**Educational Data Mining:**

**Analyze and Predict student’s academic performance**

**A research project**

**Supervisor: Guido Zuccon**

**Zhiying Zhou (9835580)**

Contents

[1. Introduction 3](#_Toc490953994)

[1.1 Background 3](#_Toc490953995)

[1.2 Motivation 4](#_Toc490953996)

[1.3 Objectives and Scope 4](#_Toc490953997)

[1.4 Methods 5](#_Toc490953998)

[1.5 Deliverables 6](#_Toc490953999)

[1.6 Significance 6](#_Toc490954000)

[2. Project Methodology 7](#_Toc490954001)

[2.1 Analysis of background 7](#_Toc490954002)

[2.2 Implementation 7](#_Toc490954003)

[2.3 Evaluation and Improvement 8](#_Toc490954004)

[3. Project Management Approach 9](#_Toc490954005)

[3.1 Project Management 9](#_Toc490954006)

[3.2 Stakeholders 9](#_Toc490954007)

[3.3 Project Controlling 10](#_Toc490954008)

[3.4 Constraints of the project 12](#_Toc490954009)

[References: 13](#_Toc490954010)

**1. Introduction**

**1.1 Background**

This project aims to analyze and predict students’ academic performance. The Mining of students’ interaction performance of education is an emerging field of application of data mining. In fact, according to (*Bunkar, Singh, Pandya, & Bunkar, 2012*), there is an increasing interest in using data mining in education. Because it may allow to identify early on struggling students with different topics in where students experienced difficulties, or on those students who are more likely to fail in the examination prior to the examination etc.

The Mining of students’ interaction performance of education is defined as the process of converting raw data from educational systems to useful information that can be used to inform design decisions and answer research questions (*Wook et al., 2009*). Data mining is an analytic approach, which can extract new insights from massive amount of data to predict hidden trends and patterns through a combination of explicit knowledge base, sophisticated analytical skills and academic domain knowledge.

There are two main research fields that related to the mining of students’ interaction performance. The first one is Learning Analytics (LA), which 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;

The second one is Educational Data Ming (EDM), it is defined as a rising order, worried with creating strategies for investigating the interesting sorts of information that originate from the educational domain, and utilizing those techniques for better comprehend of students, and their learning culture. EDM frequently stretch with the change of student models which means the student's ongoing learning, inspiration, meta perception and demeanor. EDM is using techniques of data mining and numerous research areas so as to understand about how learning is done by the student.

The difference between [Educational Data Mining](https://en.wikipedia.org/wiki/Educational_data_mining) (EDM) and Learning Analytics (LA) has been a concern of several researchers. The main differences of the two fields are mainly on different purposes. (*Brooks, 2012*) proposes that a better distinction between the EDM and LA communities is in the roots of where each community originated, with authorship at the EDM community being dominated by researchers coming from intelligent tutoring paradigms, and LA researchers being more focused on enterprise learning systems (e.g. learning content management systems).

According to the comparison, EDM is much closer to purpose of this study, because we are focused on how learning is done by the student (students’ education system) rather than enterprise learning system. Therefore, in this study we will analyze and predict the students’ academic performance based on the theory and technology of EDM.

**1.2 Motivation**

Examination plays a vital role in any student's life. The marks obtained by students in the examination affect their study and career life, especially in this competitive modern society. If the students’ performance can be predicted prior to the examination and the results shows that a student tends to fail in the examination, then extra efforts can be taken to improve his study and help him to pass the examination. It indeed benefits students a lot.

There are also many situations that the students’ performance needs to be analyzed and predicted in all levels of education. For example, to identify eligible students for scholarships or participating in placement activities, to take early action on a student who are most likely to drop out, to allocate resource with an accurate estimate of how many students will take a particular course efficiently, to help students to select courses based on other graduates’ grades and so on.

Therefore, it’s useful to analyze and predict students’ academic performance. Not only will it benefit the students, but also for schools and parents. Students can take extra efforts if there is a risk of their final exam based on the prediction of the performance.

