentioned, helps to proceed smoothly and easily. The advancement in the Project is induced.

This includes various sections like Data Analytics, Structured Query Language (SQL), Moodle database, RStudio, Highcharts, Importance of Motivation, and Principle Component Analysis (PCA).

3.2.1 Data Analytics [12]

As taken from the references, Data analytics is explained as follows.

Data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about that information.

Definition: Analysis of data is the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data analytics is distinguished from data mining by the scope, purpose and focus of the analysis. Data miners sort

through huge data sets using sophisticated software to identify undiscovered patterns and establish hidden relationships. Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known by the researcher.

Statistician John Tukey defined data analysis in 1961 as: "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

There are several phases that can be distinguished, described below. The phases are iterative, in that feedback from later phases may result in additional work in earlier phases.

3.2.1.1 Data requirements

The data which is necessary as inputs to the analysis are specified based upon the requirements of those directing the analysis or customers who will use the finished product of the analysis. The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people). Specific variables regarding a population (e.g., age and income) may be specified and obtained. Data may be numerical or categorical (i.e., a text label for numbers).

3.2.1.2 Data collection

Data is collected from a variety of sources. The requirements may be communicated by analysts to custodians of the data, such as information technology personnel within an organization. The data may also be collected from sensors in the environment, such as traffic cameras, satellites, recording devices, etc. It may also be obtained through interviews, downloads from online sources, or reading documentation.

3.2.1.3 Data processing

Data initially obtained must be processed or organized for analysis. For instance, this may involve placing data into rows and columns in a table format for further analysis, such as within a spreadsheet or statistical software.

3.2.1.4 Data cleaning

Once processed and organized, the data may be incomplete, contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that data is entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, deduplication, and column segmentation. Such data problems can also be identified through a variety of analytical techniques. For example, with financial information, the totals for particular variables may be compared against separately published numbers believed to be reliable. Unusual amounts above or below pre-determined thresholds may also be reviewed. There are several types of data cleaning that depend on the type of data. Quantitative data methods for outlier detection can be used to get rid of likely incorrectly entered data. Textual data spellcheckers can be used to lessen the amount of mistyped words, but it is harder to tell if the words themselves are correct.

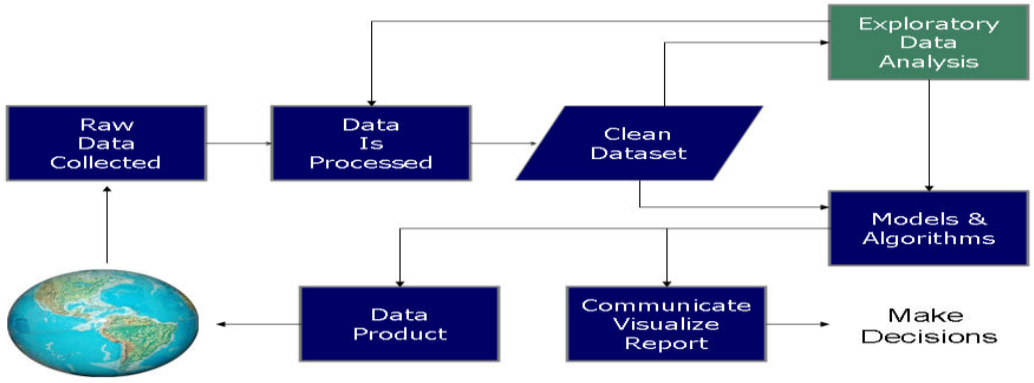


Figure 3.2 Phases in Data Analysis process

3.2.1.5 Exploratory data analysis

Once the data is cleaned, it can be analyzed. Analysts may apply a variety of techniques referred to as exploratory data analysis to begin understanding the messages contained in the data. The process of exploration may result in additional data cleaning or additional requests for data, so these activities may be iterative in nature. Descriptive statistics such as the average or median may be generated to help understand the data. Data visualization may also be used to examine the data in graphical format, to obtain additional insight regarding the messages within the data.

3.2.1.6 Modeling and algorithms

Mathematical formulas or models called algorithms may be applied to the data to identify relationships among the variables, such as correlation or causation. In general terms, models may be developed to evaluate a particular variable in the data based on other variable(s) in the data, with some residual error depending on model accuracy (i.e., Data = Model + Error).

3.2.1.7 Data product

A data product is a computer application that takes data inputs and generates outputs, feeding them back into the environment. It may be based on a model or algorithm. An example is an application that analyzes data about customer purchasing history and recommends other purchases the customer might enjoy.

3.2.1.8 Communication

Once the data is analyzed, it may be reported in many formats to the users of the analysis to support their requirements. The users may have feedback, which results in additional analysis. As such, much of the analytical cycle is iterative.

When determining how to communicate the results, the analyst may consider data visualization techniques to help clearly and efficiently communicate the message to the audience. **Data visualization** uses information displays such as tables and charts to help communicate key messages contained in the data. Tables are helpful to a user who might lookup specific numbers, while charts (e.g., bar charts or line charts) may help explain the quantitative messages contained in the data.

3.2.2 Stuctured Query Language (SQL):

Structured Query Language is a special-purpose programming language designed for managing data held in a relational database management system (RDBMS). The scope of SQL includes data insert, query, update and delete, schema creation and modification, and data access control. SQL can be broadly categorized into the DDL (Data Definition Language), DML (Data Manipulation Language) and Database Control Language (DCL).

Following are the relevant SQL commands used in the project to query the database and extracting required data to develop indicators of motivation.

3.2.2.1 Data Definition Language (DDL)

The Data Definition Language (DDL) is used to define new table schemas or alter the already existing ones. They usually take the name of the entity that needs be added or deleted to the already existing/new schema as an input and performs the action specified.

CREATE TABLE- This command is used to create a table to the specified database. It takes the columns and their data types as their input and creates the tables accordingly.

The syntax of the create table is as follows

```
CREATE   TABLE table_name
(
column_name1 data_type(size),
column_name2 data_type(size),
column_name3 data_type(size),
....
);
```

ALTER TABLE- This query is used to alter the already existing table schema by adding a new column or deleting the existing column or modifying the existing column. This takes the column that needs to be altered or added to the schema and its datatype (based on specific action) as an input. For example, here are the syntaxes for some actions that can be done by ALTER TABLE statement.

To add a column in a table:

```
ALTER TABLE table_name ADD column_name datatype
```

To delete a column in a table:

```
ALTER TABLE table_name DROP COLUMN column_name
```

To change the data type of a column in a table :

```
ALTER TABLE table_name MODIFY column_name datatype
```

DROP TABLE- This command is used to delete the table schema or a particular column from an already existing database. This cannot be reverted. A dropped table cannot be reversed.

The syntax is same as the DROP statement in the above section. (which is used along with alter command)

3.2.2.2 Data Manipulation Language (DML)

The Data manipulation language of the SQL is used to perform operations with the data in the database tables. They were the most used in the development of indicators. They are usually used to manipulate and perform data analytics.

SELECT- This query is used to select a subset of tuples (or records) from a table.This can be also used to get data based on a condition specified by the WHERE clause. This can be used in combination with other query commands like INSERT,SET etc.,

The syntax of the select statement is as follows:

```
SELECT column_name,column_name FROM table_name;
```

UPDATE- This query is used to update an existing table just by modifying the values of a certain column. This goes with the SET keyword.

Syntax:

```
UPDATE table_name
SET column1=value1,column2=value2,...
WHERE some_column=some_value;
```

INSERT- This query command is used to insert a new tuple into the table.

Syntax:

```
INSERT INTO table_name (column1,column2, column3,...)
VALUES (value1,value2,value3,...);
```

A combination of the above two sets of queries were used to create special indicator tables and populate them suitably.

3.2.2.3 Joins

Another very important group of queries worth mentioning is the joins. Joins are used to join two tables based on a key.

There are 4 kinds of joins namely the inner join, the left outer join, the right outer join and the full join. The 4 joins are explained below. Let us consider two tables A and B. The intuition behind these groups of queries is that the tables are treated as mathematical Relations. These queries were useful in selecting the data required for the indicators.

INNER JOIN -The inner join selects only those tuples common between tables A and B based on a certain condition usually that the Key of A matches the Key of B, i.e. inner join is A ∩ B.

Syntax:

```
SELECT column_name(s)
FROM table1
JOIN table2
ON table1.column_name=table2.column_name;
```

LEFT OUTER JOIN -The Left outer join selects all the rows of the Left table (A) and those columns that match in the right table (B). The missing values are replaced with NULL. Left outer join mathematically is A – B.

Syntax:

```
SELECT column_name(s)
FROM table1
LEFT JOIN table2
ON table1.column_name=table2.column_name;
```

RIGHT OUTER JOIN - The Right outer join selects all the rows of the Right table (B) and those columns that match in the Left table (A). The missing values are replaced with NULL. Right outer join mathematically is B – A.

Syntax:

```
SELECT column_name(s)
FROM table1
RIGHT    JOIN table2
ON table1.column_name=table2.column_name;
```

FULL JOIN - The full join performs the Cartesian product of the tables. Mathematical interpretation of Full join is A x B (A cross B).

Syntax:

```
SELECT column_name(s)
FROM table1
FULL OUTER    JOIN table2
ON table1.column_name=table2.column_name;
```

3.2.3 Importance of Motivation Index:

Berhenke (2011) summarize motivation definition in an elegant phrase: "Motivation is that, which activates and directs behavior towards certain goals." Moreover, Gage and Berliner (1984) describe motivation as the intensity of behaviour, the direction of behaviour, and the duration of behaviour.

Before describing the indicators of motivation that is developed as a part of my TFG, it is of utmost importance to describe why indicators are important at this juncture. Let us for an example consider a tool that is used to monitor the health conditions of a person. The health conditions can be monitored by a variety of parameters, for an instance Blood Pressure, body temperature, previous medical history etc. Thus if a person is healthy or not can be answered by assessing these parameters. But in this case, the inLab's Learning Analytics project aims at measuring the motivation of students'. Motivation cannot be directly measured but can be made a function of an array of indicators. Hence, it is of a necessity to develop those indicators that characterize the motivation of students. In my TFG I have developed some of the indicators proposed. They are explained clearly in the later sections.

The first step in any analytics project is to identify the objective of the analysis, the next being identification of suitable indicators and features that aids in reaching the objective of the project, here motivation. Thus development of indicators is a crucial step in this process. Only after this step come the other intelligent data mining methods. In this project some of the indicators and their design and implementation are discussed. In the next section the various indicators that were developed in my TFG are defined and the foundations of motivation are described in the sections to follow.

Indicator design is a very creative activity and it needs to be done very carefully, by making proper and meaningful assumptions. Indicators are statistical features that are obtained from the data which can be used to characterize the data. In the learning Analytics context it is worth mentioning the following points.

1. Indicators rely on monitoring of the learning actions and the learning context. Eventually the learning patterns are decided.

2. Indicators have to adapt according to the learner's goals, actions, performance and history as well as to the context in which the learning takes place. In other words Indicators should be correct and should capture the sense of the entire data.

3. Indicators are responses to learner's actions or to change in the context of the Learning process, where the response is not necessarily immediate.

In essence Indicators identify and capture the traits of motivation from the data and can be used to represent motivation as a function of these indicators.

According to the literature (Chelladurai 2006, Scholl 2015), motivation can be decomposed into three major components, the ones regarding activity, persistence and intensity.

Activation motivation refers to a part of a motivation linked to initiate behaviour. This is motivation to start.

Persistentional motivation refers to a part of motivation linked to effort to move toward the goal even though the obstacles exist. This is motivation to persist.

Intensifying motivation refers to a part of motivation linked to the concentration and vigour that goes into pursuing a goal. This is motivation to stake of one's own effort.

Students who are not motivated think of always having other priorities. Procrastinating, Prolonging, Bad emotion associated while working, Boredom, Negative perceptual bias. The task is perceived more difficult than it is.

On the other hand this is how motivated people behave. It is a priority. I want to do this first. I want to start now. Quick. I want to finish now. Good emotion associated while working, Excitement, fulfilment, Positive perceptual bias. The task is perceived easier than it is.

3.2.4 Highcharts

Highcharts is a charting library written in pure JavaScript, offering an easy way of adding interactive charts to your web site or web application. Highcharts currently supports many chart types. It works in all modern browsers using SVG for the graphics rendering.

