**Educational Data Mining:**

**Using R markdowns to analyze and predict student’s academic performance**

**A development project**

**Supervised by Guido Zuccon**

**Zhiying Zhou (9835580)**

**1. Introduction**

**1.1 Background**

Data mining concepts and methods have been applied in various fields like marketing, medicine, real estate, customer relationship management, engineering, web mining etc. Educational Data Mining is a new emerging technique of data mining that can be applied on the data related to the field of education. There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining. (Bunkar, Singh, Pandya, & Bunkar, 2012)

In the educational area, data mining was 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” [15]. According to [16] data mining is an analytic approach that “capitalizes on the advances of technology and the extreme richness of data in higher education for improving research and decision making through uncovering hidden trends and patterns that lend them to predicative modeling using a combination of explicit knowledge base, sophisticated analytical skills and academic domain knowledge”.

EDM defined as per The Educational Data Mining community website- “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 [3]. EDM is that are of research which uses techniques of data mining and numerous research areas so as to understand about how learning is done by the student.

Educational Data Mining seeks to use these data repositories to better understand learners and learning, and to develop computational approaches that combine data and theory to transform practice to benefit learners. It has emerged as a research area in recent years for researchers all over the world from different and related research areas. (Romero & Ventura, 2010)

**1.2 Motivation**

Examination plays a vital role in any student's life. The marks obtained by the student in the examination decide his future. Therefore, it becomes essential to predict whether the student will pass or fail in the examination. If the prediction says that a student tends to fail in the examination prior to the examination then extra efforts can be taken to improve his studies and help him to pass the examination.

Using data mining techniques many kinds of knowledge can be discovered. Prediction about the students' performance is an integral part of an education system, As the overall growth of the education system is directly proportional to the success rate of the students in their examinations. Therefore, there are many situations where the performance of the students’ needs to be predicted, for example to identify eligible students for participating in placement activities, to identify students eligible for scholarships and to find the weak students so that remedial action can be taken for their betterment. The discovered knowledge can be used for prediction regarding enrollment of students in a particular course, prediction about student's performance and so on.

Overall, it’s very useful to analyze the educational data of students. Not only it will 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 activity to motivate student’s studying enthusiasm based on the results of the analysis.

**1.3 Objectives and Scope**

This study will develop R markdown codes in processing the Student’s academic performance datasets, which is download from the Kaggle website.

The dataset consists 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.

What I will do:

Using Studio R to analyze this datasets in R programming language

The targets of the study are the following:

1. Exploration of the datasets. Distribution of the data in different features: gender, nationality, grade, topic, parent…… (like girls raises more hand, more discussions in high school)
2. Find underlying relationships. Like parent who are not satisfied and not answer survey, connection with study activity and performance (raising hand, absence, content)
3. Build prediction model, like decision tree or neural network to predict the student’s academic performance.

# In this study, we make prediction about fail and pass ratio of students based on final exam. Decision trees model will also be used to identify the students who are likely to fail. These students can be considered for proper counseling so as to improve their result. The model can be easily integrated into a recommender system that can help students in their course selection, based on their and other graduated students’ grades.

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

**1.4 Methods**

This analysis part 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 [2]. 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. Firstly, we will describe the previous related work. Detailed methodology is discussed in the following part. Discussion will be presented after the methodology part. 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 knowledge:

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; 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 learners and learning, and to develop computational approaches that combine data and theory to transform practice to benefit learners. It has emerged as a research area in recent years for researchers all over the world from different and related research areas.

**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 data analysis method to analyze the student’s academic performance?
2. How can we process and analyze?
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, 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. Exploration of the datasets. Distribution of the data in different features: gender, nationality, grade, topic, parent.

2. Find underlying relationships. Like parent who are not satisfied and not answer survey, connection with study activity and performance.

3. Build prediction models, like decision tree or regression model to predict the student’s academic performance. we will analyse the overall results, create meaningful visualisations, construct classification and regression models

As part of my solution I will also define a number of detailed questions and hypotheses about the data 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. Compare the results of different models and improve them. 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**

Since the objectives of this project are mainly on research on Hadoop, and the time is fixed, I will user Waterfall approach to managing this project. Waterfall approach is easy to use and manage which is good for small projects that contain clear requirements.

The detailed process of the approach is:

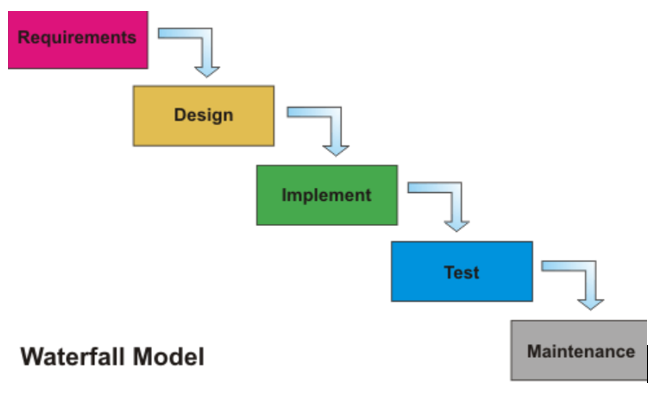
1. Planning and Requirements.

2. Design and modeling the project.

3. Follow the plan to study the Hadoop technology and big data knowledge and process the data using Hadoop.  
4. Test and evaluate the results to ensure the quality of the project.

5. Analyze, evaluate and improvement.

*Figure 4.2. Waterfall methodology*



**3.2 Stakeholders**

|  |  |
| --- | --- |
| **Stakeholder** | **Involvement** |
| Student | Develops the project. |
| Academic Supervisor | Provides guidelines and recommendations. |
| Project Coordinator | Gives feedback on presentation of the project plan.  Assesses the progress of the project. |

**3.3 Project Controlling**

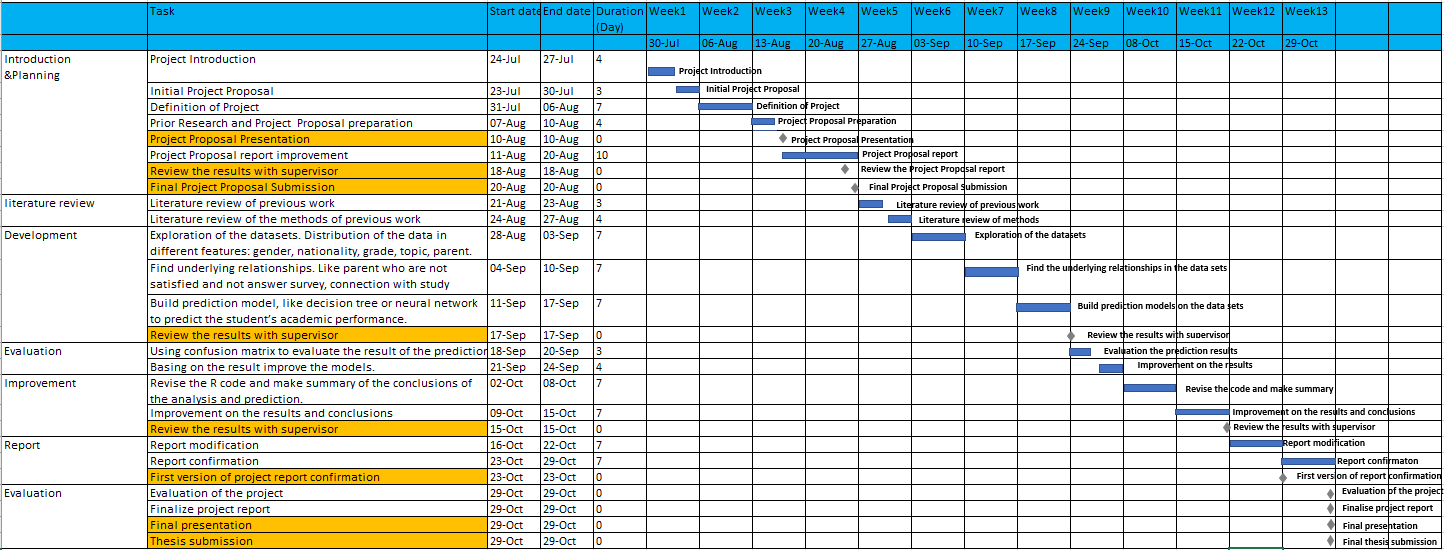
**a) Project scope**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Scope | In | Out |
| Data sets | Collect educational data sets. |  |  |
| Pre-processing the students‘ academic performance data sets. |  |  |
| Development | 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 also made Project Gantt Chart manage the project and control risk.

**d) Communication plan**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Communication Type** | **Objective of Communication** | **Medium** | **Frequency** | **Audience** | **Deliverable** |
| Project progress meeting | Review the progress and improvement for the next week | Meeting | Every week | Academic supervisor | Agenda meeting minutes |
| Phase Report | Summarize 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 |
| 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 | -- Seek the similar problems online and find out the solutions.  -- Communicate the problems with the project team timely and solve it together.  -- Seek help from teachers. |
| The time for the development of the project is relatively short what may cause that not all the features are delivered. | Medium | Low | -- Prepare project plan  -- Manage the project in different phases  -- Follow the timetable of the project plan |
| Not enough cooperation with project team | Low | Medium | -- Build communication plan  -- Routine meeting |

**g) Constraints**

|  |  |
| --- | --- |
| 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. |