Parents can take more measures to supervise the student’s study timely. Schools also can develop more efficient activities or remedial programs to motivate student’s studying enthusiasm based on the results of the analysis. In addition, it will also benefit all levels of education from lower level education to higher education institutions.

**1.3 Objectives and Scope**

This study will process, exploration, analyze and predict Student’s academic performance based on a XAPI Educational Mining Datasets.

The XAPI Educational Mining Datasets is download from the Kaggle website. It was collected from an e-Learning system called Kalboard 360 using Experience API Web service (XAPI).

The datasets consist of 480 student records and 16 features. The features are classified into three major categories:

1. Demographic features such as gender, grade levels, topic and nationality.
2. Academic background features such as educational stage, grade Level and section.
3. Behavioral features such as raised hand on class, opening resources, answering survey by parents, and school satisfaction.

The students are classified into three numerical intervals based on their total grade/mark:

Low-Level: interval includes values from 0 to 69,

Middle-Level: interval includes values from 70 to 89,

High-Level: interval includes values from 90-100.

We will use R Studio to analyze this datasets in R programming language, and aim to achieve the following targets:

1. Preprocess of the datasets. Clean the dataset, remove the useless columns or rows; and
2. Explore the datasets. Explore the distribution of the datasets in different features: gender, nationality, grade, topic, parental satisfaction etc. (like girls raises more hand, more discussions in high school etc.); and
3. Find underlying relationships. Like parent who are not satisfied and not answer survey, connection with study activity and performance (raising hand, discussion, absence, parental satisfaction, answering survey etc.); and
4. Build prediction model, like decision tree or regression model to predict the student’s academic performance.
5. Evaluate the predictive results of models and improve the models by comparing the accuracy.

In this study, we make prediction of students’ academic performance based on the educational datasets, like the ratio of fail or pass of students. Decision trees model will also be used to identify the students who are more likely to fail the examination. These students can be considered for proper counseling so as to improve their academic performance prior to the examination.

The datasets used in this project is just a sample, we can use the similar method to analyze many kinds of educational performance to suit all kinds of students in different grade levels, gender, nationality, and even suitable for university students.

**1.4 Methods**

This overall analysis of study will use the statistical analysis on the student’ performance datasets to explore the distribution of the data in different features.

The prediction part of this study is based on classification techniques. Classification generally refers to the mapping of data items into predefined groups and classes (*Dunham,2003*). It is also termed as supervised learning. The data classification process involves learning and classification.

In the evaluation phase of the study, the training data are analyzed by classification algorithm and during classification phase the test data are used to estimate the accuracy of the classification rules.

The organization of the study is as follows:

1. Firstly, we will summarize the previous related work about EDM.
2. Detailed methodology and project management approach will be designed in the following part.
3. The main part will be focus on the project implementation and results, which will include detailed procedure of analysis and mining of the datasets, visualization and conclusion of the educational datasets.
4. Evaluation of the project results will also be presented after the implementation part.
5. The final conclusion and future work of the study will be discussed in the last part.

**1.5 Deliverables**

There will be 2 main types of deliverables in this study:

1. A final report that contains the following main items:

1. Literature review on the related work previously; and
2. Detail steps of analysis and predictionof the dataset; and
3. Summary of the results and relationships; and
4. Discussion of this study; and
5. The final conclusion and future work.

2. One single zip archive folder which contains the following files in this study:

1. Rmarkdown file, Rmarkdown in PDf and HTML format, which will show the visualization of the analysis results of the dataset; and
2. The student’s academic performance dataset.

**1.6 Significance**

Analysis of Educational Data seeks to use these data repositories to better understand students and their study, and to develop computational approaches that combine data and theory to transform practice to benefit students. It has emerged as a research area in recent years for researchers all over the world from different and related research areas.

Overall, not only will it benefit the students, but also for schools and parents. In addition, it will also benefit all levels of education from lower level education to higher education institutions.

**2. Project Methodology**

There are 3 main questions that we want to solve in this study:

1. Is it feasible to get useful results by using EDM method to analyze the student’s academic performance?
2. How can we process and analyze the educational dataset and produce useful insights?
3. What kind of outcome we can get from the results?