High charts have the ability to set defaults. Yet it is easily override-able. Suppose if you want to create a dashboard and want to create a lot of similar looking charts but with different data. Highcharts API has set of Plot Options. You can set the default look and feel setting for a particular type of chart in its plot options. We can also change this default look and feel just before creating the chart or even dynamically.

All the developed indicators are supposed to be graphically visualized using highcharts and integrated with the other components in the dashboard. But, for time being, the plots for indicators are shown using ggplot in RStudio.

Highcharts may be used for the same purpose in the future work of the Learning Analytics Project.

3.2.5 R and RStudio:

The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. R is a GNU project.

In my project, R supports the connection with databases, easy execution of logic with the fetched data, graphical visualizations using 'ggplot' and PCA analysis to find the correlation between the developed indicators and helps to decide which has to be taken for measuring motivation index.

RStudio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics.

Having this information, now it becomes easier and much faster to dive into the project and use the required platforms easily.

3.2.8 Moodle Database:

Moodle is the acronym for Modular Object-Oriented Dynamic Learning Environment. Moodle is a learning platform designed to provide educators, administrators and learners with a single robust, secure and integrated system to create personalised learning environments as it is declared in their website moodle.org. Moodle's standard database contains about 300 tables to store all information regarding the contents of a course, the interactions of the users and the grades. The database structure is defined, edited and upgraded using the XMLDB system. Performing analytics on such huge databases is complex and time consuming. Thus the larger databases are broken down into simpler databases which are easier to work with. Following shows the relevant tables from "Westeros Agora Moodle" database used for our project.

Resources
mdl_logstore_satndard_log
mdl_log
mdl_course
mdl_course_module
mdl_forum

ModuleType
mdl_assign
mdl_quiz
mdl_resources
mdl_hotpot
mdl_questionnaire

CHAPTER 4

PROJECT MANAGEMENT

The project management is the important phase of the project development. By this, we can come to know what the project should be and the planning helps us to have an idea about the project.

4.1 Scope

In this project we aim at identifying new indicators from the data available and enhance the already existing Learning analytics platform and indicators. The data is obtained from the students' moodle logs and they are stored in the moodle data base. This data is used to perform the analytics.

1. Justification: Once the project is completed, the same can be used to study the learning rates and the learning behaviour of students. The project also visualizes the results of the data into pleasing and informative graphs which makes it easier for the teachers and the analysts to draw inferences. The most important aspect of the project being the availability of indicators.

2. Product Scope: The outcome of my project is a part of the main Project. It provides a clean backend for data extraction. After ETL, which is the most important and primary step in the main project, my work starts with doing analysis, computations and visualization. Dashboards can then be made by other team members. As a result a complete product is obtained which can be used to perform learning analytics on secondary school data.

3. Acceptance Criteria: The product is to be widely used by secondary school teachers. The main aim of the system being able to extract the data, analyse and develop a friendly interface for the teachers that aid them in understanding the student's learning behaviour better. In addition the data is interpreted on various factors and the results are summed up. On successful development of the product it tested with the actual data of all the schools and the results are verified with the secondary school teachers and any change requested is appended to the already developed software.

4. Deliverables: As mentioned earlier the outcome of my project is indicators and the calculated motivation Index. This outcome is used to make a final product,

which is a web based platform that performs all the activities from the extraction of data to visualization of the results.

5. Constraints: The project's success mainly depends on the quality of the data that is obtained. This being a data analysis project the outcomes are better when the data is sufficient there by avoiding unnecessary assumptions which, leads to better promising results and inferences. Another constraint is the timeframe. All the phases of the project need to be done within the deadline.

4.2 Planning

4.2.1 Methodology

The most important step in the process is to choose the process model that we would like to follow. With the requirements of the project we choose the agile project management methodology.

Agile management or agile project management is an iterative and incremental method of managing the design and build activities for engineering, information technology, and new product or service development projects in a highly flexible and interactive manner, for example agile software development. It requires capable individuals from the relevant business, with supplier and customer input. This process model is chosen because we are enhancing an already existing system and this method allows a dynamic adaptation to any changes as per the requirements.

More about agile methodology is discussed earlier in the Methodology and Theoretical Framework chapter.

4.2.2 Task Description

Based on the objectives and observations during the course of work, I define various tasks in my project.

1. ETL Process: It stands for Extract, Transform and Load. This is the first process done on the server side.Perl Scripts are used to populate data into local Server. It is done by Jordi Casanovas on inLab, Learning analytics team. However this is the basic task that is done on the server side. Only if this is done, we are able to get the

populated data. So each member of the project should have basic knowledge of this task.

2. Developing Indicators: It is the most crucial part of my project. I developed seven key indicators, which are broadly classified into"Persistence and Choice" indicators. The development is done using SQL and R Scripts.

3. Data Visualization: To make the indicators available to the teachers and let them know their students' motivation index, the plots are made with duration of the course vs each indicator. As for time being, they are displayed locally in server of inLab. But this can be made global i.e. can be launched into the web server.

4. PCA and Motivation Index: The principle component analysis is done for all the developed indicators and the correlation between them is examined to find the index of the motivation. It is better described in the following sections.

4.2.3 Requirements Engineering

In requirements engineering, all the requirements of the project are classified into Functional and Non-functional requirements. The requirements of my TFG according to this classification are being presented in this section.

Functional requirements:

R-1 Design and development of indicators

Indicators are the statistical features that are useful to draw a conclusion on the learning style of the students in the learning platform. All the indicators are ultimately used to compute a new wholesome indicator called Motivation Index. The algorithms are designed and implemented properly. The indicators that are developed are listed below.

1. Number of Logs executed
2. Resilence Level
3. Persistence Level
4. Number of Attempts
5. Break Time
6. Curiosity Rae
7. Forum Participation

The definitions and explanation of all these indicators are clearly stated in the chapter 5.

R-2 Visualization of indicators

Graphical visualisation makes it easier and faster to have a look at each indicator and thus derive the students' information. The graphs are plotted for all the developed indicators an also for the calculated motivation index with duration of the course on x-axis and values of indicators on y-axis. The visualisation is proposed to be done using Highcharts .But finally done with ggplot of R.

R-3 Calculation of Motivation index

This is the ultimate objective of the project. After developing all the indicators, the motivation index is calculated by taking the mean of all the indicator values for a particular student in the particular course. These values are stored against each day in the course duration corresponding to each user in all enrolled courses.

Non-Functional Requirements

NR-1 Effectiveness of indicators

The indicators designed as a part of the project should be statistically efficient and correct and produce meaningful results. Not all data available may be very useful. Thus, those that are used to develop the indicators should be sensible and meaningful.

NR-2 Efficiency of Algorithms

Indicators themselves are algorithms. The best method should be to extract the statistical inference from the available data and the algorithms developed should be scalable with a minimum computational cost.

NR-3 Intuitive and interactive interface

Data visualization primarily concerns displaying huge amounts of raw data in a simple and an intuitive way to the client. Interaction of the client with the data in real time to interpret data in multiple dimensions is a key requirement of the project.

An interface that satisfies these requirements and which has a good user experience has to be designed and developed.

NR-4 Handling Large Amounts of data

This project involves large amount of data. To give an estimation of the project deals with data of around 534 students that currently contains more than one lakh rows (145248) in just one table. The table exponentially grows. The tables are updated every day by choosing the required data, the daily interactions, from the MOODLE databases by the ETL process. Thus the software developed should be capable of handling large amounts of data.

4.2.4 Gantt Chart

Initial Gantt chart:

This is only a tentative plan and the Gantt chart shows only the first few steps that the project takes. Based on the further developments a few other steps and additions can be made. This is because this being an incremental project the exact steps were not known in the beginning. A few changes have been made in the course of progress of the project. The clear distinction between the initial and final planning has been produced here. The Gantt chart for initial planning is as follows.

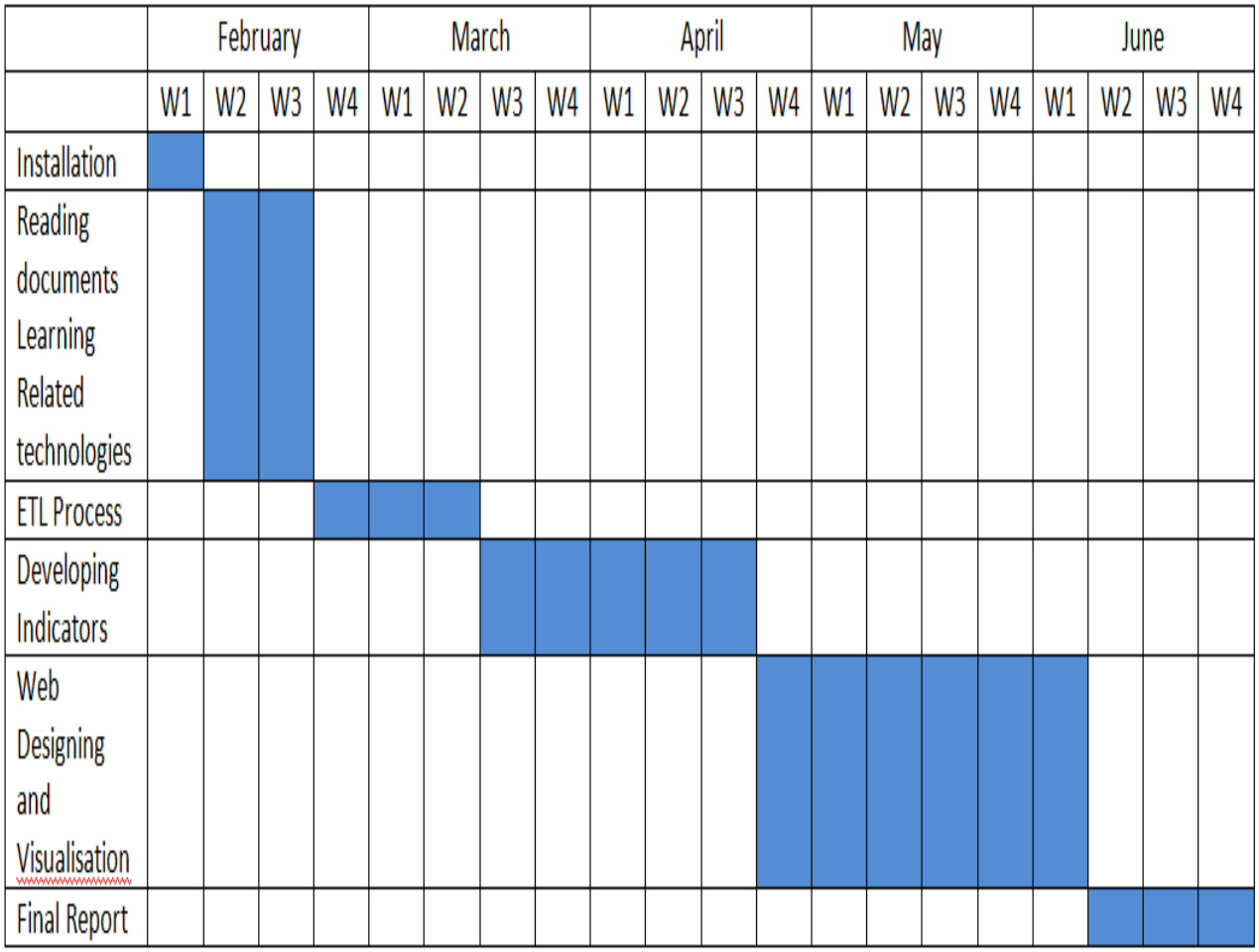


Figure 4.1 Initial Gantt chart

Final Gantt chart:

You can see that the initial planning is so linear and straight forward. But that is not the case happened in the real scenario. The tasks are happened to be simultaneous. Also the tasks described are only the proposed ones and they got change during the course of time. The Gantt chart for final planning is as shown below.

