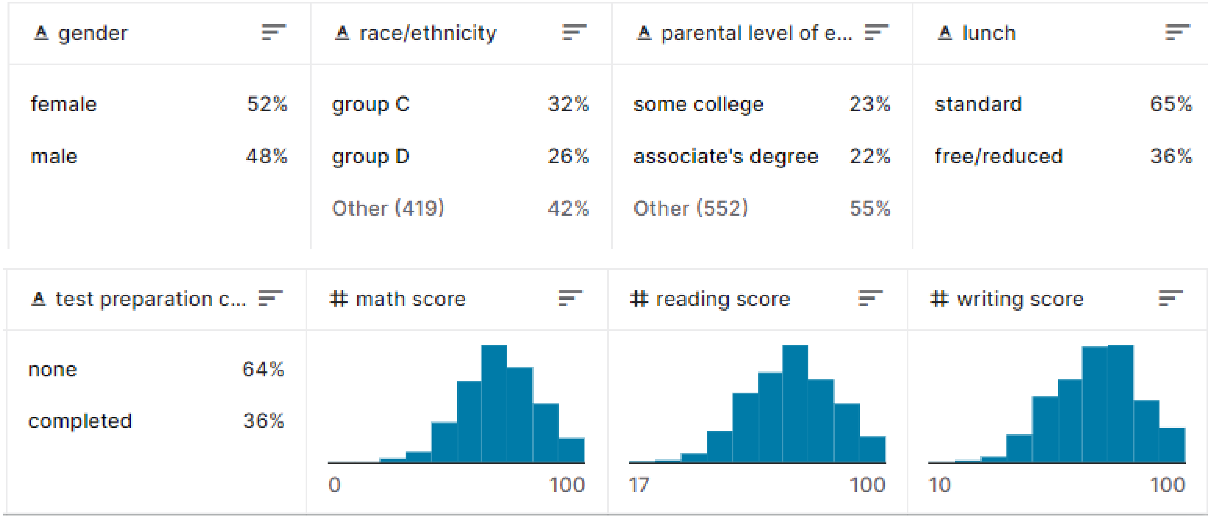
# 最後一組 Midterm Report

M11215032 葉品和  M11215052 陳奕帆  M11215066 鄭宜珊

## 1. Dataset

We use the dataset “**Students Performance in Exams**” from Kaggle. There are a total of 8 attributes. The first 5 attributes are quasi-identifier, and the last 3 attributes are sensitive attributes.



| gender | | race/ethnicity | | parental level of e... | | lunch | |
|---|---|---|---|---|---|---|---|
| female | 52% | group C | 32% | some college | 23% | standard | 65% |
| male | 48% | group D | 26% | associate's degree | 22% | free/reduced | 36% |
| | | Other (419) | 42% | Other (552) | 55% | | |

| test preparation c... | | math score | reading score | writing score |
|---|---|---|---|---|
| none | 64% | | | |
| completed | 36% | | | |

## 2. Anonymizing Algorithm

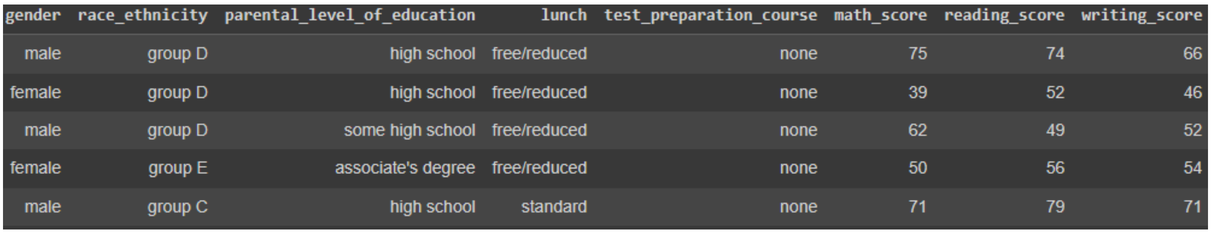
In the "race/ethnicity" attribute, there are a total of 5 categories, namely "group A" to "group E." Among them, "group C" and "group D" appear the most frequent, accounting for 32% and 26%, respectively. As a result, we have modified the remaining 3 categories to "other" to achieve anonymity.

