**VIETNAMESE – GERMAN UNIVERSITY**

FACULTY OF ENGINEERING

COMPUTER SCIENCE DEPARTMENT

PROJECT REPORT

LOCATE EFFICIENTLY

FACTORS AFFECTING EXAM RESULT THINGY

&

ESTIMATE FUTURE EXAM RESULT

THINGY

**(L.E.F.A.E.R.T. & E.F.E.R.T.)**

***Module 25: Data Mining***

1. Le Duc Nghia - 14706

2. Nguyen Ba Le -14261

Abbreviation

LEFAERT **L**ocate **E**fficiently **F**actors **A**ffecting **E**xam **R**esult **T**hingy

EFERT **E**stimate **F**uture **E**xam **R**esult **T**hingy

List of Figures

List of Tables

Table 1: List of attributes 5

Table 2: List of new attributes 11

Table of Contents

I. Introduction................................................................................................................5

II. Data Understanding..................................................................................................5

III. Data Cleaning ........................................................................................................ 6

IV.Data Preprocessing...................................................................................................7

V. Modeling...............................................................................................................

VI. Experiments ...............................................................................................................

VII. Conclusion and Future Works..................................................................................

1. **INTRODUCTION**

When a student take an exam, there are many factors that can contribute to how the students perform in their exam. It would be extremely hard to manually find out what factor is affecting the student the most. Therefore, we aim to locate which is the most influential factor on the exam performance of students with this project.

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

* The dataset was downloaded from: <https://www.kaggle.com/spscientist/students-performance-in-exams>
* Dataset file is saved as StudentPerformance.csv
* There are 8 columns and 1001 rows (including the title row).
* The size of dataset is 71 KB.
* The dataset is a combination of strings and numbers.

1. **Attribute:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Name*** | ***Type*** | ***Description*** | ***Example*** | ***Operation*** |
| gender | Nominal | Gender is the range of characteristics pertaining to, and differentiating between male and female. | Male | mode, entropy, contingency correlation, X2 test |
| race/ethnicity | Nominal | Differentiate which group of people each student belong to. | Group A | mode, entropy, contingency correlation, X2 test |
| parental level of education | Nominal | The level of education that the parents received. | Bachelor's degree | mode, entropy, contingency correlation, X2 test |
| lunch | Nominal | The type of lunch that the student eat. | Standard | mode, entropy, contingency correlation, X2 test |
| test preparation course | Nominal | Special course to prepare student for their exam. | None | mode, entropy, contingency correlation, X2 test. |
| math score | Interval | The score that the student get in their math exam between 0 and 100 score. | 72 | mean, standard deviation, Pearson's correlation, t and F tests. |
| reading score | Interval | The score that the student get in their reading exam between 0 and 100 score. | 74 | mean, standard deviation, Pearson's correlation, t and F tests. |
| writing score | Interval | The score that the student get in their writing exam between 0 and 100 score. | 100 | mean, standard deviation, Pearson's correlation, t and F tests. |

***Table 1: List of attributes***

1. **Objects**

We have 1000 objects in our dataset.

1. **DATA CLEANING**

*[*

*Analyze and do the statistics of dataset on missing values, duplicate, redundant, irrelevant data in your dataset.*

There is only one redundant attribute in the dataset is the gender. Because this is the project writing about the education, so the gender attribute is too sensitive in many ways. Because of that we decided that this attribute will not be used to analyze and do the statisics.

* *How you remove them, modified them (you can add code here)*

We removed the gender attribute directly on the csv file

* *Why you keep them*

We decided not to keep anything that not relative to our process ( Might not be needed , you can delete or adjust this line if you need.)

* *…*

*]*

1. **DATA PREPROCESSING**

*[*

* *Reduce the objects, the attributes.*

We just only removed the gender attribute.

* *Discuss on the size, attributes and the objects of the subset and*

-We adding some attributes and save it to the file name StudentResults\_and\_StudentResolves.csv.

data.to\_csv('StudentResults\_and\_StudentResolves.csv')

-The data size is 109KB.

+New attribute:

* math/reading/writing tier (3 attributes): ,
* total tier
* extra study math/reading/writing percentage (3 attributes)
* math/reading/writing min (3 attributes), math/reading/writing max (3 attributes).
* *How do you do it (you can add code here)*

-We use python language and import the pandas,numpy libraries to adding and set the conditions to these attributes.

-Attributes:

+ Math tier:

* Definition: Math evaluation based on the math score.
* Explaination: If the attribute name ‘math score’ >= 50 then it will written pass, else will be fail.
* Code:

conditions1 = [(data['math score']>=50 ),(data['math score']<50)]  
values1 = ['pass','fail']  
data['Math tier'] = np.select(conditions1, values1)

+ Reading/Writing tier: the definition, explaination and code is similarity to the “math pass” attribute.

+ Total tier:

* Definition: Total evaluation based on 3 subject math, reading and writing.
* Explaination: This attribute will be written pass if all math/reading/writing pass are written pass, else will be fail.
* Code:

conditions = [((data['math score']>=50) & (data['reading score']>=50) & (data['writing score']>=50)),((data['math score']<50) | (data['reading score']<50) | (data['writing score']<50))]  
values = ['pass','fail']  
data['Total tier'] = np.select(conditions,values)

+ Extra study math percentage:

* Definition: Calculate the time for extra study in percentage base on the math result
* Explaination: The percentage will be calculated following the function: D = 100 – (data[‘writing score’]-49)\*100/22 (If D > 100 then D = 100, If D < 0 then D = 0)
* Code:

conditions\_math= [(data['math score']<50),  
 (data['math score']>70),  
 (data['math score']>=50) & (data['math score']<=70)]  
values\_math=[100,0,round(100-(data['math score']-49)\*(100/22))]  
data['Extra study math percentage'] = np.select(conditions\_math,values\_math)

+ Extra study reading/writing percentage: the definition, explaination and code is similarity to the “Extra study math percentage” attribute.