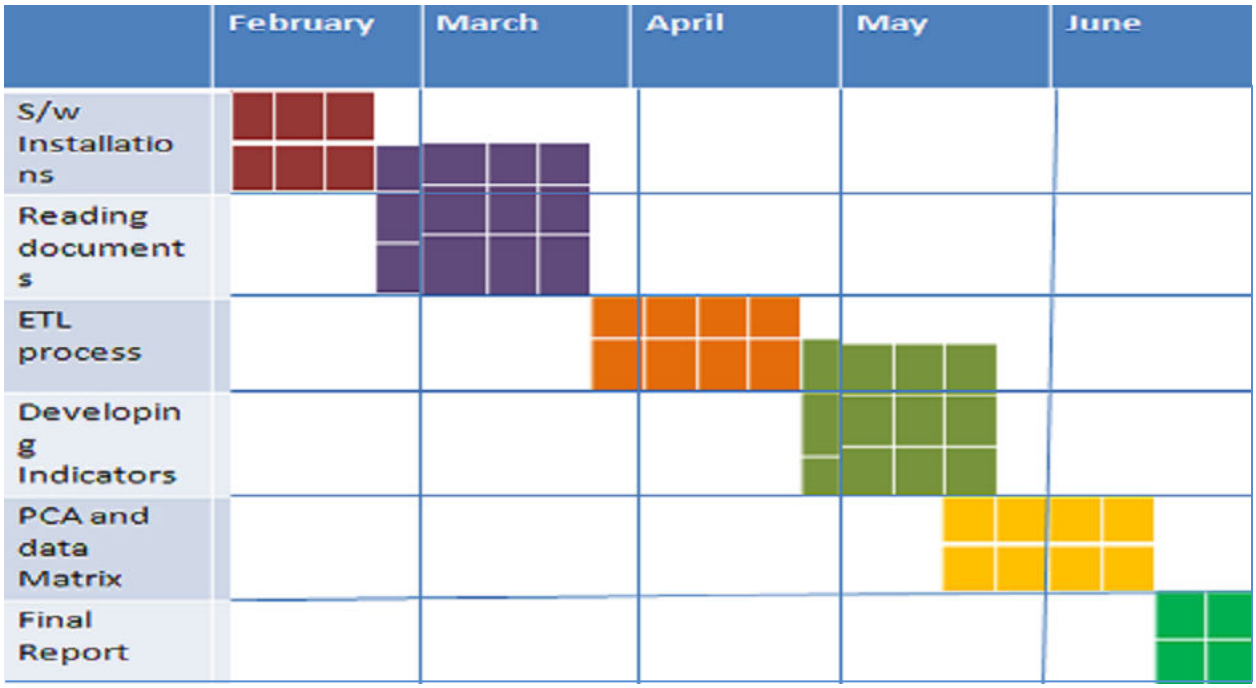


Figure 4.2 Final Gantt chart

Initially the web designing and visualization using Highcharts was proposed. But as the development of the indicators started, it is realized that the designing algorithm and logic implementation of each and every indicators takes much time. Also it is realized that the Principle Component Analysis is to be done definitely to find the correlation between the indicators. Thus the task of Graphical visualization using Highcharts of JavaScript is replaced by ggplot of R. And the Web designing part is replaced by doing PCA and calculating motivation index.

Hence the justification for change in the planning.

4.2.4 Action Plan

The various design technologies are studied and the best methodology is chosen. The data is available and the required data is alone extracted from the main database. The construction phase involves the actual coding of the software. The construction and testing may introduce a delay in the delivery of the project because they involve the actual engineering of the software. Similarly the indicator design is an activity that involves a keen observation of data and this also introduces a delay in the product delivery.

4.3. ECONOMIC MANAGEMENT

The Economic Management is one of the most important aspects to be considered for the successful implementation of any project. This document presents the analysis of estimated budget and its sustainability.It is to be noted that the budget shown here is an estimated budget and may be changed due to systems used and unexpected factors. To calculate the amortization the following factors namely, first the overall life of the hardware or software in use is considered and the project will be completed in 6 months. Hence, the amortization cost comes one eighth of the actual life of the component.

4.3.1 Identification of costs

The project's requirements, functionalities and the final output are keenly studied and all the elements that affect the project development (directly and indirectly) are considered. Elements like labour costs, licence costs, infrastructure costs, maintenance costs etc. are taken into account. The total expenditure is then divided into three parts. They are 1.expenditure on Software resources 2.expenditure on hardware resources and 3.expenditure on human resources.

Expenditure corresponding to each task

Some of the software's installed are open source and some are licensed. For learning different technologies, I have made use of some online tutorials which are paid resources. For the ETL process, most of the work is done in already installed softwares.For designing new indicators, a lot of human resources are spent along with hardware and software resources. Finally, for visualisation and calculation of motivation index, free software's are used and is deployed in the main server.

4.3.2 Budget Estimation

In each category, a detailed description of costs for every component is estimated with an objective to analyse the budget accurately and optimize it. To calculate the amortization we consider the useful life of component and total project time.

4.3.2.1 Software Resources Cost Estimation:

The first table contains the estimated costs connected with licenses for used software. The majority of software used is open source and for free. However cost of even a few software is high as seen in the table.

Software Component	Total cost	Amortized cost
Microsoft Windows 8.1	150 Euros	25 Euros
Database Browser	Open source	-
RStudio	Opensource	-
Microsoft Word	70 Euros	11,70 Euros

Table 4.1 Software Cost Estimation

4.3.2.2 Hardware Resources Cost Estimation:

Servers with high performance are required for the project. A desktop PC, along with all the required peripherals is provided for each person of the team. Table 4.3.2.2 shows total cost, life span and the amortized cost of computers and the servers.

Hardware Component	Useful life	Total cost	Amortized cost
Personal computers	4 years	1000 Euros	125 Euros
Servers and disk space	8 years	2500 Euros	156,25 Euros

Table 4.2 Hardware Cost Estimation

4.3.2.3 Human Resource Cost Estimation:

The last but not least issue are the human resources. I have estimated costs of my work during 18 weeks and two professors of UPC with whom I was consulting the project

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

Table 4.3 Human Resource Cost Estimation

4.3.3 Budget Control management

All the cost estimation is done based on the current state of affairs. The material resources are standard and there would be no deviations. In case of any hardware failure, repair or replacement must be done. It is unlikely that we need any hardware component apart from the resources that have been listed in the budget estimations. In the case of human resources, the cost will be constant during our project tenure. After that, other interns may take up the task of enhancing the learning platform.

4.4 SUSTAINABILITY

The success of the project hinges on the ability to intelligently evaluate and prioritize sustainability investments to achieve the greatest impact. In this section we are going to evaluate the sustainability of our project in three areas: economic, social and environmental.

4.4.1 Economic sustainability

Economic sustainability deals with using various strategies for employing existing resources optimally to ensure maximum benefit. The specified budget plan takes into account all the factors and estimates a feasible project cost. It is impossible to finish this project at a lower cost. It involves mostly open source software and basic hardware equipment. Also the time spent for each task is based on its priority. As stated before, this project involves enhancing the existing learning analytics platform and it is a part of what is being carried out already.

4.4.2 Social sustainability

The project aims at enhancing e-learning platform used by all secondary schools in Catalonia. It analyses students' study patterns and provides feedback to teachers and students. This improves quality and purpose of education. The current platform has no features to provide any motivation or feedback which is a big limitation considering the amount of data that is available from Moodle logs. This project improves current learning scenario and does not harm any collective.

4.4.3 Environmental sustainability

The servers and the computer hardware that are being used throughout the project, will be reused by other people and at the end of the hardware's lifespan, the components would be taken to a recycling centre. Assuming the power consumed by a single computer is 250 watts, total energy expended would be 125KW. This amounts to 48.125 kg of CO2, which is within the permissible limits. So the ecological footprint is deteriorating, and it is inevitable in a Computer Engineering project. But the environmental loss is kept to a minimum. All the project code is open source and is reusable for any further enhancements or studies.

4.4.4 Sustainability Matrix

A sustainability matrix is prepared in accordance with guidelines provided by Christian Felber and scores are given for the three perspectives: Economic, Social and Environmental for the planning phase as shown in the table 4.4. The scores assigned are in accordance with the description given on sustainability in

each area. This reveals that the total score is 47 out of 60. So the project is highly sustainable.

Sustainable? **Planning**	**Economic** Economic viability	**Social** Improved quality of life	**Environment** Resource analysis
assessment	8	8	6
Outcomes	Final costs vs forecast	Impact on social environment	Resource consumption
assessment	9	10	6
Risks	Adapting to changes of scenery	Social damage	Environmental damage
assessment	0	0	0
Total	47		

Table 4.4 Sustainability Matrix

CHAPTER 5

DESIGN AND IMPLEMENTATION

This section is the most important section. All the technical work, I have done is included in this section. I discussed mainly the development of seven indicators, visualizing them, integrating them into a single table in the database and finally calculating the motivation index in this section.

Note that the mandatory tasks, (which I am going to refer in this section) are the tasks that must be done by the students. These tasks will have the deadline value. The modules like assignment, quiz and hotpot are considered as mandatory tasks. The opposite is the non-mandatory tasks. These are optional. These can be done or not. There is no compulsion in doing them. All the modules other than assignment, quiz and hotpot are considered as non-mandatory tasks.

The following is the table containing set of indicators that I have calculated and their brief details.

Rows no.	Ñame of te Indicador	Formal Definición	Adicional Información	Statistical definition
1.	Number of logs executed.	The total logs executed by a student in a course.	It is calculated per all tasks per each day.Comes under Persistence Indicators.	#sum(logs)
2.	Resilence Level	Maximum number of tasks	It is calculated for all tasks in	#Max(tasks)in 2 hrs

		performend in an interval of two hours in a day	each course for each student per day. comes under Persistence Indicators.	
3.	Number of attempts	Total number of attempts to finish a task	It is calculated per mandatory tasks ,not per day. comes under Persistence Indicators.	#Sum(attempts)before closing
4.	Persistence Level	The mean of the timespent in completing a task	It is calculated for mandatory tasks,per course.Not per day. comes under Persistence Indicators.	$\frac{\text{Timespent}}{\text{No.Of.Attempts}}$
5.	BreakTime	The total time doing nothing in between the first access and deadline	It is calculated for only mandatory tasks,per day. comes under Persistence Indicators.	Deadline-Firstaccess-Timespent
6.	Curiosity Rate	Number of accesses to	It is calculated for	#Accesses to non.man taks

		non-mandatory tasks to all assigned non-mandatory tasks	only non mandatory tasks,per day. Comes under Choice indicators	Total assigned non-man.tasks
7.	Forum Participation	Measures the participation in a forum related to activities	It is calculated for the forum activities,per day.Comes under Choice indicators	0, not participated 1,viewed 2, activities other than viewed

Table 5.1 List of Indicators developed

The Design and implementation of each indicator is as follows:

5.1 Design and Implementation of Indicators

5.1.1Number of Logs Executed

The number of logs executed is the simple but important indicators. It indicates the number of logs (any activity) executed by a student on a particular day, in any enrolled subject is taken as the number of logs.

The heuristic followed for the number of logs executed indicator is “more the number of logs executed, more motivated the student is.”The justification for this is as follows. If this indicator value is high, it means that the student is doing more number of activities in that course. Thus he is more interested to do the tasks in that course. But there are some opposing cases for this heuristic. For example, let a student is distracted by some other work while doing a particular task. He resumes the task and again he went to do some other task. If the same repeats, then the value for number of logs executed will be high but he is not actually motivated.

However, the possible cases that justify the above mentioned heuristic are more than those that oppose. Thus it is the believed heuristic.

The algorithm for calculating the number of logs executed is simple. It is as follows

```
NoOfLogs()

1. Begin
2. For Each Day Di ∈ mdl_logstore_standard_log l
    a. For Each course Cj
        i. For Each Student Sk ∈ Cj
           1.InsertIntoLogs(Di,Cj,Sk,count(activities))
        End Loop
    End loop
End Loop
End
```

Algorithm 5.1 NoOfLogsExecuted

Note that the mdl_logstore_satndard_log is the table in the database into which the datails of the student are updated every day. So for calculating almost all the inducators, data is taken from this table. As per the algorithm,this indicator value is calculated for all the courses that have entry into the mdl_logstore_satndard_log on that particular day.
The SQL code snippet for this insertion is as follows:

```
INSERT INTO NoOfLogs SELECT
from_unixtime(m.timecreated,m.courseid,m.userid,
m.count(*)) GROUP BY m.date,m.courseid,m.userid FROM
mdl_logstore_satndard_log m
```

Date	courseid	userid	Cmid	noOfLogs
9/15/2015	399502	399341	3994461	14
9/23/2015	399502	399341	3994470	1
9/29/2015	399502	399341	3994470	1
9/30/2015	399502	399341	3994478	2
10/6/2015	399502	399341	3994461	18
10/13/2015	399502	399341	3994461	7
10/14/2015	399502	399341	3994461	8
10/28/2015	399502	399341	399181841	4
11/24/2015	399502	399341	3994505	10
12/1/2015	399502	399341	3994461	10

Table 5.2 NoOfLogs Simulation

The above table is the simulated table for this indicator using the above algorithm. The graph plotted with duration of the course vs this indicator values is shown below.