There are 6 different categories in attribute “ parental level of education”. “some college” appears the most frequently(23%), “associate's degree” is the second, it has about 22%. We combine the rest of the categories into two combinations. “some high school” and “high school” to “before college”. “bachelor's degree” and “master's degree” to “bachelor's and master's degree”.

We use the mondrian algorithm to produce the data based on different k, l and t. Initially, the Mondrian algorithm assigns all data records to one group. It then iteratively partitions the data on the attribute values to split each group into two subsets. The split subsets are added back to the list for further partitioning if the partitioning meets the user-defined “k”, “l”, “t” constraints. Otherwise, another attribute is selected for partitioning the group.

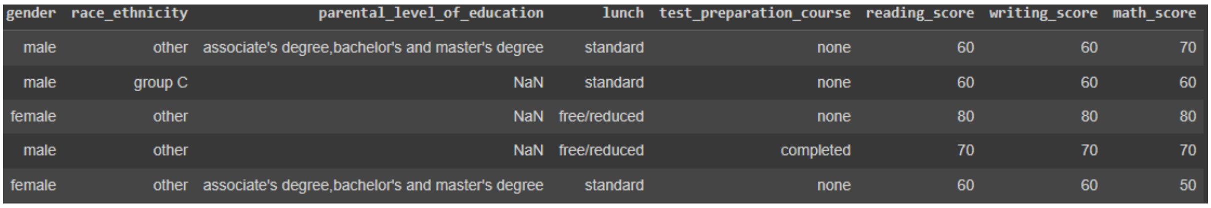
If no attribute yields a valid partition that meets the constraints for a group, that group is added to the output dataset.

The iteration continues until no more groups can be further partitioned. Finally, generalization is applied on the attribute values in the output groups to meet the k-anonymity, l-diversity and t-closeness requirements defined by the user.



| gender | race_ethnicity | parental_level_of_education | lunch | test_preparation_course | math_score | reading_score | writing_score |
|---|---|---|---|---|---|---|---|
| male | group D | high school | free/reduced | none | 75 | 74 | 66 |
| female | group D | high school | free/reduced | none | 39 | 52 | 46 |
| male | group D | some high school | free/reduced | none | 62 | 49 | 52 |
| female | group E | associate's degree | free/reduced | none | 50 | 56 | 54 |
| male | group C | high school | standard | none | 71 | 79 | 71 |

(a) original data



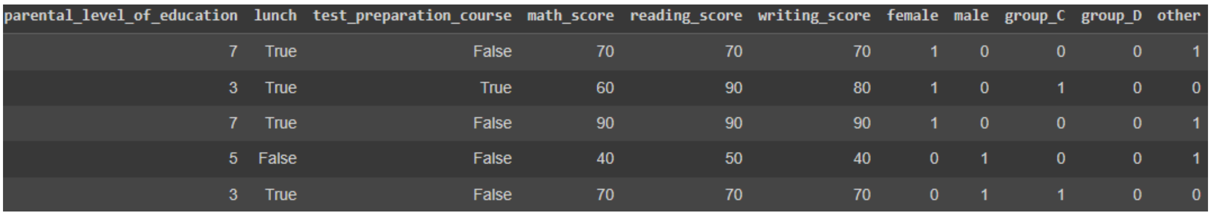
| gender | race_ethnicity | parental_level_of_education | lunch | test_preparation_course | reading_score | writing_score | math_score |
|---|---|---|---|---|---|---|---|
| male | other | associate's degree,bachelor's and master's degree | standard | none | 60 | 60 | 70 |
| male | group C | NaN | standard | none | 60 | 60 | 60 |
| female | other | NaN | free/reduced | none | 80 | 80 | 80 |
| male | other | NaN | free/reduced | completed | 70 | 70 | 70 |
| female | other | associate's degree,bachelor's and master's degree | standard | none | 60 | 60 | 50 |