To answer 3 questions above, we will apply this study according to the following 3 main steps:

**2.1 Analysis of background**

In this part, it’s neccessary to do a Literature review to understand the data mining in educational datasets, data analysis and data mining methods, what did people do in this area, what kind of methods did they use, and how did they do?

It’s also neccessary to understand the results of these articles and get a comprehensive understanding of the technologies and findings in educational data mining and how they apply these findings in real life.

**2.2 Implementation**

I this phase, I will code in R programming language to develop a program, which includes the following parts.

1. Preprocess of the datasets. Clean the dataset, remove the useless columns or rows; and
2. Explore the datasets. Explore the distribution of the datasets in different features: gender, nationality, grade, topic, parental satisfaction etc. (like girls raises more hand, more discussions in high school etc.); and
3. Find underlying relationships. Like parent who are not satisfied and not answer survey, connection with study activity and performance (raising hand, discussion, absence, parental satisfaction, answering survey etc.); and
4. Build prediction model, like decision tree or regression model to predict the student’s academic performance.

As part of my solution I will also define a number of detailed questions and hypotheses based on each part mentioned above.

Exploration Part:

|  |  |
| --- | --- |
| ID | Questions |
| 1 | How many students counts in different topics |
| 2 | How many students counts in different birth places |
| 3 | How many students counts in different genders |
| 4 | How many students counts in different grade ID |
| 5 | How many students counts in different section ID |
| 6 | What’s students counts distribution in parent satisfaction of the school |
| 7 | What’s students counts distribution in absence for over 7 days or not |

Find underlying relationships part:

|  |  |
| --- | --- |
| ID | Questions |
| 1 | What’s students counts distribution in different topics and nationalities |
| 2 | What’s students counts distribution in different topics and classes |
| 3 | What’s the connection between parents’ satisfaction level and answering the survey |
| 4 | What’s the connection between student’s grades and absence |
| 5 | What’s the distribution of raising hands counts based on parents’ satisfaction level |
| 6 | What’s the distribution of raising hands counts based on different semesters |
| 7 | What’s the distribution of raising hands counts based on different grades |
| 8 | What’s the distribution of raising hands counts based on different topics |
| 9 | What’s the distribution of raising hands counts based on different section ID |
| 10 | … |

Prediction Part:

|  |  |
| --- | --- |
| ID | Questions |
| 1 | The results reported in the confusion matrix, with respect to true and false positives. |
| 2 | If you think the classifiers you have created are acceptable in terms of their effectiveness. |
| 3 | Why you think the models have made the predictions they did: reflect on this especially with respect to the distributions of class variables. |

**2.3 Evaluation and Improvement**

Based on the analysis results, test and evaluate the results to ensure the quality of the analysis. Decison tree algorithmn will be mainly used in this part. We will compare the results of confusion matrix that generated by different models, and improve the accuracy of the predictive results by updating the settings of each model. Finally, find out the best model and summarize the conclusions from those models and generate useful insights.

**3. Project Management Approach**

**3.1 Project Management**

It is widely recognized that the waterfall model can be suited to projects where requirements and scope are fixed, the product itself is firm and stable, and the technology is clearly understood.

Waterfall approach will be used for this project. The objectives of this project are mainly on analysis and prediction of students’ academic performance, the requirements, scope and project period are fixed (for 13 weeks). It’s much easier to use and manage than agile approach, and it’s quite suitable for small projects that contain clear requirements.

Following the waterfall approach, this project will be processed in these steps:

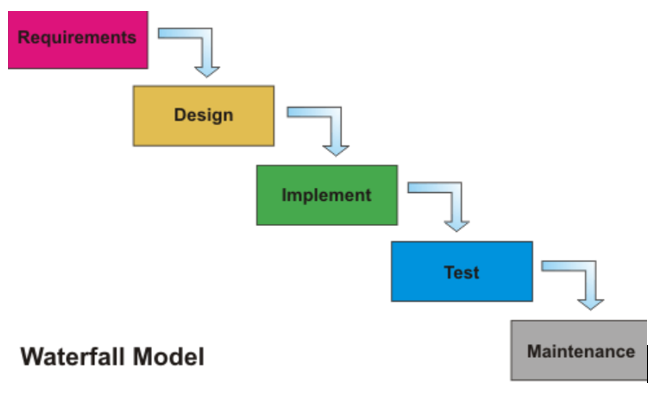
1. Definition of project requirements and make project plan.