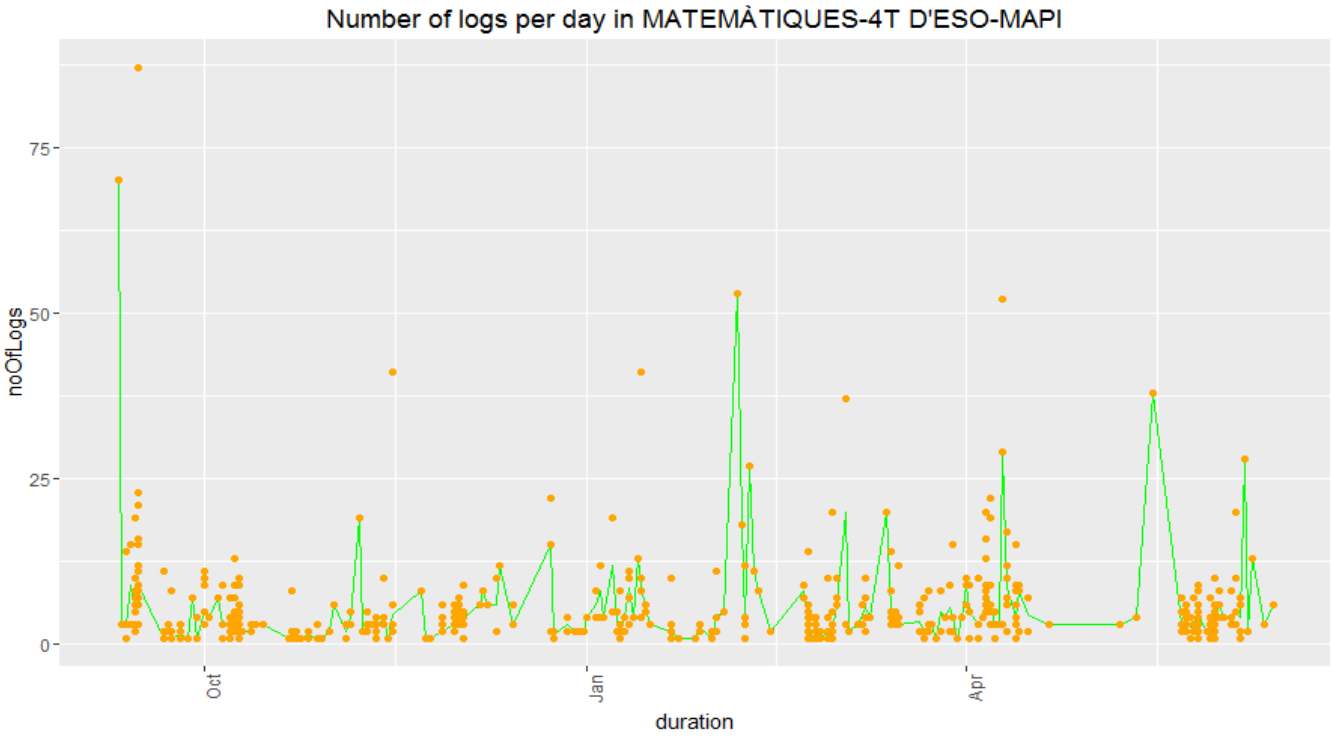


Figure 5.1 Visualization of NoOfLogs indicator

5.1.2 Persistence level

Persistence Level, as the name indicates, is used to see how persistent a student is in doing the tasks. Persistence level can be measured as the ratio of total time spent by a student in a course task to the total number of attempts he made to complete the same task. In other words, Persistence can be taken as the mean time a student spent on one task.

```
Persistence Level = time spent/number of attempts
```

According to the formula, more the time spent by a student and in less attempts, that means that he has a tendency to work continuously, taking short number of breaks. Thus he is more persistent in completing the task and hence we can conclude that he is more motivated.

In the case, if the total time spent is less and the number of attempts is also more, the persistence level will be moderate. We may conclude that he is moderately motivated. But that need not be the only conclusion. Other possibilities are also there which are against our heuristic. Those are discussed in the later sections.

Other possible case is, more the time spent and more number of attempts made, then the student is not enthusiastic in completing the task. He wants to procrastinate and not motivated to do the task.

Considering all the cases, the heuristic followed for this indicator is "more the persistence level, more the student is motivated."

However there are limitations to this heuristic, which are discussed in the following chapter.

Coming to the design of this indicator, the persistence level is developed only for the mandatory tasks. Mandatory tasks are the tasks which must be completed by the student once assigned. They will have deadline and must be completed within that deadline.

The total time spent by the student on the moodle platform is made available and updated every day. The time spent table has entries in a day only when on that day, the task is completed.

Each day the time spent table of the database is checked if there are any new entries. For those entries alone the number of attempts is taken from the number of attempts table and persistence level is calculated using the above proposed formula. This value is assigned for that day. Likewise it is calculated for all the days in the academic year. As mentioned several times, the date is dynamically updated. I.e. every day the calculation of persistence is to be done for the newly added entries and integrated with the old values.

The algorithm for calculating the persistence level is as follows.

Algorithm 4: Persitence Level

```
PersistenceLevel ()
1.Begin
2. For Each Course Ci  and Mandatory task t1
```

```
        i.For Each Student Sj ∈ Ci
            1.Check 'Timespent' for any new entry
            2. IF(entry exists)
                a. Check No.of.attempts table for
                corresponding Sj,Ci,
                b. If (value exits)
                    i. persistence= timespent/no.of
                    attemps
                    ii. INSERT into persistence (Ci,
                    tk, Sj,persistence)
                c. Else
                    i. Report Error
                    ii.Break
                d.End If
            3. Else
                a.Break
            End If
    End loop
End loop
End.
```

Algorithm 5.2 PersistenceLevel

From the algorithm, it is clear that it is calculated for each student in each enrolled subject and every day from the beginning of the academic year.

The table creation is done using SQL commands. The important code snippet for this table is as follows.

```
INSERT INTO P_Persistence1
SELECT DISTINCT b.course,b.cmid,b.userid, b.time/p.attempts
FROM Balaji_timespent b LEFT JOIN P_NoOfAttempts1 p
ON b.course=p.courseid and b.cmid=p.cmid and b.userid=p.userid
```

The final output table for the persistence table is as follows:

Courseid	userid	Cmid	persistence
1518184	15181368	151817201	3.5
1518184	15181400	151817201	1.25
1518184	15181607	151817201	0.75
1518184	15181645	151817201	1.5

1518184	15181681	151817201	4.5
1518184	15181684	151817201	2
1518184	15181703	151817201	3.5
1518184	151827	151817201	7
1518184	15181368	151817201	2.5
1518184	15181382	151817201	1
1518184	15181400	151817201	1.75

Table 5.3 Simulation of Persistence values

As only the records which have the corresponding entries in the timespent and number of attempts table are take into this table, the persistence value for other dates can be considered as NA.

Another important thing to be noted is that since the number of attempts is the total number of times a student work irrespective of date, i.e a student may complete a task in 2 days, other in 5 days, some other in 2 days but not consecutive days. So it can't be possible to include date in the number of attempts table. Same is the case with timespent table also. Thus, while combining these two table to calculate persistence, it is not possible to correlate persistence with the dates. That means persistence cannot be corresponded to a particular date. It corresponds to particular student in enrolled course. However to have the graphical representation of this indicator, the date when a new entry is added into timespent table is corresponded to that task's persistence in a course.

The graph visualization for this indicator is as follows.

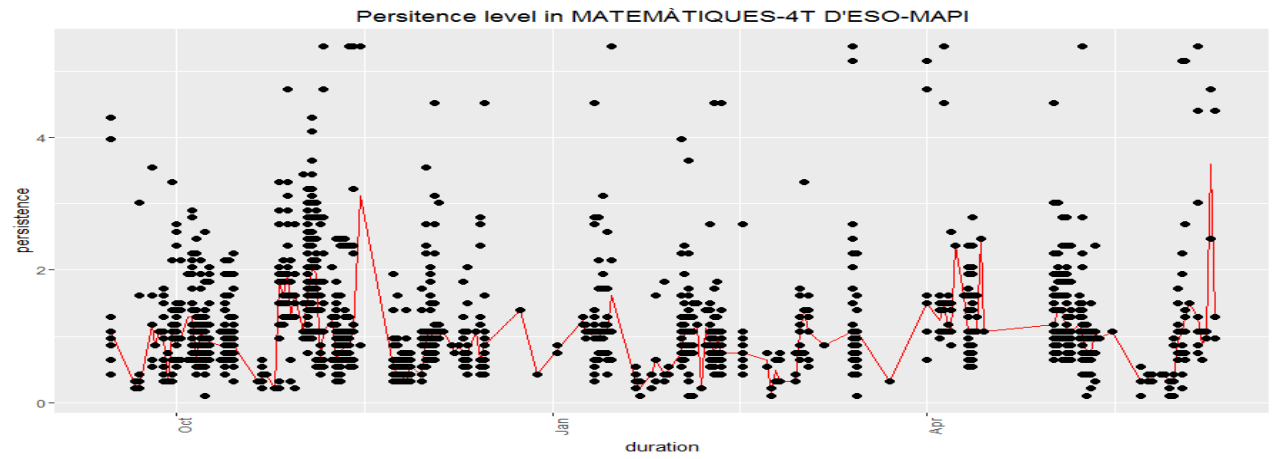


Figure 5.2 visualization of Persistence Level

5.1.3 Number of Attempts:

The number of attempts indicator is a very good factor for deciding how persistent a student is. If the number of attempts are more, that means that there are lot of breaks taken and lot of obstacles in completing a task. This is the assumption and happens in most (not all) of the cases. Thus the heuristic followed is "more the number of attempts to complete a task, more the possibility of not being motivated".

The number of attempts indicator is calculated only for the mandatory tasks. Taken a task in a course, done by a particular student, if the action in the mdl_logstore_standard_log table is 'close task', the total count of actions before closing the task is computed. It is taken as the number of attempts value. But the limitations in computing the number of attempts is that the attempts made after the deadline of the assigned tasks are ignored. since it is the total count of attempts in task, it is irrespective of the date. The problem of date and number of attempts correspondence is discussed in the validation section clearly.

The algorithm for calculating number of attempts :

NoOfAttempts()

```
1.Begin
2. For Each Course Ci and Mandatory task tkin
mdl_logstore_standard_log
        i.For Each Student Sj ∈ Ci
            1. Count=0
            2. while (l.action='close task')
                a. Count=Count+1
                    b. INSERT INTO NoOfAttempts
    Values(Ci, tk, Sj,Count)
            End While
        End Loop
    End Loop
End.
```

Algorithm 5.3 NoOfAttempts

The query used for the simulation of number of attempts is :

```
>INSERT INTO NoOfAttempts select courseid, userid, cmid,
count(*) FROM mdl_logstore_standard_log GROUP BY
coursed,userid,cmid ORDER BY courseid WHERE action NOT LIKE
'%close%'
```

The simulated table for the same algorithm is as follows.

courseid	userid	Cmid	attem pts
399502	399341	3994461	10
399502	399341	3994470	7
399502	399341	3994471	11
399502	399341	3994485	2
399502	399341	3994486	2
399502	399341	3994490	1
399502	399341	3994491	6
399502	399341	3994492	2
399502	399341	3994494	1
399502	399341	3994497	1
399502	399341	3994502	1
399502	399341	3994503	4
399502	399341	3994504	1
399502	399341	3994505	1
399502	399341	3994510	1

Table 5.4 NoOfLogs Simulation

As there are no corresponding date values for the number of attempts in this table, the graphical representation of the number of attempts is calculated against the day to start on values of each course task. However for performing PCA, this indicator value is again calculated per day. The Graphical visual result is given below.