(b) anonymized data (k=10, l=10, t=0.3)

## 3. Machine Learning

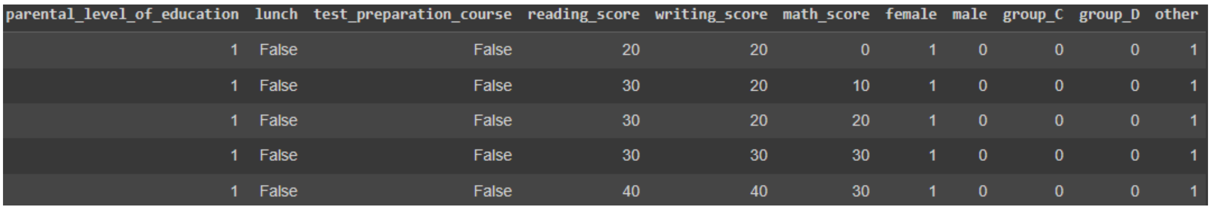
In the preprocessing phase, we applied one-hot encoding to map the categorical attributes "race/ethnicity" and "gender". The attribute "parental level of education" was mapped to numeric values from 1 to 7, corresponding to the 7 levels of education. The "lunch" and ”test preparation course” attributes were encoded into boolean values.

We've used the SVM algorithm to predict the attribute "writing score". Then, we used 3 metrics, including "Mean Square Error(MSE)", "Root Mean Square Error(RMSE)", and "Explained Variance Score", “R-Squared” to evaluate our model. The figures below show the results of the original data and the anonymized data after preprocessing.



| parental_level_of_education | lunch | test_preparation_course | math_score | reading_score | writing_score | female | male | group_C | group_D | other |
|---|---|---|---|---|---|---|---|---|---|---|
| 7 | True | False | 70 | 70 | 70 | 1 | 0 | 0 | 0 | 1 |
| 3 | True | True | 60 | 90 | 80 | 1 | 0 | 1 | 0 | 0 |
| 7 | True | False | 90 | 90 | 90 | 1 | 0 | 0 | 0 | 1 |
| 5 | False | False | 40 | 50 | 40 | 0 | 1 | 0 | 0 | 1 |
| 3 | True | False | 70 | 70 | 70 | 0 | 1 | 1 | 0 | 0 |

(c) original data (preprocessed)



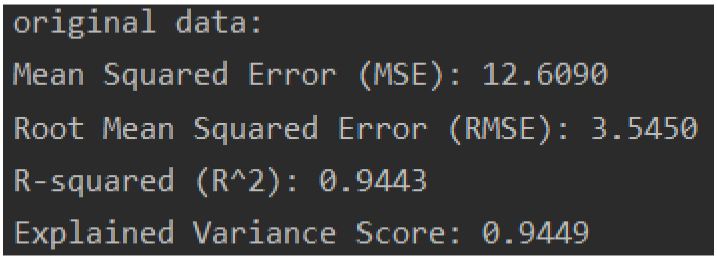
| parental_level_of_education | lunch | test_preparation_course | reading_score | writing_score | math_score | female | male | group_C | group_D | other |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | False | False | 20 | 20 | 0 | 1 | 0 | 0 | 0 | 1 |
| 1 | False | False | 30 | 20 | 10 | 1 | 0 | 0 | 0 | 1 |
| 1 | False | False | 30 | 20 | 20 | 1 | 0 | 0 | 0 | 1 |
| 1 | False | False | 30 | 30 | 30 | 1 | 0 | 0 | 0 | 1 |
| 1 | False | False | 40 | 40 | 30 | 1 | 0 | 0 | 0 | 1 |

(d) anonymized data (preprocessed)

## 4. Results

(1) Predict attribute : *”writing score”*

(2) train with original data:

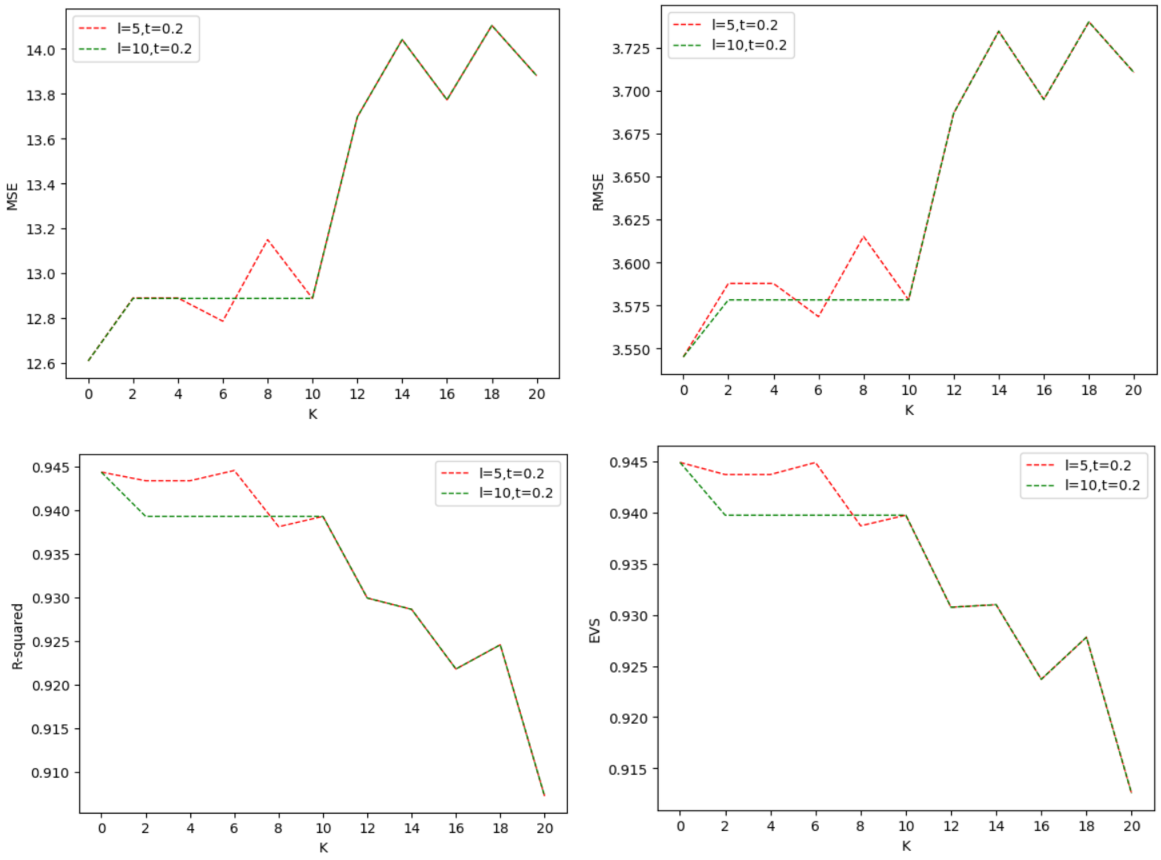


```
original data:
Mean Squared Error (MSE): 12.6090
Root Mean Squared Error (RMSE): 3.5450
R-squared (R^2): 0.9443
Explained Variance Score: 0.9449
```

(3) train with anonymized data (different k, l, t):

| ***l \ t*** | 0.2 | 0.3 |
|---|---|---|
| 5 | red | blue |
| 10 | green | black |

(red line coincides with blue line and black line)



(e) result

### 5. Responsibilities

| | |
|---|---|
| 葉品和 | data collection, Mondrian algorithm(l-diversity), integrating, report, result plot |
| 陳奕帆 | Mondrian algorithm(k-anonymity), ML(preprocessing), report, video |
| 鄭宜珊 | Mondrian algorithm(t-closeness), ML(result measurement), report |

### 6. Video

https://youtu.be/PWxChaRz_0o