2. Design the process of the project and make models.

3. Follow the project plan and implement the project from data preprocessing, exploration of the datasets, to prediction of the datasets.

4. Test and evaluate the results, and compare the accuracy of different prediction models.

5. Finally, make improvement and maintenance of the performance of the analysis and prediction based on the results, then summarize the conclusion.



**3.2 Stakeholders**

|  |  |
| --- | --- |
| **Stakeholder** | **Involvement** |
| Student | Develops and implements the whole project. |
| Academic Supervisor | Provides suggestions, guidelines, technical support and recommendations. |
| Project Coordinator | Gives feedbacks and advices on the presentation of the project proposal. Assesses the assignment of the project. |

**3.3 Project Controlling**

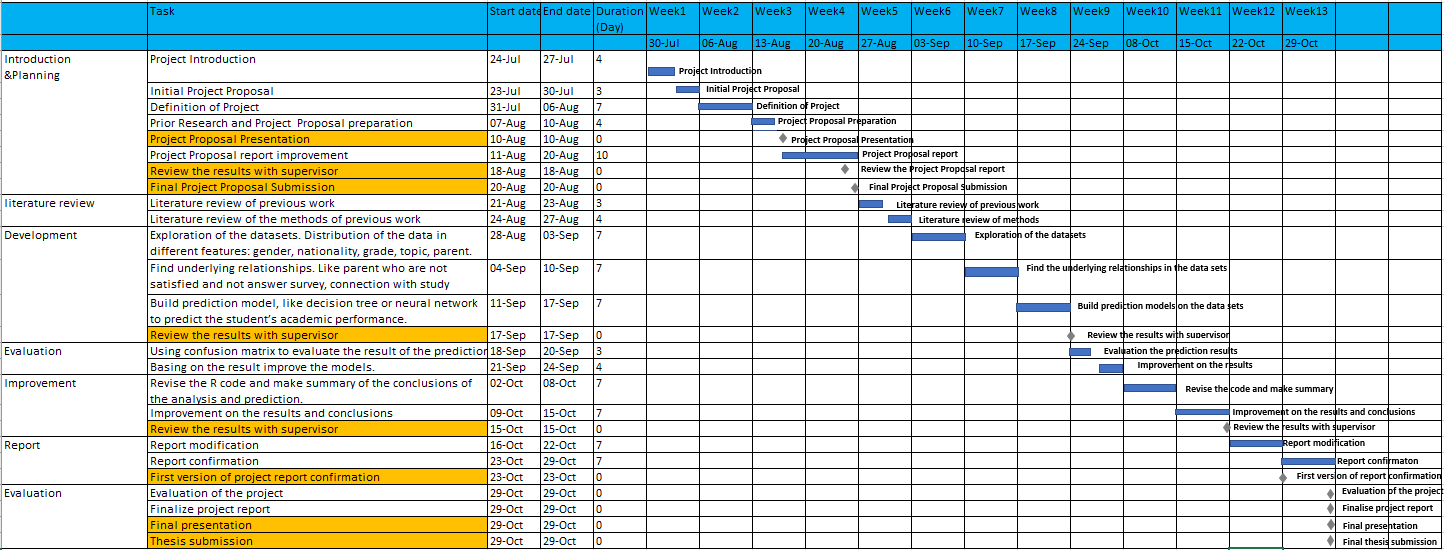
**a) Project scope**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Scope | In | Out |
| Data sets | Collect educational data sets. |  |  |
| Pre-processing the students‘ academic performance data sets. |  |  |
| Development | Preprocessing the datasets |  |  |
| Exploration of the datasets. |  |  |
| Find underlying relationships of the datasets |  |  |
| Build prediction models, like decision tree or regression model to predict the student’s academic performance. |  |  |
| Evaluate the predictive results of the datasets |  |  |
| Implement other more algorithms, like machine learning, pattern recognition. |  |  |
| Implement more different classification techniques like Naive Bayes, Bayesian Network, NB Tree and K Nearest Neighbor, and compare the efficiency of the classifiers. |  |  |