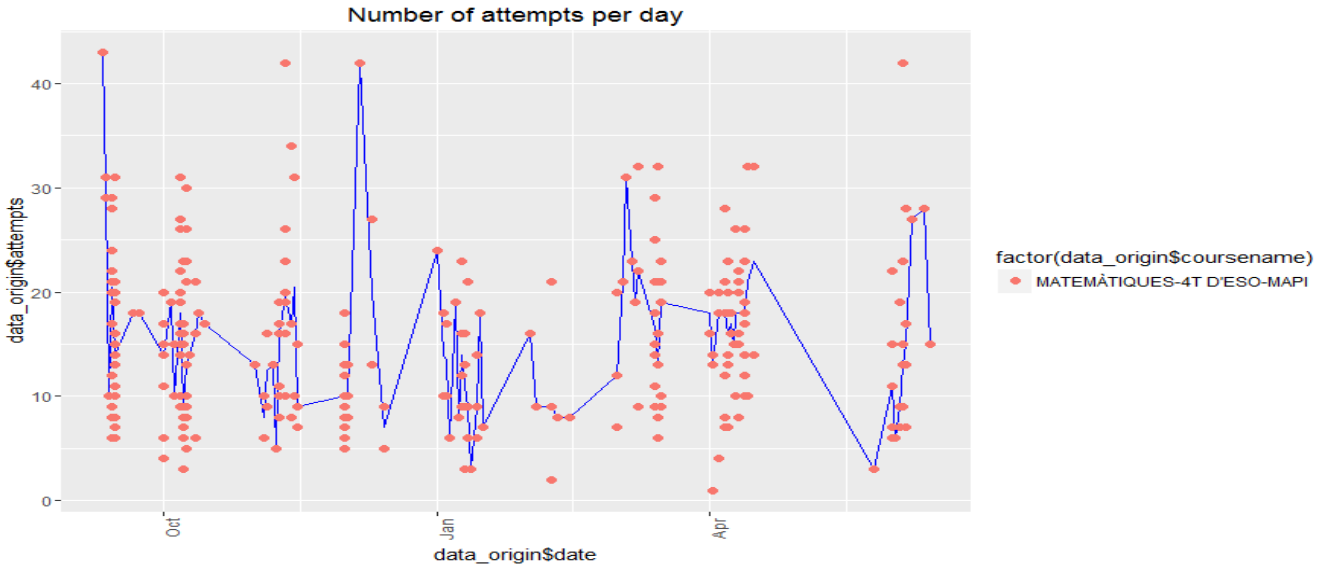


Figure 5.3 NoOfAttempts visualization

5.1.4 BREAK TIME

Break time can be defined as the total time a student didn't do anything after starting the task and before submitting the task. It is calculated only for the mandatory tasks. Mathematically, Break time can be defined as,

```
Breaktime= Deadline - timespent  -  firstaccess
```

Since the Deadline ,first access are used for the calculation of the break time,it can be inferred that the break time is calculated only for the mandatory tasks as non mandatory tasks do not have a compulsory values for these factors.

The heuristic that is followed for the break time indicator is "Lesser the break time, More the Motivated"

The Algorithm for calculating the break time is as follows

```
Breaktime ()
1.Begin
2.For Each course Ci in the course table
    a.∀ Each student Sk
        i. For mandatory tasks tj
        1.If(Deadline value exists in deadline table)
```

```
            i.DeadLine  =SELECT  Deadline  from  that
table.
        2.ELSE
            ii. DeadLine= Date_to_start_on + 7 days
        End If
    ii.End For
    b.date=Current date
    c.If(DeadLine=date)
    i.if(timespent  and  firstaccess  value  exists  for
this date)
        1.breaktime= DeadLine – timespent - firstaccess
    ii. Else
        1.Report Error
        2.Break
    End If
    d.Else
    i.Break
    End If
End Loop
End.
```

Algorithm 5.4 BreakTime

As mentioned in the algorithm 2.a.i, in each course, the deadline may exists or may not exists due to some errors in the deadline table. In the later case,the date to start on for the task is added with an estimated window of one week and taken as the deadline. Each day the deadline table is checked if any tasks have deadline on that day. If so, the break time for that is calculated for that day.

The graphical representation of break time in one course is shown below with the median line and breaktime values of one user is shown below.

The query snippet for getting these values is

```
create  table  P_Breaktime  select  d.deadLine  as
date,d.courseid,b.userid,d.cmid,d.moduleType,(d.deadLine-
b.FIRSTACCESS-  b1.timespent)/86400  as  breaktime  from
Balaji_deadLine d left join (Balaji_FA b ,Balaji_TIMESPENT b1)on
(d.courseid=b.courseid  and  d.courseid=b1.courseid  and
d.cmid=b.cmid and b.userid=b1.userid and d.cmid=b1.cmid)
```

Date	courseid	Userid	Cmid	moduleType	Braktime
9/22/2015	2863023	286353	286163566	assign	6.64023148148148
9/22/2015	2863023	286379	286163566	assign	5.39011574074074
9/22/2015	2863023	286415	286163566	assign	6.99270833333334
9/22/2015	2863023	286421	286163566	assign	6.96819444444444
9/22/2015	2863023	286429	286163566	assign	6.95126157407407
9/22/2015	2863023	286447	286163566	assign	6.99167824074074
12/12/2015	486922	486833	48679440	assign	4.85260416666667
12/11/2015	1152362	1152183	115211041	Quiz	0.58775462962963
12/13/2015	1152362	1152211	115211041	Quiz	6.69561342592593
12/30/2015	1152362	1152226	115211001	assign	1.8834375

Table 5.5 BreakTime Simulation

The graphical representation of the break time over time is as follows

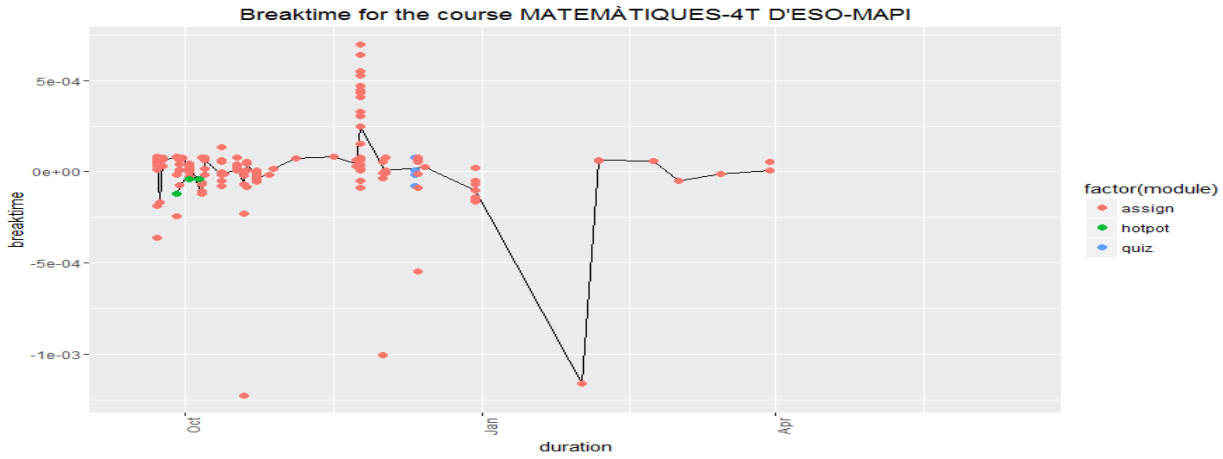


Figure 5.4 Breaktime visualization

5.1.5 RESILIENCE LEVEL

This section discusses an indicator namely the resilience level. This indicator is a measure of persistence. This is a very intuitive indicator because it is an evidence of persistence. The persistence to do an activity or a course directly reflects how well a student is motivated. Thus in this indicator the persistence is

translated into a measure of motivation per course. The working of the indicator is demonstrated below.

In this indicator the accesses of a student on a particular day is measured. It checks in every two hour window the activity of a student. That is in a given series of activities, the number of breaks a student takes is computed. The length of the longest such sequence, without any breaks, is computed. For a clearer understanding refer to the algorithm and the simulation of a sample output.

The peak two hours is calculated taking into account the starting time values of each tack in that day, in a course for a student. Starting with the first task (earliest task),its start time is taken and 2 hrs is added to it.In this period the total number of tasks performed is counted.Again next task's starting date is taken and the same process is repeated each time till the last task.Each time the count of tasks is checked and if it is greater the max. value,it is assigned to max value.Initially the max value is assigned to zero.

This also has small design issue. This actual efficiency and the performance of this indicator could be justified only if a student has enrolled into more than one course, because the persistence could be checked if the student sticks on to one course before he moves to the next one. In the inLab's project only a pilot version of the platform is created i.e. only one course from each school is used for the verification of the platform.

A student has taken a break, if anyone of the two things happen,

1. The difference between the current and the next activity time stamp is greater than two hours.
2. If the current and the next activities belong to different subjects.

```
Resilience=max(tasks performed in peak 2 hours)
```

The resilience level is measure for each course, each student on a daily basis. The algorithm for Resilence level is as follows

```
Resilence Level()
1.Begin
2.For Each course Ci
    a.For Each Student Sj € Cj
    b. For each day Dk
        i. count= number(tasks in Ci )
        ii for l in 1 to count
```

```
            1.temp[l] =start_time (task_l)
        End Loop
        iii. max=0
        iv. for l in 1 to count
            1.time=temp[l]
            2. time2=temp[l]+2
            3.c1=0
            4.while(time<time2)
                a. if( any action of a task exists)
                    i.c1=c1+1
                End if
            End while
            5.if(c1>max)
                a.max=c1
            End If
        End  loop
    End Loop
End Loop
End.
```

Algorithm 5.5 Resilence Level

The simulated table is as follows.

date	Courseid	Userid	Resilience
9/10/2015	286345628	2863041	0
9/10/2015	286345628	2863301	1
9/10/2015	48376989	4835621	1
11/10/2015	78676789	7859930	4
11/10/2015	115178098	11517234	0
11/10/2015	86456890	864262	0

Table 5.6 Resilience Level Simulation

The computation is done entirely in R and the simulated table for resilience level shown above. And thus the code snippet is not included.
The graph for Resilence level is as below.

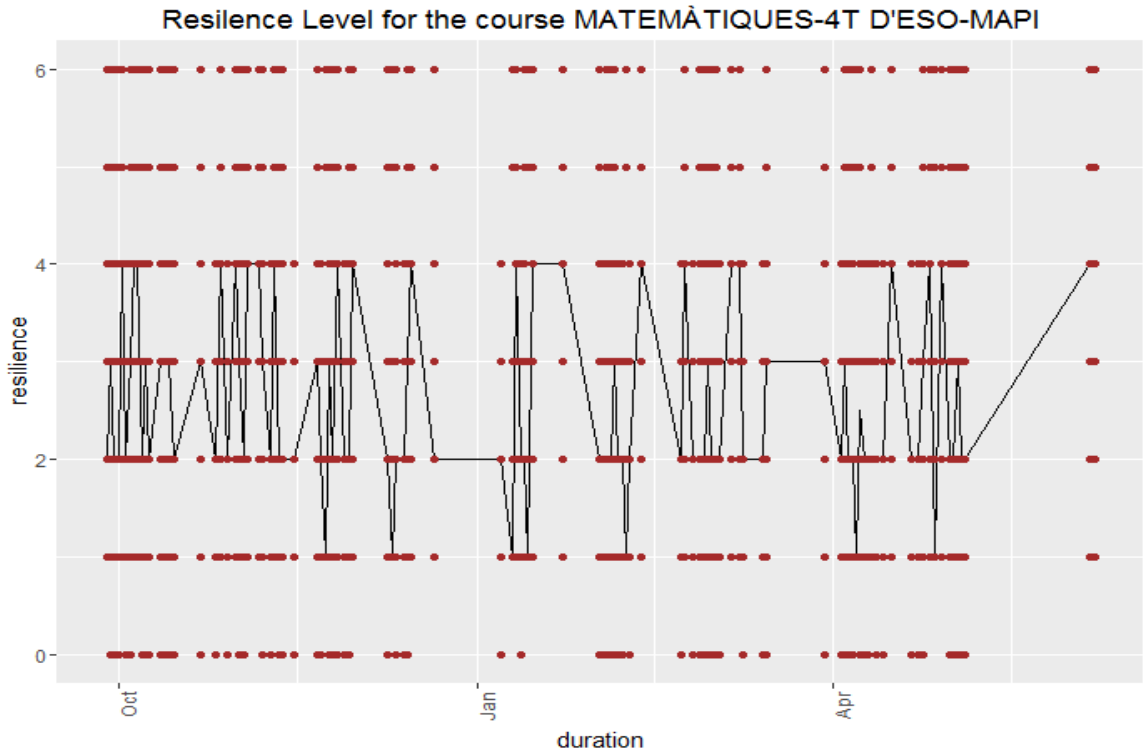


Figure 5.5 Resilience Level Visualization

5.1.6 Curiosity Level

This is the other indicator helping the teacher know how curious a student is in a subject.It is computed as the ratio of all accesses to the non-mandatory tasks to the assigned non mandatory tasks in a course. This indicator is very useful as the students who are performing non mandatory tasks obviously are interested and motivated in that subject.

