Abbreviation

KDD **K**nowledge **D**iscovery in **D**atabase

… …

List of Figures

Figure 1 The accuracy between X and Y 10

List of Tables

Table 1: A comparison between X and Y 10

Table of Contents

I…. 3

II… 4

III…. 5

*(Removing this italic red paragraph below after you read it)*

*Student please use font Times New Roman, size 13, Justify Text Alignment Mode for the report content.*

*If you insert the code into the report, please keep its original format with font, size and colors.*

*Please insert only the code useful for the report in explanation or clarification, and limit lengthening the report by pasting unnecessary code. You can insert the path of the code files, and I will search it in your attached project code to check them.*

**You will improve the below sections time by time until the final submission.**

1. **INTRODUCTION**

*[In this section, students please insert the detailed description of your project with its objectives and data mining purposes. You can give the samples to make it clearer.]*

1. **DATA UNDERSTANDING**
2. **Dataset**

*[*

* *Stored in the flat file .csv or in the relational database MySQL, …*
* *Dimensional description: number of attributes, number of objects.*
* *If there is any hierarchical structure in any attribute value, that is, the value is a structure, not a single value, even you need another dataset to store its values.*
* *If there is extra storage for media files such as image, video, audio, etc.*
* *Dataset size (KB, MB,…)*
* *Dataset type*
* *…. (students insert more details in here)…*

*]*

1. **Attribute**

*[Make a table to list the attributes and their attribute type, description, example, operations, … (see the lecture 2)*

*Which attributes have flaws like empty, missing, …. Make a discussion on these attributes if they are important or have important meaning to the data mining because it affect to the result?*

*……… students fill more with your own ideas at here…..*

*]*

1. **Objects**

*[*

* *How many objects do you have in the dataset?*
* *…*

*]*

1. **DATA CLEANING**

*[*

* *Analyze and do the statistics of dataset on missing values, duplicate, redundant, irrelevant data in your dataset.*
* *How you remove them, modified them (you can add code here)*
* *Why you keep them*
* *…*

*]*

1. **DATA PREPROCESSING**

*[*

* *Reduce the objects, the attributes.*
* *Discuss on the size, attributes and the objects of the subset and*
* *How do you do it (you can add code here)*
* *…*

*]*

***--- Students please think a bit and add more contents into the below sections. The red lines are just some hints for you to understand what that sections are about. So, you need to add more your own ideas.***

***--- You can change the position of the hints between the sections if you think their new positions are more rational.***

***--- Please add more your own ideas on the report contents that help you gain higher score.***

***--- IN THIS REQUEST, you can start with 1 algorithm first, the other two can be done in the next REQUEST (request 4)***

1. **MODEL BUILDING**

In this section, our project will use decision tree algorithms to show the effectiveness of the external attributes to the student result. These attributes are:

+ Ethincity.

+ Parent level educaiton.

+ Test preparation and lunch.

Base on these attributes this algorithms will train and give the statistic values that prove the effectiveness of them to the student current result show in 4 attributes:

+ Math tier.

+ Reading tier.

+ Writing tier.

+ Total tier.

We will skip the attributes that related to the guessing result like:

+ Extra math reading time.

+ Extra reading reading time.

+ Extra writing reading time.

+ Math guess min.

+ Reading guess min.

+ Writing guess min.

+ Math guess max.

+ Reading guess max.

+ Writing guess max.

Next, we will replace the values in Ethincity, Parent level of education, Test preparation and lunch into simplicity number:

+ Ehincity: Group A-E are simplified into 0-4 respectively.

+ Parent level of education:

- We will change: Associate’s degree, College, Master’s degree and Bachelor’s degree into Higher Education.

-Then will convert the Higher Education and Highschool into 0,1 respectively.

+ Test preparation course: simplfied completed and none into 1 and 0, respectively.