**b) Timeline tasks**

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Phases** | **Milestone** | **Due Date** | **Duration** |
| Introduction | Project Proposal presentation | 10th Aug | 3 weeks |
| Planning | Final project proposal submission | 20th Aug | 1 week |
| Literature Review | Review literatures to find out what has been researched about educational performance and what the methods they used. Explore the student’s academic performance dataset, and find out what can be done using R on these datasets. | 27th Aug | 1 week |
| Development | Based on the knowledge, developing the R code to analyse and predict the student’s academic performance. | 24th Sep | 4 weeks |
| Revise and Summarize the results | Revise the R code and make summary of the conclusions of the analysis and prediction. | 15th OCT | 2 weeks |
| Evaluation | Final presentation | 22th OCT | 2 weeks |
| Thesis submission | 29th OCT |

**c) Project Gantt Chart**

We generate the Project Gantt Chart to manage the project, which indicates the specific time schedule on each phase of the project. As a result, by following Project Gantt Chart step by step, we can deliver this project with all the required features on time.

**d) Communication plan**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Communication Type** | **Objective of Communication** | **Medium** | **Frequency** | **Audience** | **Deliverable** |
| Project progress meeting | Review the project progress and give feedbacks | Meeting | Weekly | Academic supervisor | Agenda meeting minutes |
| Phase Report | Identify the progress and problems in each phase | Reporting | Phase | Academic supervisor & Project Team | Phased Report |
| Daily notice | Receive all the mission or seek help from academic supervisor | Communication in Slack | Casual | Academic supervisor & Project Team | Slack communication records |
| Weely attendance to Lab | Communication with other student who work on the similar projects | Lab | Every week | Project Team | Lab attendance records |
| Final | Represent all outcome of this project | Communication in Slack | Final | All stakeholders | Project Report |

**e) Risk control**

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Description** | **Impact** | **Likelihood** | **Mitigation strategy** |
| Have some difficulties in programming R language for some complicated problems | Medium | High | -- Search the similar problems online and look for the solutions.  -- Discuss the problems with the other team members and solve it together.  -- Seek some help from supervisor. |
| Some of the features of this project are not delivered within the project time. | Medium | Low | -- Generate reasonable and project plan  -- Check the project progress in each phase  -- Keep pace with the timetable of the project plan |
| Lack of cooperation with team members | Low | Medium | -- Attend the weekly lab on time  -- Keep touch with team |

**3.4 Constraints of the project**

|  |  |
| --- | --- |
| ID | Description |
| 1 | Since the more records the more accurate results we can get, we just have 480 records in this student’s academic performance data sets. |
| 2 | We just use R programming language to develop this project, even though Python is much stronger. |

**References:**

C. Brooks. A Data-Assisted Approach to Supporting Instructional Interventions in Technology Enhanced Learning Environments. PhD Dissertation. University of Saskatchewan, Saskatoon, Canada 2012.

Bunkar, K., Singh, U. K., Pandya, B., & Bunkar, R. (2012, 20-22 Sept. 2012). Data mining: Prediction for performance improvement of graduate students using classification. Paper presented at the 2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN).

H. Dunham Margret, 'Data Mining-Introductory and advanced topics', published by Pearson Education, 2003.

Wook, M., Yahaya, Y. H., Wahab, N., Isa, M. R. M., Awang, N. F., & Seong, H. Y. (2009, 28-30 Dec. 2009). Predicting NDUM Student's Academic Performance Using Data Mining Techniques. Paper presented at the 2009 Second International Conference on Computer and Electrical Engineering.