This is calculated per day per all non mandatory tasks. The algorithm for finding the curiosity rate is as follows.

CuriosityRate()

```
 1.Begin
2. For Each Course Ci and Mandatory task tkin
mdl_logstore_standard_log
        i.For Each Student Sj ∈ Ci
            1. Count1=sum(tasks)in
            mdl_logstore_standard_log table
            2. count2=sum(tasks)in Day to start on
            table
            3.curiosity =count1/count2
```

```
            End While
        End Loop
    End Loop
End.
```

Algorithm 5.6 CuriosityRate

Code Snippet

```
insert into P_Curiosity select
p1.date,p1.courseid,p1.userid,p2.type ,p1.c1/p2.c2 from P_Cur1
p1 join P_Curiosity2 p2 on p1.courseid=p2.courseid
```

The simulated table based on this algorithm is as follows.

date	courseid	Cmid	Curiosity	userid
9/15/2015	399502	mdl_forum	0.666666666666667	39917021
9/15/2015	399502	mdl_forum	0.333333333333333	39917021
9/15/2015	399502	mdl_forum	1.33333333333333	39914906
9/16/2015	399502	mdl_resource	0.2	399341
9/16/2015	399502	mdl_resource	0.4	39917021
9/16/2015	399502	mdl_resource	0.6	39914903
10/7/2015	399502	mdl_url	0.333333333333333	39914901
10/7/2015	399502	mdl_url	0.333333333333333	39914902
10/7/2015	399502	mdl_url	0.666666666666667	39914907
10/7/2015	399502	mdl_url	1	39914906
10/7/2015	399502	mdl_url	0.666666666666667	39914908
10/27/2015	399502	mdl_resource	1	39914903
10/27/2015	399502	mdl_resource	1	39914902

Table 5.7 Curiosity rate simulation

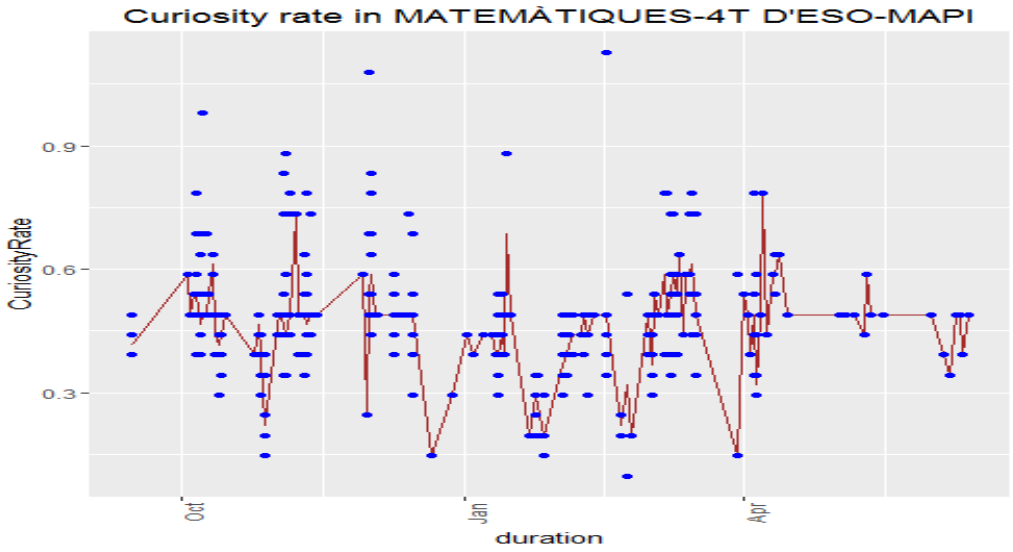


Figure 5.6 Curiosity Rate

5.1.7 Forum Participation

This is one of the important Choice indicators. It measures the accesses (participation) in a Forum related to activities of a given subject on particular day. It reflects non-expected but desired performance.

If a student is interested to participate in a Forum,it simply implies that he is more enthusiastic to learn more in that subject.He has the tendency to get self-motivated and gain extra knowledge than the peer. That means he is interested in the domain and has chosen to learn more in that particular topic. No one will access the forum every day.If a person accesses on a day, obviously he is more motivated than others on that day.

This is the simplest indicator to calculate. Yet the important indicator to be considered.

forumPaticipation=	0 , if not participated 1 , if only accessed to view 2 , if participated to post,create,discuss etc.,

Daily the log table called `mdl_logstore_standard_log` is checked for 'forum' module component. If the action is `view`, then the value assigned is 1. If there is any action other than view ,the value assigned is 2.If there is no entry in the log table for a student in that course on a day, then 0 is assigned correspondingly.

The algorithm for the forum participation is as follows.

forumParticipation()

```
1.Begin
2.For Each Day Di ∈ mdl_logstore_standard_log l
    a. For Each course Cj corresponding to Di
        i. For Each Student Sk ∈ Cj
            1. Select records where l.module=`Forum`
            2.  Check if l. action=`%view%`
            3. If (YES)
```

```
                    i. INSERT Values (Di, Cj, Sk, l.action,
                    1) into P_Forum
                4. Else
    i. INSERT Values (Di, Cj, Sk, l.action, 2) into
    P_Forum
        End Loop
    End Loop
End Loop
3. ∀ days ∉ mdl_logstore_standard_log l
    a. INSERT Values (Di, Cj, Sk, l.action , 0) into
P_Forum
4. End
```

Algorithm 5.7 ForumParticipation

The simple sql query snippet used for obtaining the table is

```
>INSERT INTO P_FORUM_P select d.date,l.course,l.userid,l.action
FROM p_all_dates d
LEFT JOIN  mdl_logstore_standard_log
ON d.date=l.date and d.courseid=l.course and d.userid=l.userid
>UPDATE P_FORUM_P set p_value=2 WHERE prtcptn NOT LIKE '%view%'
> UPDATE P_FORUM_P set p_value=1 WHERE prtcptn  LIKE '%view%'
>UPDATE P_FORUM_P set p_value=0 WHERE prtcptn IS NULL
```

The final obtained table for the Forum participation is as follows

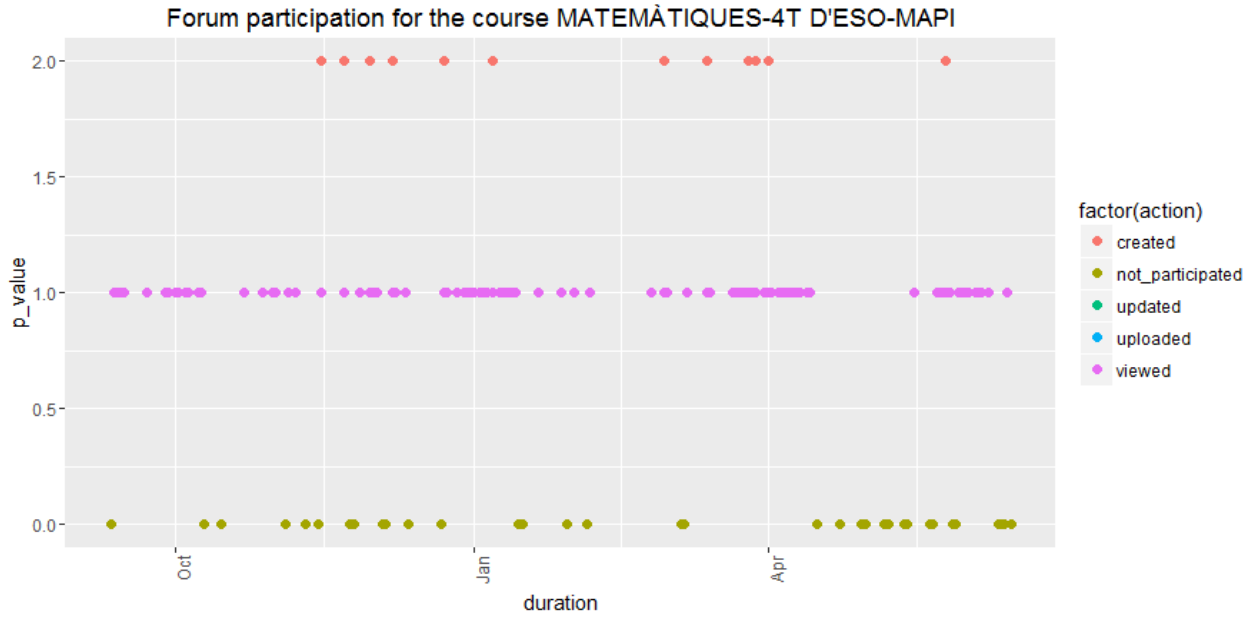
Date	course	userid	Prtcptn	p_value
9/10/2015	15181242	15182501	Viewed	1
9/11/2015	2863005	28668	not_participated	0
9/13/2015	2863023	28668	Viewed	1
9/24/2015	486922	4861962	Viewed	1
9/26/2015	486922	4866381	Viewed	1
9/28/2015	2863023	28668	Viewed	1
9/30/2015	486922	4863462	Viewed	1
10/1/2015	2863023	286374	Viewed	1

10/10/2015	486129	48614721	not_participated	0
10/25/2015	486922	48612101	Viewed	1
10/26/2015	486922	4868081	Uploaded	2
10/27/2015	486922	4868301	Viewed	1

Table 5.8 Forum Participation

The table contains values for all the days for all the students in all the enrolled courses. This data is used to represent the Forum participation graphically. The x axis is the date and y axis is the p_value from the table. The representation can be at different filters like we can see for all students in one particular course or one student in one course or one student in all courses etc.

The graph for all the students in one course is included below.



Please not that the line in all the graphs except in forum access represents the median line.Like wise the graphs can be created for any student in any course using ggplot in RStudio.

5.2 MEASUREMENT OF MOTIVATION INDEX

Thus the motivation index can now be measured as all the indicators have been developed and the motivation index can be estimated based on these results.

Based on these definitions, characteristics and constraints, a system of indicators was developed to track motivation, where motivation is a function of several indicators and can be formulated as following:

M = f(x1, x2, x3, x4, x5, x6, x7, x8)

where n is the total number of considered indicators

Moreover, the linear correlation between motivation and the indicators was tested. And the model took the form of:

$$M = \sum_{i=1}^{n} \beta i Xi$$

Here M refers to the "motivation index" and the Xi refers to the ith indicator and the βi refers to the coefficient of the indicator, defining its importance. This βi is computed by performing PCA on the big tables formed by the indicator. Supplementary variables are student, date and course. Motivation is calculated for each student, per day, within each course.

S = {s1,s2,…,ss} s – number of students

D = {d1,d2,…,dd} d – number of days

C = {c1,c2,…, cc} c – number of courses

The indicators are extracted from the suitable databases and designed that aids in the measurement of motivation.

The more about PCA is discussed in the later section.As of now, the motivation index is calculated taking the mean of the indicators.

To summarize, all the indicators corresponding to all dates,all students in all courses are taken into a dataframe .The subset of this dataframe containing only the indicators is taken and is scaled. The mean value for this scaled indicators is calculated and taken as the motivation index.These values are stored in a dataframe and this new dataframe is appended with the original dataframe.