+ Math guess min:

* Definition: Prediction the minimum result of the math in next course
* Explaination: ‘Math guess min’ = ‘math score’ – 10 (if ‘Math guess min’ < 0 then ‘Math guess min’ = 0)
* Code:

math\_guess\_conditions = [data['math score']<=10  
 ,(data['math score']>10) & (data['math score']<50)  
 ,(data['math score']>=50) & (data['math score']<70)  
 ,(data['math score']>=70) & (data['math score']<80)  
 ,(data['math score']>=80) & (data['math score']<=100)]

math\_guess\_values\_min = [0,  
 data['math score']-10,  
 data['math score']-10,  
 data['math score']-10,  
 data['math score']-10]

data['math\_guess\_min'] = np.select(math\_guess\_conditions,math\_guess\_values\_min)

+ Reading/writing guess min: the definition, explaination and code is similarity to the “Math guess min” attribute.

+ Math guess max:

* Definition: Prediction the maximum result of the math in next course
* Explaination: ‘Math guess max ’ = ‘math score’ + 20 + 5\* ‘extra math study percentage’atttibute in % unit (if ‘Math guess max’ > 100 then ‘Math guess max’ = 100)
* Code:

math\_guess\_conditions = [data['math score']<=10  
 ,(data['math score']>10) & (data['math score']<50)  
 ,(data['math score']>=50) & (data['math score']<70)  
 ,(data['math score']>=70) & (data['math score']<80)  
 ,(data['math score']>=80) & (data['math score']<=100)]

math\_guess\_values\_max = [data['math score']+25,  
 data['math score']+25,  
 round(data['math score']+20+0.05\*data['Extra study math percentage']),  
 data['math score']+20,100]

data['math\_guess\_max'] = np.select(math\_guess\_conditions,math\_guess\_values\_max)

+ Reading/writing guess max: the definition, explaination and code is similarity to the “Math guess max” attribute.

New attribute in table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Name*** | ***Type*** | ***Description*** | ***Example*** | ***Operation*** |
| Math tier | Nominal | Pass or Fail base on the math result | Pass | mode, entropy, contingency correlation, X2 test |
| Reading tier | Nominal | Pass or Fail base on the reading result | Pass | mode, entropy, contingency correlation, X2 test |
| Writing tier | Nominal | Pass or Fail base on the math result | Fail | mode, entropy, contingency correlation, X2 test |
| Extra study math percentage | Interval | The extra time study of math between 0-100 in percentage unit | 56 | mean, standard deviation, Pearson's correlation, t and F tests |
| Extra study reading percentage | Interval | The extra time study of reading between 0-100 in percentage unit | 45 | mean, standard deviation, Pearson's correlation, t and F tests |
| Extra study writing percentage | Interval | The extra time study of writing between 0-100 in percentage unit | 80 | mean, standard deviation, Pearson's correlation, t and F tests |
| math guess min | Interval | The minimum score prediction in math from 0 to 90 (because base on current score – 10) | 36 | mean, standard deviation, Pearson's correlation, t and F tests |
| reading guess min | Interval | The minimum score prediction in reading from 0 to 90 (because base on current score – 10) | 64 | mean, standard deviation, Pearson's correlation, t and F tests |
| writing guess min | Interval | The minimum score prediction in writing from 0 to 90 (because base on current score – 10) | 82 | mean, standard deviation, Pearson's correlation, t and F tests |
| math guess max | Interval | The maximum score prediction in math from 25 to 100 (because base on current score + 25 if student current score < 50) | 42 | mean, standard deviation, Pearson's correlation, t and F tests |
| reading guess max | Interval | The maximum score prediction in reading from 25 to 100 (because base on current score + 25 if student current score < 50) | 85 | mean, standard deviation, Pearson's correlation, t and F tests |
| writing guess max | Interval | The maximum score prediction in writing from 25 to 100 (because base on current score + 25 if student current score < 50) | 92 | mean, standard deviation, Pearson's correlation, t and F tests |

***Table 2: List of new attributes***

1. **MODEL BUILDING**
2. **EXPERIMENTS**
3. **CONCLUSION AND FUTURE WORKS**

*[*

* *Conclude what you have done. The works’ pros and cons*

**DUTY ROSTER**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Task** | **In Charge** | **Start** | **End** | **State** | **Note** |
| 1 | Select topic | Everyone | 17/11/2021 | 20/11/2021 | Done | Had to change topic after reviewing |
| 2 | Find dataset | Nguyễn Bá Lễ | 18/11/2021 | 18/11/2021 | Done | Find something related to student exam performance |
| 3 | Review dataset | Everyone | 18/11/2021 | 20/11/2021 | Done |  |
| 4 | Write report 1 | Everyone | 26/11/2021 | 26/11/2021 | Done |  |

**REFERENCE**

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

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

1. 14706
2. 14261

Lecturer: Ngoc H. Tran, Ph.D.

Binh Duong, WS2021