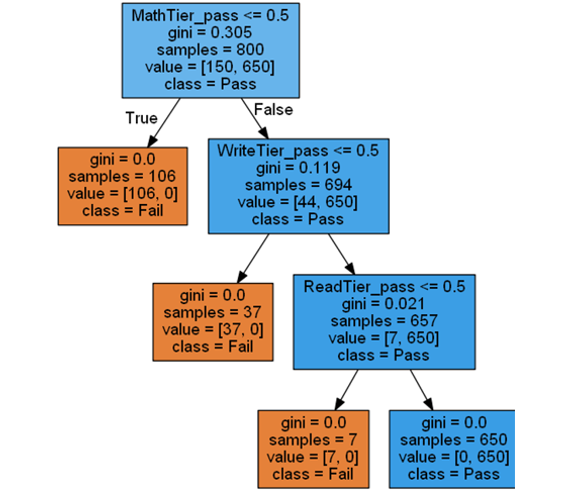
+ Lunch: simplified standard and free/reduced into 1 and 0.

+ Math tier, Reading tier and Writing tierL simplified pass and fail into 1 and 0, respectively.

On the decision tree algorithm, the time work will go into 2 parts:

+ Training: 80% total time.

+ Testing: 20% total time.

**

*[*

*In this section, students choose 3 algorithms to run on the preprocessed data.*

*V.1. Algorithm 1’s name*

*- Describe the algorithms (algorithm, its used attributes, expected predicted or described result, …)*

**-** *Analyse why the algorithm can help you in pursuing the data mining purposes of this project. What the algorithm can make and cannot make for your data mining purpose and on your dataset.*

**- …***.*

*V.1. Algorithm 2*

*- The same content instruction as above*

*V.1. Algorithm 3*

*- The same content instruction as above*

***]***

1. **EXPERIMENTS**

Physical resources:

+ 1 PC

+ CPU:

- Speed: 0.7 GHz

- Processes: 4

- Threads: 40

+ 0.1-0.2 GB RAM

* *Experiements description: training and testing data description, purpose, …*

We

* *Visual Data (Charts or Tables – coding to build up charts using data visualization can be referred to the coding textbook I shared with your last week)*
  + *Accuracy visual data of 3 algorithms. Analyse why it is good or bad (what factor such as dataset, params, ….)*
  + *Other visual data analyse the data mining purpose (e.g., heart disease is effected by attributes X1, X2, …. More specificially, when X1, X2, … increases, the heart disease of these ages … are affected seriousely, etc…*
  + *Compare three of them using charts or tables.*

1. **CONCLUSION AND FUTURE WORKS**

- Pros:

+Easy to construct

+Quick run and response

+It can be used with a wide range of objects.

+Given a general assessment of the student classification, finding and giving time to low-scoring students, and guessing the point using simple math and simple algorithms,This is a common misunderstanding.

+The data can be displayed easily by bar chart and pie chart

- Cons:

+The data depends mostly on mathematic.

+Not diversity (the data focuses solely on classifying, guessing points, and assisting the weak student, and has yet to release some strategies to assist the student in improving as well as fostering the good and excellent students).

-Future works:

+Adding more requests to increase the diversity of data solutions

+Using the code to determine other data like VGU or some other schools, colleges, universities, etc.

+Given some specific direction for students based on students, create algorithms to foster the good and excellent students.

**DUTY ROSTER**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Task** | **In Charge** | **Start** | **End** | **State** | **Note** |
| 1 | Design Class A, B, C | Nguyen Van A  Le Thi B | 2-Dec-18 | 1-Jan-19 | Done |  |
| 2 | Code Function 1 | Nguyen Van A | 01-Jan-19 | ~~17-Jan-19~~  20-Jan-19 | Delayed | Overlapped time with the other course project |
| 3 | Report Section II | Le Thi B | 02-Jan-19 |  | In progress |  |
| … | … | … | … | … | … | … |
| n | … | … | … | … | … | … |

**REFERENCE**

1. Tutorial Page, Oracle https://...
2. …

*[Students, please put here whatever sources you referred or used in the project]*