The database table simulated by this calculation is as follows.

Date	SubjectID	SubjectName	StudentID	motivationIndex
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	151824	1.37576243622889
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181296	1.37576243622889
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181344	1.37576243622889
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181365	1.37576243622889
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181389	1.37576243622889
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181427	1.37576243622889
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181444	1.37576243622889
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181464	1.37576243622889
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181465	1.37576243622889
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181480	4.31572001228296
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181544	1.37576243622889
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181573	2.65497079075975
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181575	1.37576243622889
2015-09-15 00:00:00	15181242	400 Informàtica 4t ESO. Projecte mSchools. (català)	15181633	1.37576243622889

The graphical representation of the motivation index over time is as follows. This is calculated for one course for all the dates in the academic year.

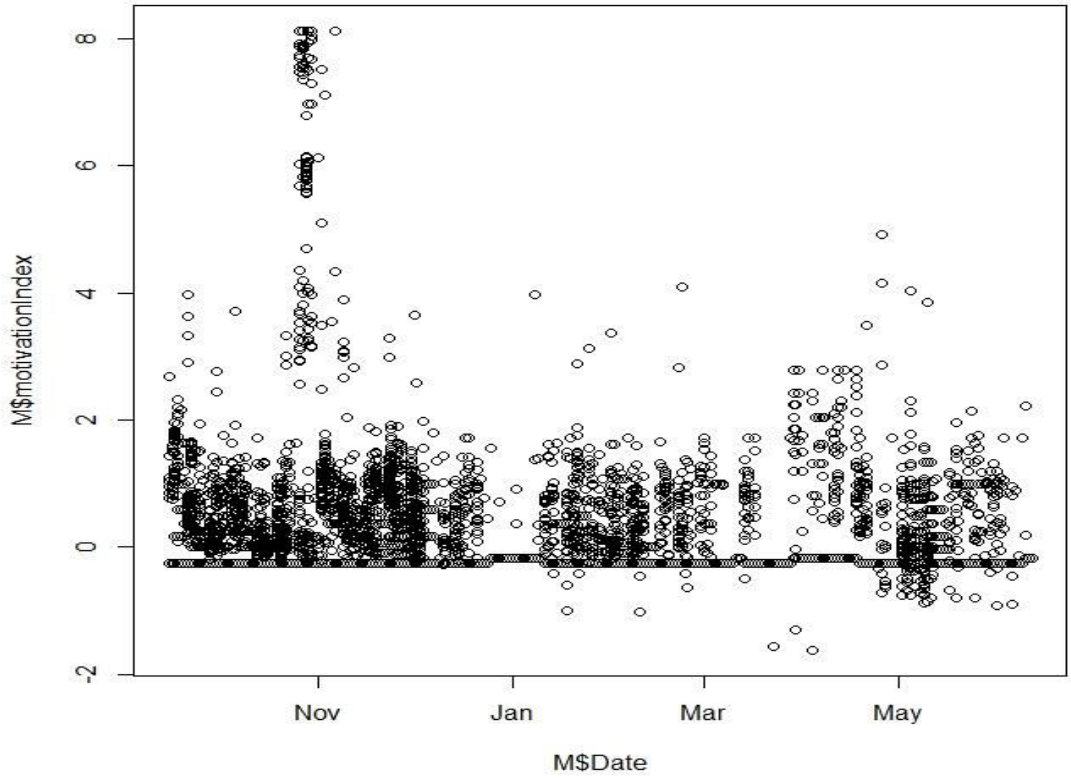


Figure 5.8 Motivation Index

CHAPTER 6

TESTING AND VALIDATION

6.1Testing

Testing is one of the crucial type of testing to check the correctness of the project or the satisfaction to the user to his requirements specification. It is useful to know if the correct project is developed as well as if the project is developed correctly.

The Unit testing for each indicator is done then and there immediately after developing them. Now it is essential to discuss about integration testing.

6.1.1 INTEGRATION TESTING

Integration is one of the most important aspects to be taken care of in the group projects. As the number of different parts of the small projects are integrated into a big project, the effectiveness and the correctness of the final project obviously depends on the factor of correct of each and every single small project and the way of integrating these module projects. As mentioned earlier, Learning Analytics is big research project going on in inLab, FIB, and my TFG is a part of the Learning analytics.

Definition: Integration testing is the phase in software testing in which individual software modules are combined and tested as a group. It occurs after unit testing and before validation testing. Even though the unit modules are tested separately, they have to be tested after integrating into a larger module.

Integration testing is normally done after the unit testing where each module is tested for its functionality and accuracy. All these verified units in unit testing are then integrated.The ingrated module as whole is tested in this type of testing.

As my part of the TFG, I have calculated seven indicators as discussed in the earlier section. Another seven indicators namely Agility Rate, Speed Delivery, Time spent, transition time, Competitive level, Engagement Level, Delivery time are developed by the inLab team mate N Balaji. All these 14 indicators along with date,userid,coursed,course name are integrated into a big table.

Now in my TFG, since all the dates for all the courses and all the students, the rate of accuracy of all the values may change. As a result of integration,on each day, starting from September 15th 2015 to 12th june 2016, all the 534 students are included among 15 courses.

Some of the indicators are calculated only for the mandatory tasks. Some days, only a few students worked and others didn't work.Like wise considering all factors , total number of NAs for the different indicators in the big table increases thus leading to the difference in the calculation of motivation index.

Here the top down approach of integration testing is done. i.e, the dataflow and high level logic is tested right from the earlier stages.The low level utilities like finding percentage of NAs being tested relatively late.

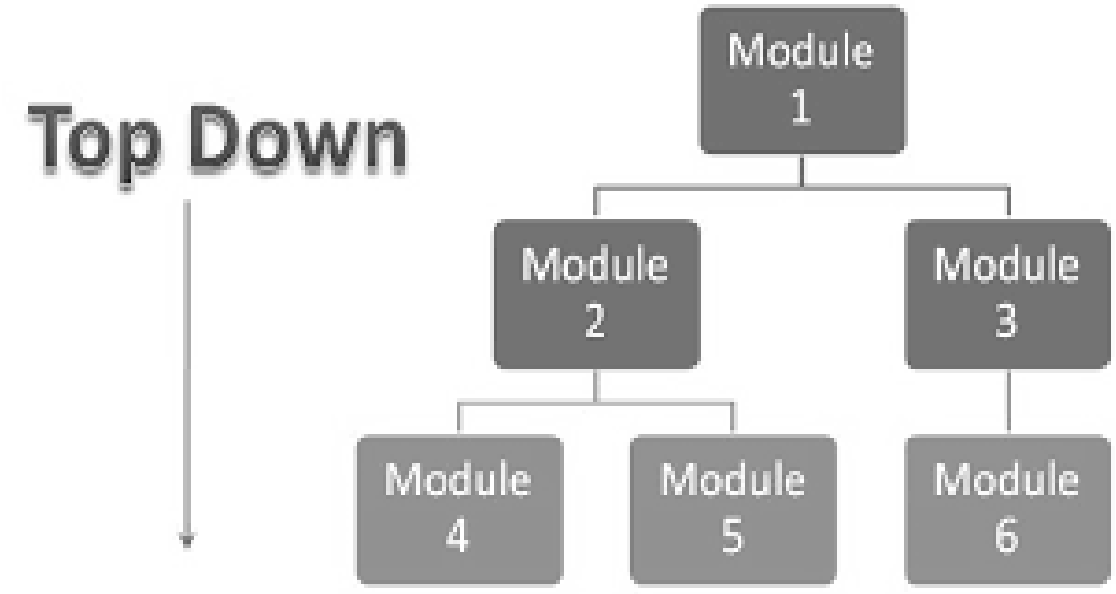


Figure 6.1 Top down approach of testing

However the final table values are seen to be robust though most of the times,the indicators are NAs.

6.2 VALIDATION:

The validation testing is the way of testing if the product meets the users' requirement or not, how correct our project is. This is the important because we can come to know whether the product developed is correct or not before delivering the product itself.

As a part of validation,the big table is taken for principle component Analysis.Though the Motivation index is calculated as the mean of all the indicators, it is necessary to check the accuracy of the logic we used and the values we obtained .Thus the correlation between the indicators is to be estimated.

6.2.1 PCA

Before going into detail,we will discuss what PCA is.Accoridng to the references taken,

Principal component analysis (**PCA**)[13] is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called **principal components**. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. The principal components are orthogonal because they are the eigenvectors of the covariance matrix, which is symmetric. PCA is sensitive to the relative scaling of the original variables.

The PCA is done in R for the big table considering all the indicators.Then the factor variable map obtained is as follows.

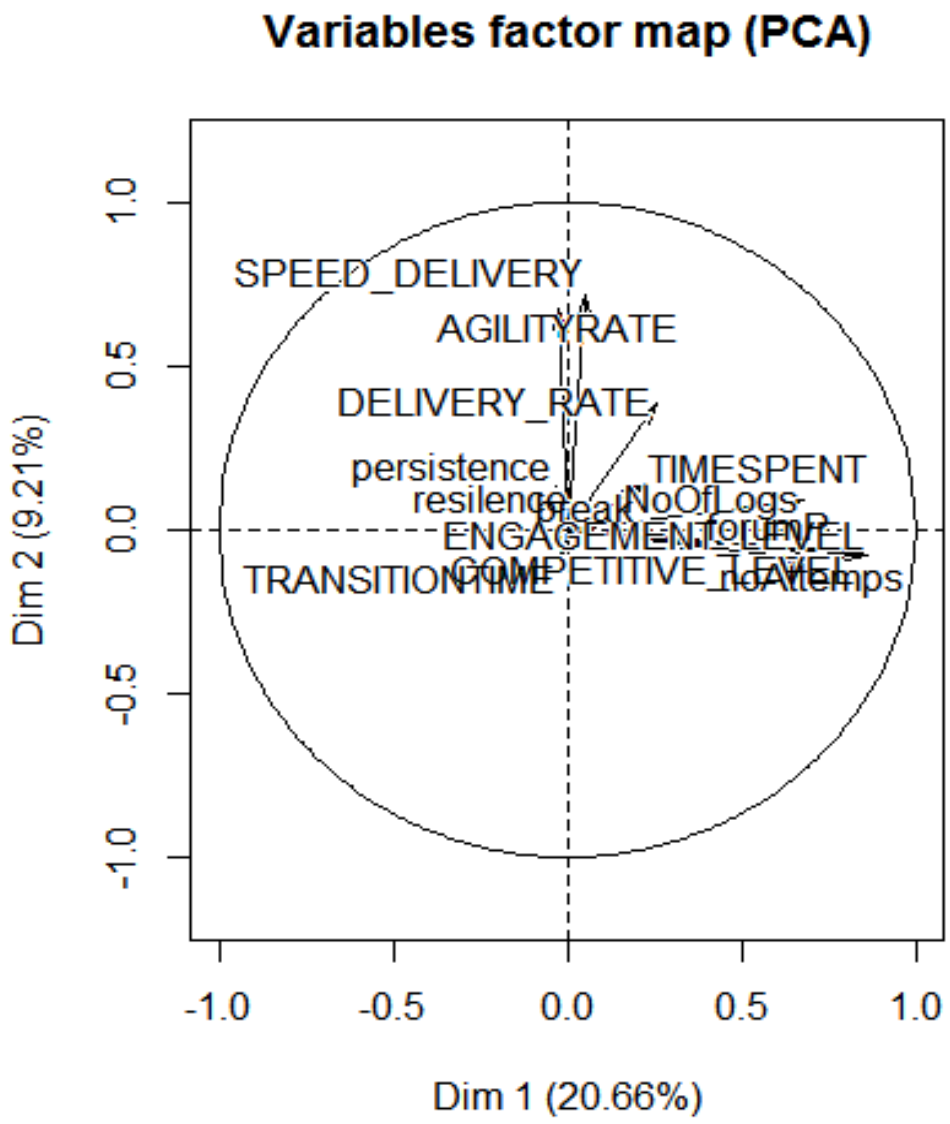


Figure 6.2 Factor Map (PCA)

```
> nd = 10
> varimax(pc1$var$coord[,1:nd])
$loadings

Loadings:
              Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7  Dim.8  Dim.9  Dim.10
AGILITYRATE                                                 0.996
TIMESPENT                   -0.976
TRANSITIONTIME                                       0.999
SPEED_DELIVERY           0.992
DELIVERY_RATE       0.110                                  -0.992
ENGAGEMENT  0.873                                           0.132
COMPETITIVE  0.837                                          0.155
NoOfLogs          0.695        -0.300                       -0.208
resilence                                    0.999
break                              0.998
noAttemps         0.717            -0.178                    0.117
persistence                        0.987
```

```
forumP          0.203                                   0.954

                Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6 Dim.7 Dim.8 Dim.9 Dim.10
SS loadings    2.521 1.003 1.047 1.017 1.002 1.002 1.002 0.998 1.000  1.011
Proportion Var 0.194 0.077 0.081 0.078 0.077 0.077 0.077 0.077 0.077  0.078
Cumulative Var 0.194 0.271 0.352 0.430 0.507 0.584 0.661 0.738 0.815  0.893

$rotmat
             [,1]       [,2]       [,3]       [,4]       [,5]       [,6]       [,7]       [,8]       [,9]      [,10]
 [1,]  0.9539970404  0.03492230 -0.1472182 -0.01216926  0.03147554  0.01156644 -0.012430991 -0.14676183 -0.01480859  0.20931182
 [2,] -0.0798037096  0.66221963 -0.1448685  0.12201438  0.01112082  0.04097942 -0.083949580 -0.35414351  0.61799777 -0.05507029
 [3,] -0.0011818416  0.22015900  0.6369014  0.35344244 -0.04707167  0.01349873 -0.068438755  0.22862021  0.01867881  0.60105384
 [4,] -0.0621919995 -0.04959205 -0.2823540  0.61901606  0.05722680  0.59163651  0.363784139 -0.09666423 -0.18699857  0.02842741
 [5,] -0.0208338832 -0.01157358 -0.1571406  0.23256980  0.72653160 -0.01944715 -0.599258801  0.16158630 -0.07684360 -0.03602180
 [6,]  0.0007351846  0.12549106  0.3316515 -0.54253377  0.49407917  0.50932541  0.255735715 -0.10234589 -0.01454080  0.01738966
 [7,] -0.0242699279  0.06600066 -0.1579569  0.08088331  0.43494266 -0.53658012  0.646388260  0.19243581  0.11133629  0.13865043
 [8,]  0.0688738594 -0.08528823 -0.2538687 -0.14250978 -0.11924179  0.31573482 -0.009605862  0.74365619  0.48110249  0.06885094
 [9,]  0.1567842942 -0.51185760  0.4340413  0.28163652  0.12388683 -0.01276152  0.070744883 -0.13897929  0.50935681 -0.38266731
[10,] -0.2220433050 -0.46883463 -0.2439165 -0.14919810  0.05187400  0.02026842 -0.091369663 -0.38141320  0.27041819  0.64736648
```

From the results it is clear that the some of the indicators are not that much correlated with each other . From the rotation matrix, we can see that some of the indicators are orthogonal to each other. Thus PCA can not be used for calculating motivation Index. Hence the mean method is used for calculation which give better results. Thus we can say that the output of the project is validated.

6.2.2 FILLING MISSING VALUES

It is noticed that there are several NAs for most of the indicator values when integrated into the big table. These Nas are tried to be filled with approximated values before performing the PCA.

Two methods of filling these NAs has been proposed by me and Balaji,inlab mate.In any case, the approximation is done using

Na.approx() Of ‘Zoo’ package in R.

In the first method,without checking any conditions,where ever there are NAs,they are filled with the nearest value corresponding to that NA position.But it seems to be inefficient .The percentage of NAs in each indicators is very large and if we do approximation,very huge number of entries will have the same approximated value.Thus it appears to be having equal value for almost everything thus leading to the incorrect results. Hence , a better method is proposed.

In the second method,for each indicator,the binary value of true or false is taken and stored in a temperorary dataframe. Now if the difference between two Consecutive TRUEs is less then 7,then the approximation is done.Else not done.

This imples that only maximum of 7 consecutive NAs can be filled with the same value which is the value of nearest. Thus it ensures that huge amount of entries donot have the same value and the results can be accurate.

FINAL OUTPUT

The following is the sample of set of motivation values obtained by doing the mean the of all the final decided list of indicators (my indicators and Balaji's indicator

```
> summary(op)
      Date              SubjectID        SubjectName          StudentID
 Min.   :1.442e+09   Min.   :  399502   Length:145248      Min.   :  151824
 1st Qu.:1.448e+09   1st Qu.:  862222   Class :character   1st Qu.: 1152372
 Median :1.454e+09   Median : 1152204   Mode  :character   Median : 8621580
 Mean   :1.454e+09   Mean   : 2778213                      Mean   :10456491
 3rd Qu.:1.460e+09   3rd Qu.: 1518184                      3rd Qu.:15181464
 Max.   :1.466e+09   Max.   :15181242                      Max.   :48615141

  AGILITYRATE      SPEED_DELIVERY    DELIVERY_RATE      resilence        persistence
 Min.   :-8.52    Min.   :-6.83    Min.   :-0.42    Min.   :-1.21    Min.   :-0.73
 1st Qu.: 0.01    1st Qu.:-0.24    1st Qu.:-0.42    1st Qu.:-0.97    1st Qu.:-0.61
 Median : 0.33    Median : 0.20    Median :-0.42    Median :-0.02    Median :-0.29
 Mean   : 0.00    Mean   : 0.00    Mean   : 0.00    Mean   : 0.00    Mean   : 0.00
 3rd Qu.: 0.48    3rd Qu.: 0.39    3rd Qu.:-0.42    3rd Qu.: 0.49    3rd Qu.: 0.22
 Max.   : 5.93    Max.   : 3.72    Max.   :22.13    Max.   : 4.61    Max.   :10.10
 NA's   :141366   NA's   :141332   NA's   :80885    NA's   :144946   NA's   :144181
     forumP          ENGplusCOMPET      MOTIVATIONINDEX
 Min.   :-0.0816   Min.   :-0.2589   Min.   :-1.92131
 1st Qu.:-0.0816   1st Qu.:-0.2589   1st Qu.:-0.25274
 Median :-0.0816   Median :-0.2589   Median :-0.17022
 Mean   : 0.0000   Mean   : 0.0000   Mean   :-0.03761
 3rd Qu.:-0.0816   3rd Qu.:-0.2589   3rd Qu.:-0.17022
 Max.   :19.1470   Max.   : 5.6505   Max.   :12.39872
 NA's   :534
>
```

CHAPTER 7

CONCLUSION

In this section the conclusion and the various results are produced. A bit of context might help here. The TFG has been concentrated on Learning Analytics. The Moodle based Agora data is analyzed to identify the patterns and perform analytics on the data. A system of indicators are identified. The various indicators that were identified are designed and developed.Also these developed indicators are visualized and Motivation is calculated.

The various design issues and the implementation aspects was provided in the previous section. The significance of the indicators can be realized only when the entire study undertaken is thoroughly understood. The indicators developed are highly instrumental in the measurement of a student belonging to a particular course in a particular day. The whole point about indicators is that the essential factors related to motivation are captured. A concrete result about the correctness of indicators cannot be produced because the results have yet to be experimentally verified .The measurement of motivation is also included here. These indicators were captured from theory and the most important aspect is realizing these indicators from the data that is available in hand. But from the results it was identified that the indicators performed well and gave satisfactory results.

Another major thing that needs to be understood well is that only a pilot version of the project is developed. Certain indicators perform well only in a competitive environment. For example the resilience level's actual behavior can be understood only when there are more than one course for a particular student.

7.1 LIMITATIONS:

Even though ,the project is aimed to be a fulfledged and very satisfactory product, there are some limitions that are to be discussed.

1.The data is stored in incremental way.That is data is taken everyday and is stored .As the data increases day by day,there is a problem of data handling.The garbage

collection possibilities are high and may be at some time,the storage leaks happen and the data may be lost.This must be taken care of.

2. The project can be transformed into a big data system for the same reasons as the previously discussed. The system performs better when a big data style is undertaken. An online processing system needs to be used.

3.The algorithms developed as a part of the system is pretty simple and the complexity may be improved better. The indicators are computed for everyday for every student and every course. This was done this way because the statistical part of the project required this schema. A better computational methodology may be designed that captures the same essential data but at a much lower cost

4.All the indicators values are not calculated for all the days.For example,persistence level and number of attempts values are not calculated for per day.But for PCA,they are modified to get values per day.Thus the results obtained may vary slightly when compared with the original data.

7.2 FUTURE WORK OF THE PROJECT:

One simple strategy is to consider the limitations ,can be solved and can be taken as the Future work. Some of the possibilities for future scope work is as follows.

1.The Front end development can be done using the output of my TFG.Thus the dashboards can be designed which can be delivered to the teachers for their reference.

2.As mentioned earlier,this is a pilot version and can be extended to all schools in Catalonia and also in Spain.

3. The big data architecture may be implemented to have easy ways of handling the increasing data.

4. The algorithmic complexity and efficiency may be improved to support scalability of the platform.

7.3 LEARNING OUTCOMES:

This section summarizes the various learning outcomes of the project. This is the most important section of my TFG because this shows the things that I have learnt by carrying out this study.

1. Design and construction of indicators for aspects for which no potential indicators exist.

2. Data handling. Extraction of required data from huge databases and proper organization of the same.

3. Using R platform to perform analytics and visualize the results in an easy and a comprehensive manner.

4. Project management that ensure smooth running of the project.

5. Measuring of the motivation index of each and every student whose data is available in the database.

6. Making sense out of massive data that give raise to interesting conclusions and observations.

This being my first internship experience, I have learnt how things work in a research development and the process flow of building a software platform. Apart from these, this project has improved me as a person by giving an exposure on good team work and project management by setting a conductive work platform.

REFERENCES

1. Page name: Learning analytics, Author: Wikipedia contributors, Publisher: *Wikipedia, The Free Encyclopedia*, Date of last revision: 1 June 2016 21:40 UTC, Date retrieved: 20 June 2016 21:44 UTC, Permanent link: https://en.wikipedia.org/w/index.php?title=Learning_analytics&oldid=723240324,Primary contributors: Revision history statistics,Page Version ID: 723240324

2. http://www.idescat.cat/economia/inec?tc=3&id=dc01&lang=en

3. "Call for Papers of the 1st International Conference on Learning Analytics & Knowledge (LAK 2011)". Retrieved February 2015

4 Michael Bostock, Vadim Ogievetsky, Jeffrey Heer. “IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)” From http://vis.stanford.edu/papers/d3. Retrieved March 2015.

5. Murray, Scott. “Interactive data visualization for the web” From http://chimera.labs.oreilly.com/books/1230000000345/index.html Retrieved May 2015.

6. Dinu, Jonathan “Data visualization and d3.js.” From https://www.udacity.com/course/ud507 Retrieved May 2015. 5 Github repository. Mbostock d3. From https://github.com/mbostock/d3/wiki/Tutorials. Retrieved April 2015.

7. Kimberly E. Arnold and Matthew D. Pistilli. “Course signals at Purdue: Using learning analytics to increase student success” In Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge, LAK '12, New York, NY, USA, 2012. ACM.

8. Helena Dierenfeld and Agathe Merceron. “Learning analytics with excel pivot tables” Moodle Research Conference, pages 115-121, 2012.

9. Wengang Liu “Using Data Mining to Dynamically Build Up Just In Time Learner Models.” Master's thesis, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, 2009.

10.Liaqat Ali, Marek Hatala, Dragan Gasevi_c, and Jelena Jovanovic.”A qualitative evaluation of evolution of a learning analytics tool. Computers & Education”, 2012.

11 https://en.wikipedia.org/w/index.php?title=Agile_software_development&oldid=725247434

12. Page name: Data analysis, Author: Wikipedia contributors, Publisher: *Wikipedia, The Free Encyclopedia*, Date of last revision: 1 June 2016 15:01 UTC, Date retrieved: 20 June 2016 17:42 UTC, Permanent link: https://en.wikipedia.org/w/index.php?title=Data_analysis&oldid=723186222 , Primary contributors: Revision history statistics, Page Version ID: 723186222.

13. https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf