FPT Education

FPT UNIVERSITY

A University Student Dropout Detector based on Academic Data
A case study at FPT University

Ngo Quang Hai; Nguyen Hoang Giang; Trinh Nhat Minh

Supervisor: Mr. Ngo Tung Son

Outline

1. Introduction
2. Literature review
3. Methodology
4. Experimental and result
5. Conclusion

1. Introduction

1. Introduction

1.1 Definition

Student dropout is when students leave their university program before completing it without obtaining a degree.

Dropout can occur for various reasons, including financial constraints, academic difficulties, lack of motivation…

For universities, this issue can lead to revenue loss and reputational damage.

=> Detecting students who are at risk of leaving school prematurely at an early stage is crucial in mitigating the issue and directing appropriate interventions

1. Introduction

1.2 Research problem

- **Class imbalance**: number of dropout students is a minority
- **Multifactional**: Many factors contribute to student dropout. Therefore, hard to determine a particular time that student dropout
- **Limited data availability**: reduce availability to develop a predictive model
- **Lack of public dataset**: researcher unable to perform their work

1. Introduction

1.3 Research question

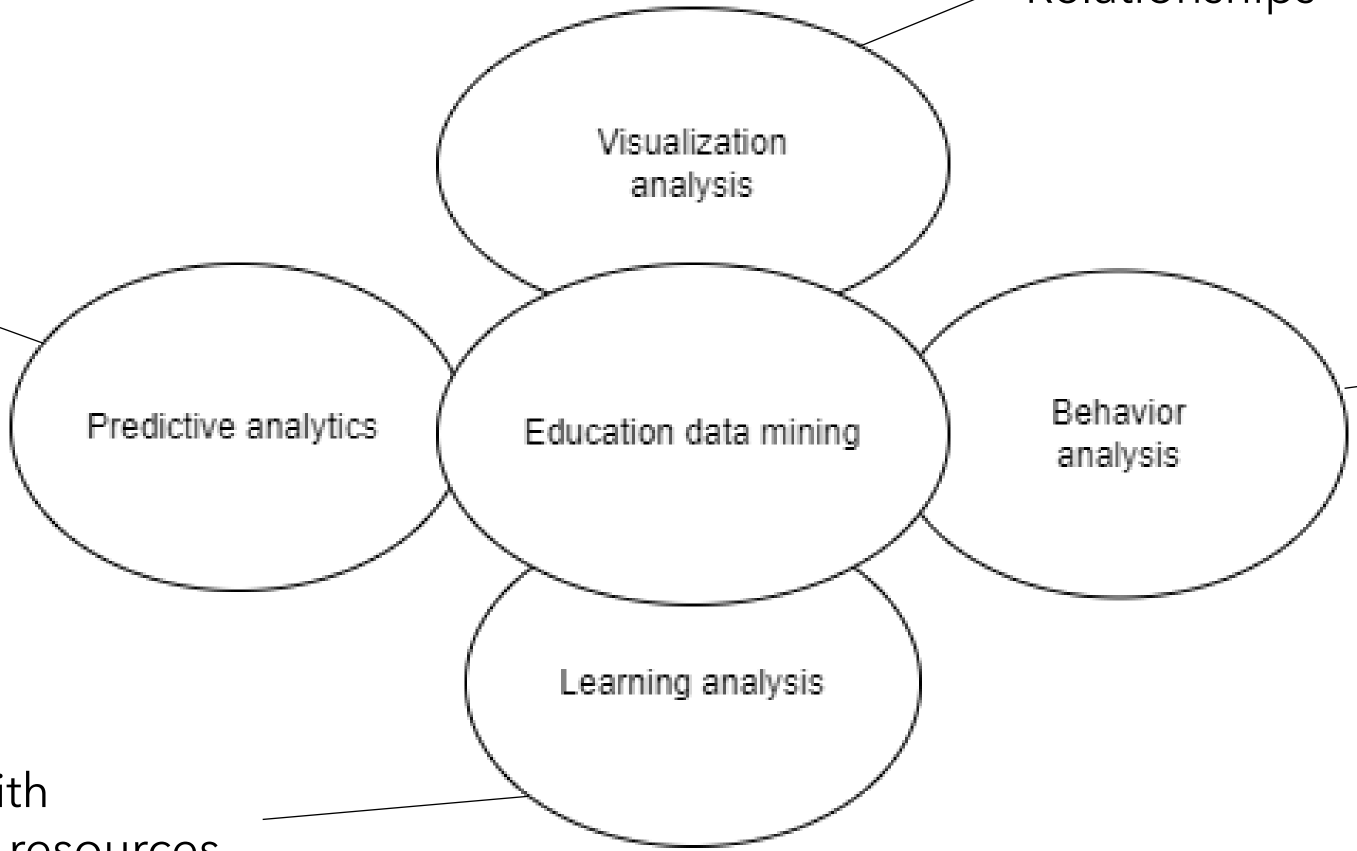
- **Question 1**: How does academic performance influence student dropout?
- **Question 2**: If the defined attributes in structured data are not sufficient, are there any hidden features that can help distinguish the characteristics of dropout students?

1.4 Research objective

- Implement data cleaning
- Use suitable feature selection algorithm
- Using sampling techniques, modify the loss function

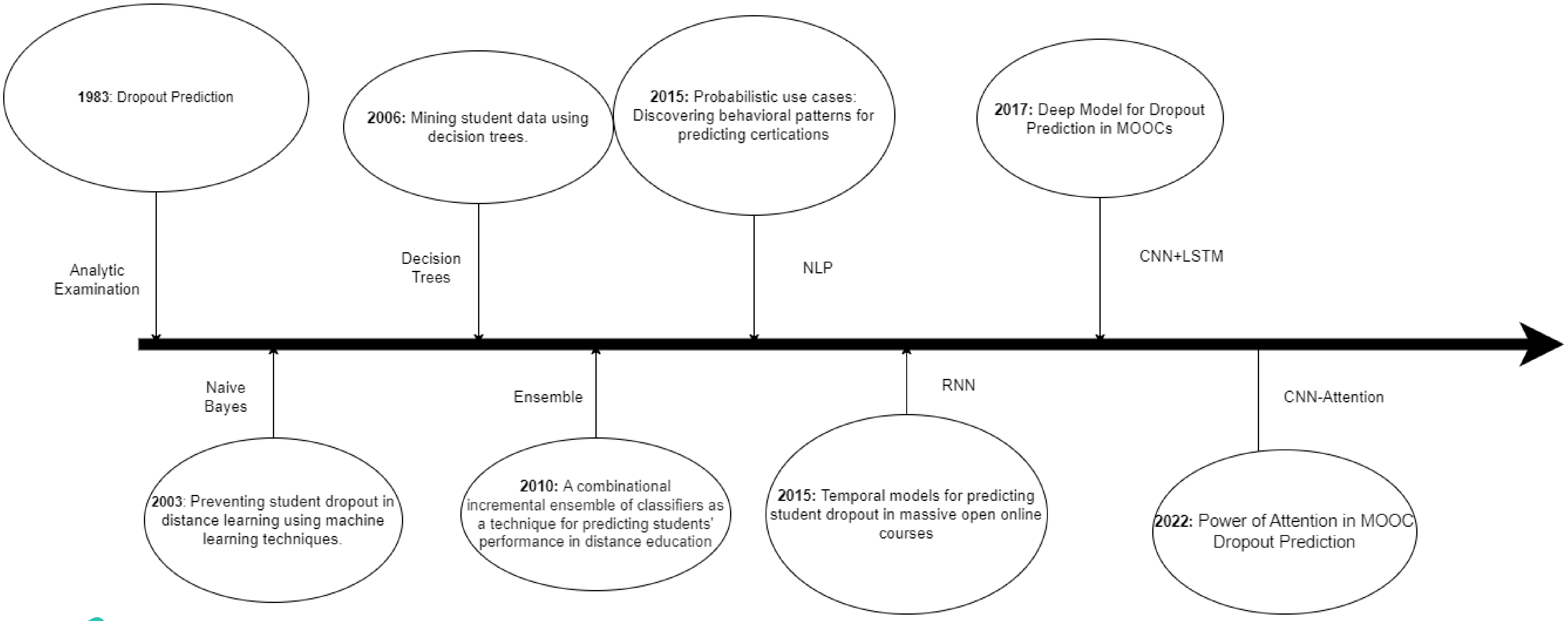
2. Literature review

2. Literature review



Data mining scheme in higher education

2. Literature review

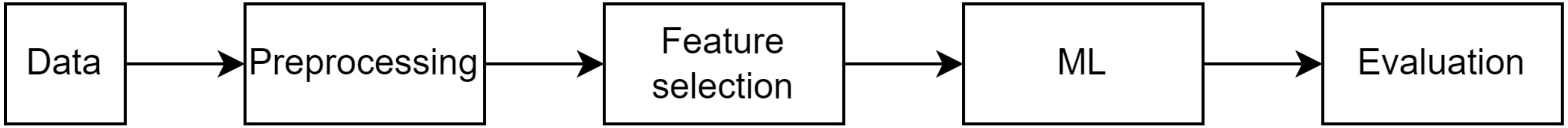


2. Literature review

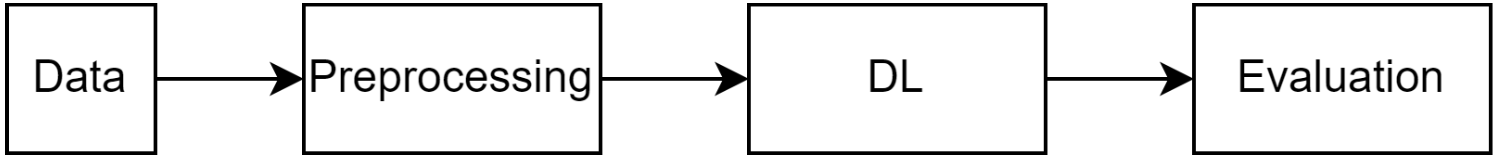
- **Traditional supervised algorithm**: Compare between methods; Tree-based give promising result; Combine with Feature Selection to increase performance
- **Basic Deep Learning**: Because the traditional algorithm cannot learn hidden features, the Deep learning method can help find hints about relations between features and the latent space
- **Sequence-based model**: Since student data is constantly updated, we can consider this a time series problem
- **Unsupervised algorithm**: Clustering data to find relationship

2. Literature review

Method 1 – Machine Learning Approach



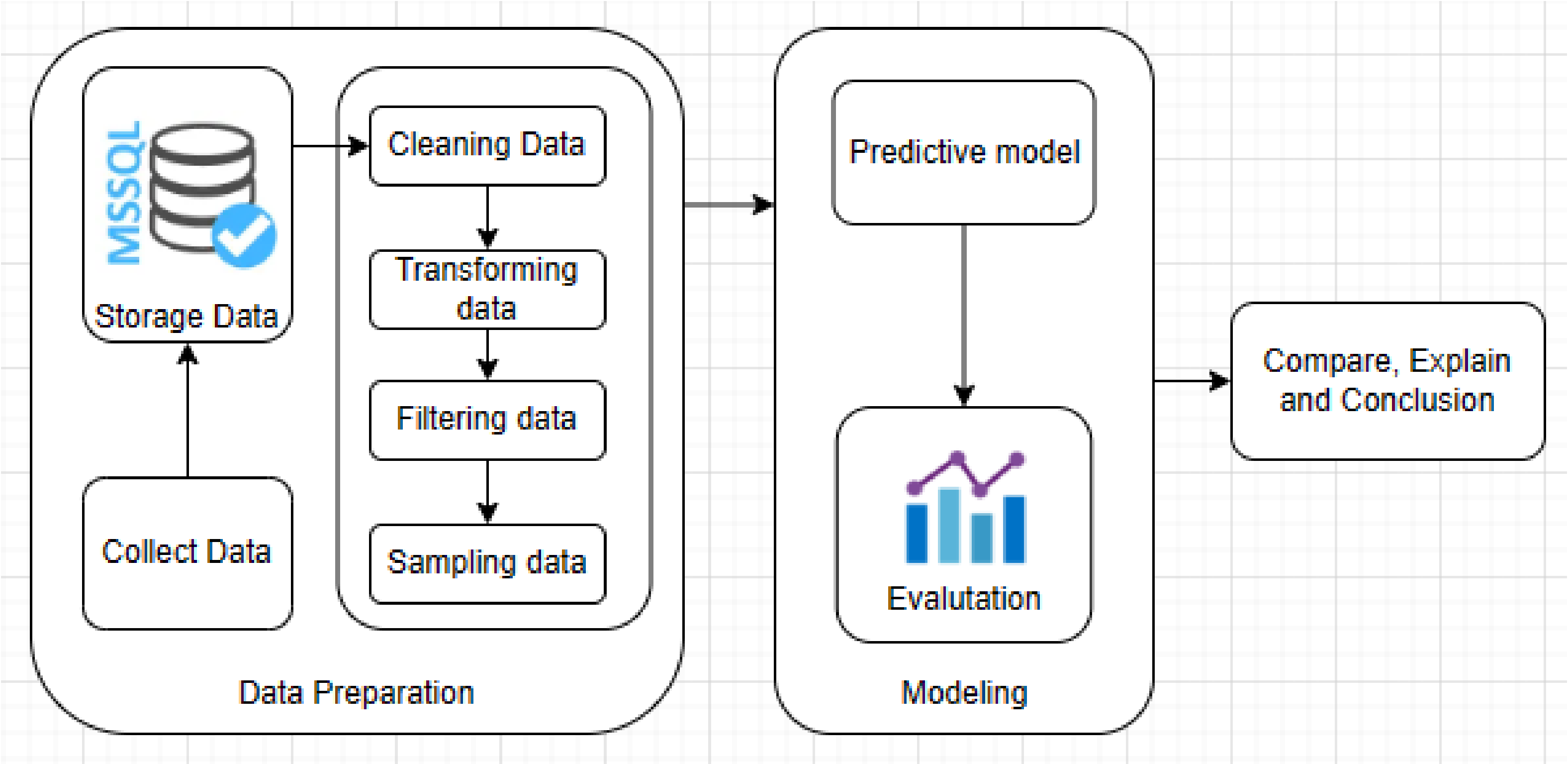
Method 2 – Deep Learning Approach



3. Methodology

3. Methodology

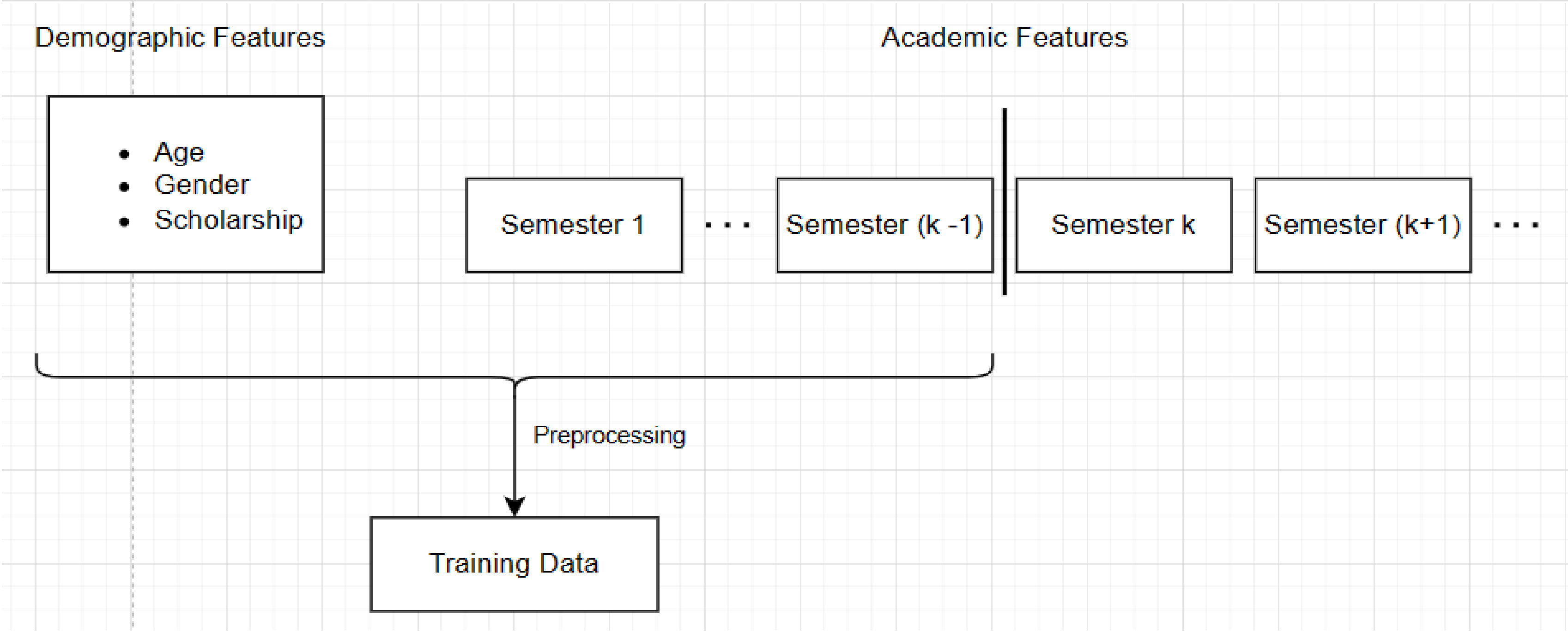
3.1 Workflow



3. Methodology

3.2 Model modeling

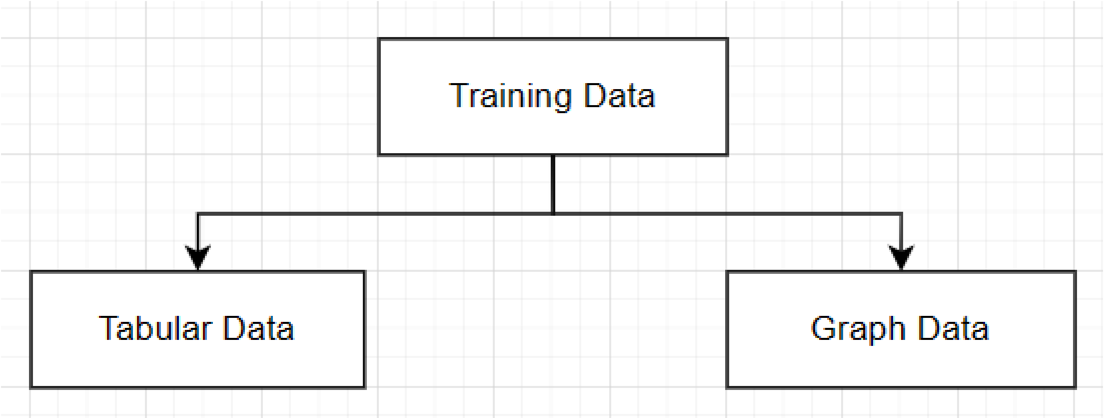
The model will use the demographic features combine with academic features, which was extracted from $(k-1)$ previous semesters to predict whether students drop out after $(k-1)$ semester.



- k is chosen semester for prediction.

3. Methodology

3.2 Model modeling



Tabular Data: Each student features will be represented by a row, and each column has a feature name that humans can understand.

Graph Data: Each student will be represented by a graph, where each node is a subject studied, Graph data is only used for the Graph classification model approach.

3. Methodology

3.3 Data preparation

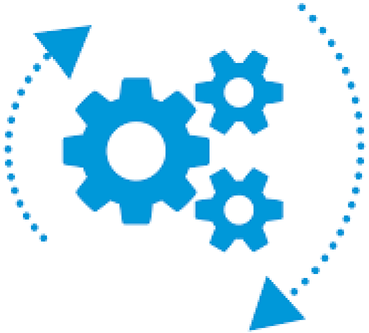
Data collection and storage:

- Our data is provided directly by the instructor, who helped describe the system for us to build data schema and crawler. After that, we send it back to you for crawler so that you can directly crawl as well as ensure the security of personal information and important data.
- The collected data was then stored in an MSSQL database.

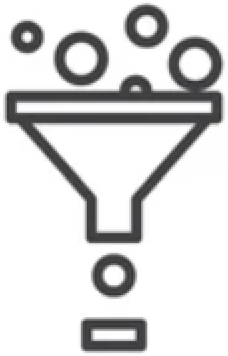
3. Methodology

3.3 Data preparation

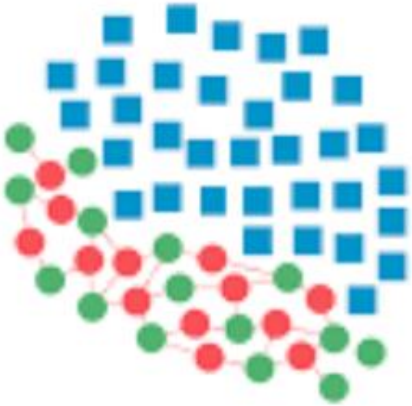
Cleaning: Remove outlier student that doesn't contribute and may harm to dataset



Transforming : create new represent similar features. Represent structure of data.



Filtering: Select information to use.



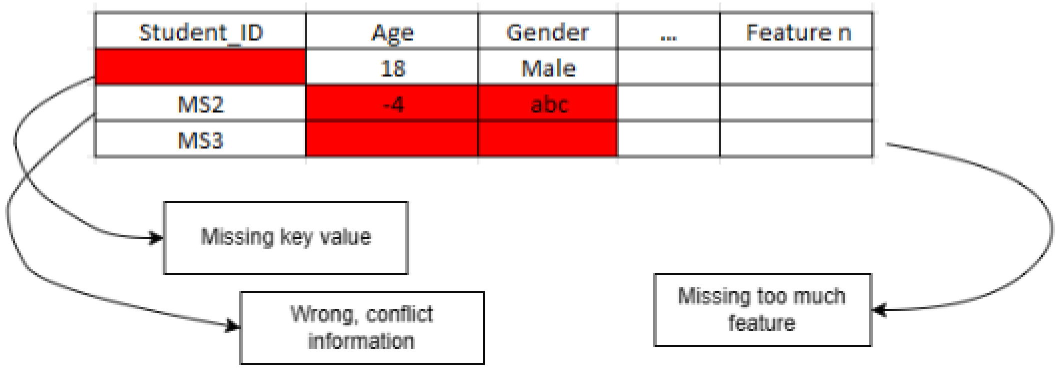
Sampling: sampling the dataset.

3. Methodology

3.3 Data preparation

Data Cleaning

Remove



Fix by re-fill mean value

Student ID	Feature 1	...	Feature n
MS1	0.8		
MS2	0.82		
MS3	0.8675		
MS4	0.85		
MS5	1		

3. Methodology

3.3 Data preparation

Data Transforming

Group GPA: The list of features contains the GPA of each department and the total average grade of all subjects in the selected department.

$$s(d) = \frac{\sum_1^{N(d)} avg_{d,i} * credit_{d,i}}{\sum_1^{n(d)} credit_{d,i}}$$

$$s0(d) = \frac{\sum_1^{N(d)} avg_{d,i}}{N(d)}$$

Group Average Component Grades (ACG): Group ACG composes the average of each type of component grade in the responding department.

$$avg(j) = \frac{\sum_1^{N(d)} score(j)_{d,i} * coef(j)_{d,i} * credit_{d,i}}{\sum_1^{N(d)} credit_{d,i}}$$

Group Coefficient Component Grades (CCG): Group ACG composes the average of each type of component weight in the responding department Data transform.

$$coef(j) = \frac{\sum_1^{N(d)} coef(j)_{d,i} * credit_{d,i}}{\sum_1^{N(d)} credit_{d,i}}$$

Group Ratio: Student learning grade rank by the following rule: 9.0–10.0 (A+), 8.0–9.0 (A), 7.0–8.0 (B+), 6.0–7.0 (B), 5.0–6.0 (C+), 4.0–5.0 (C), 3.0–4.0 (D), <3 (F). The group ratio is the list of each rank ratio.

$$ratio(r) = \frac{rank(r)_d}{\sum_1^r rank(r)_d}$$

Where:

- d: is the representative of the department.
- N(d): is the number of subjects in d^{th} department
- avg_i : the average grade of the subject i^{th}
- $credit_i$: the credit of subject i^{th}

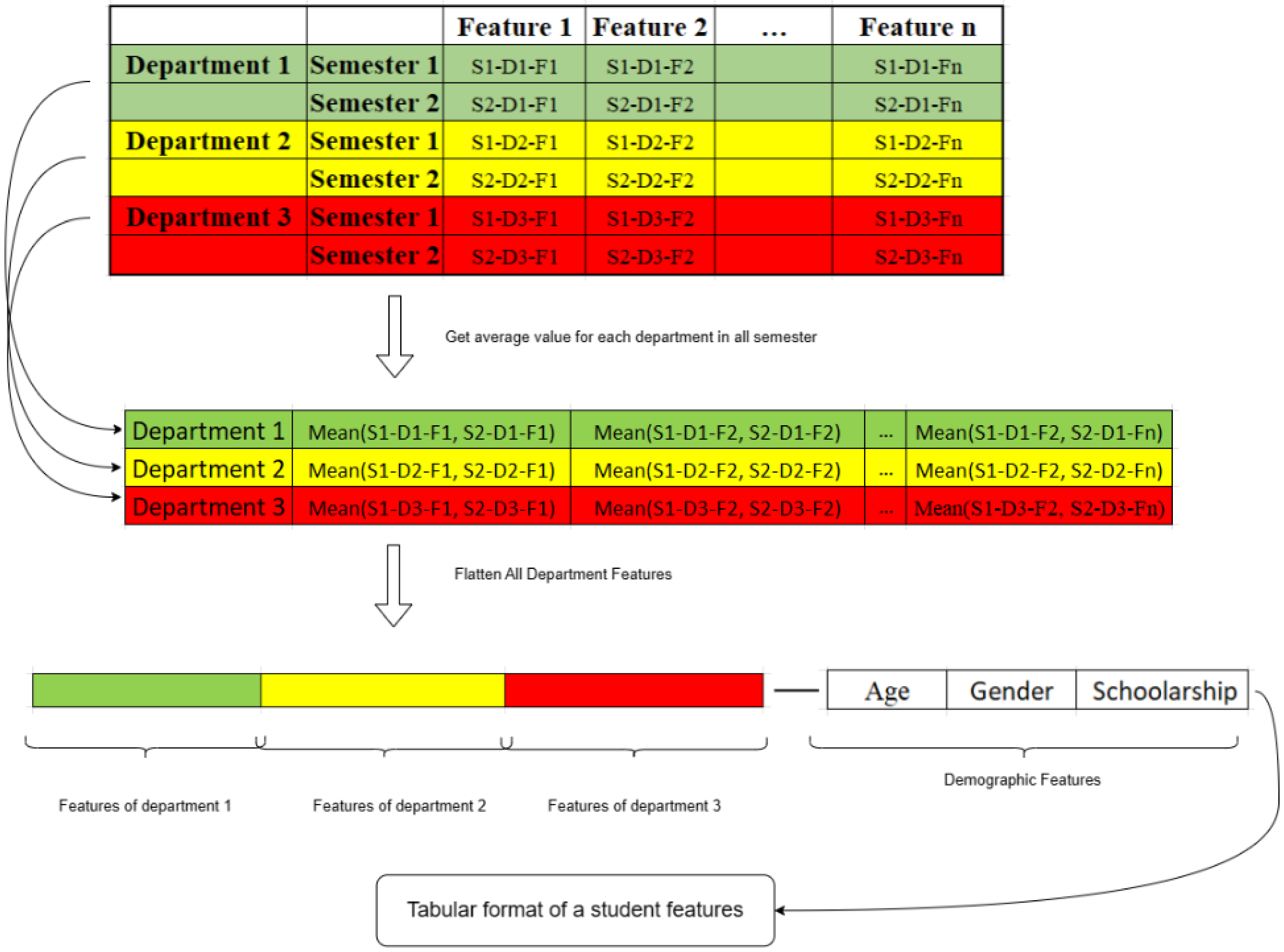
- $score(j)_{d,i}$: the j^{th} component grade for subject i^{th} in department d^{th}
- $coef(j)_{d,i}$: the j^{th} component weight for subject i^{th} in department d^{th}
- $rank(r)$: the number of grades that have rank r in the department d^{th}

3. Methodology

3.3 Data preparation

Data Transforming

Tabular Format

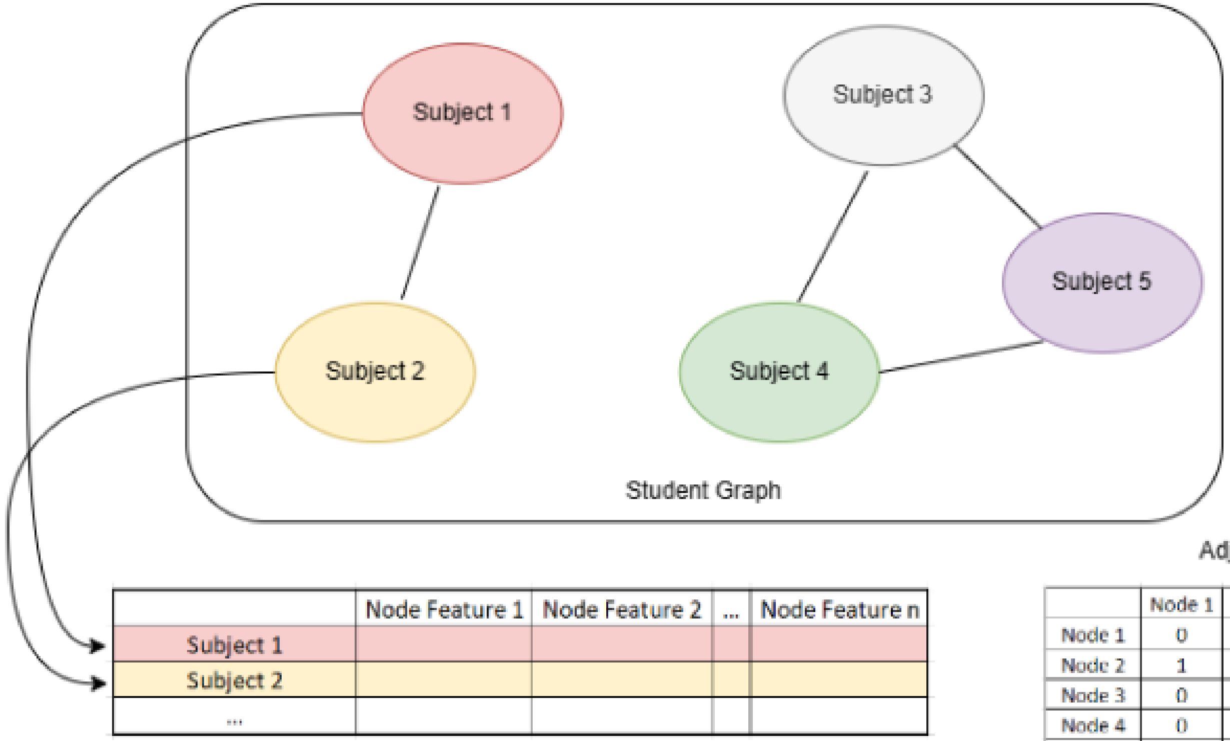


3. Methodology

3.3 Data preparation

Data Transforming

Graph Format



Adjacency matrix

	Node 1	Node 2	Node 3	Node 4	Node 5
Node 1	0	1	0	0	0
Node 2	1	0	0	0	0
Node 3	0	0	0	1	1
Node 4	0	0	1	0	1
Node 5	0	0	1	1	0

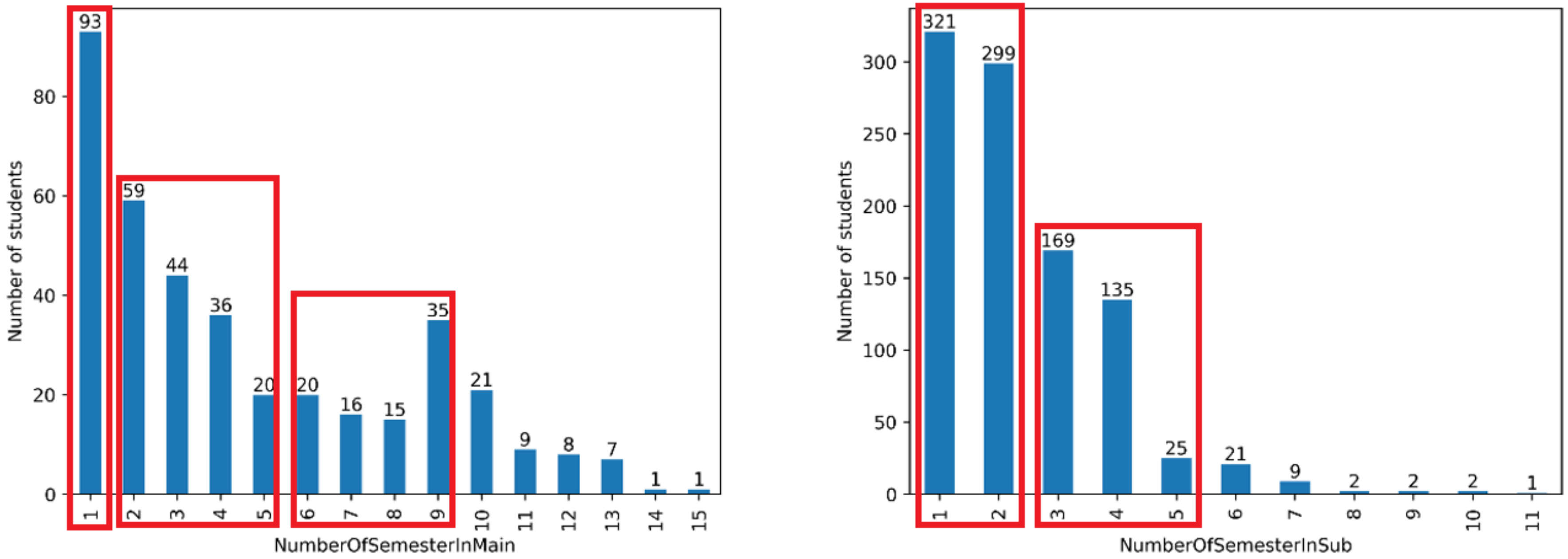
Two subjects in the same department has connected edge

3. Methodology

3.3 Data preparation

Data Filtering

- Select Semester 1 for prediction
- Separate into 2 main dataset:
  - English preparation dataset of all Student.
  - IT Student dataset from K13 only.

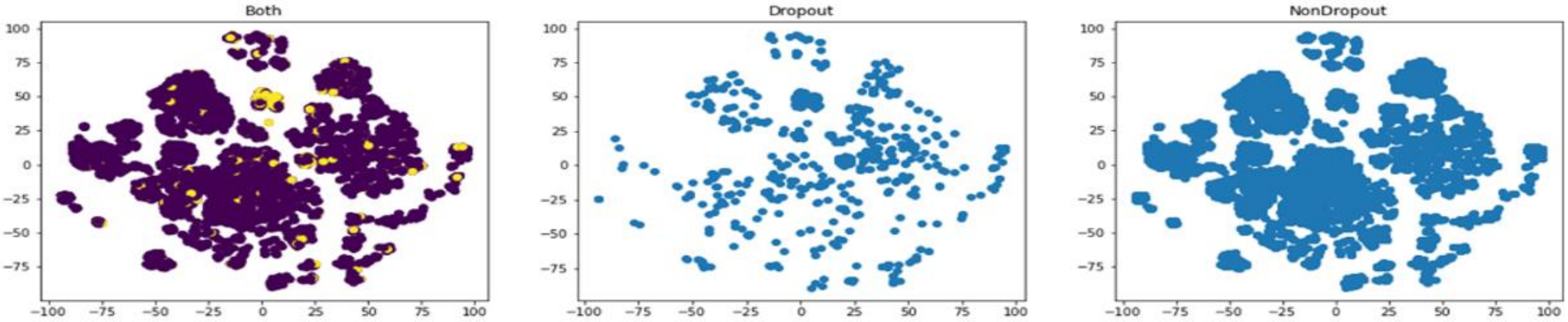


Number of Dropout Students

3. Methodology

3.3 Data preparation

Data Sampling



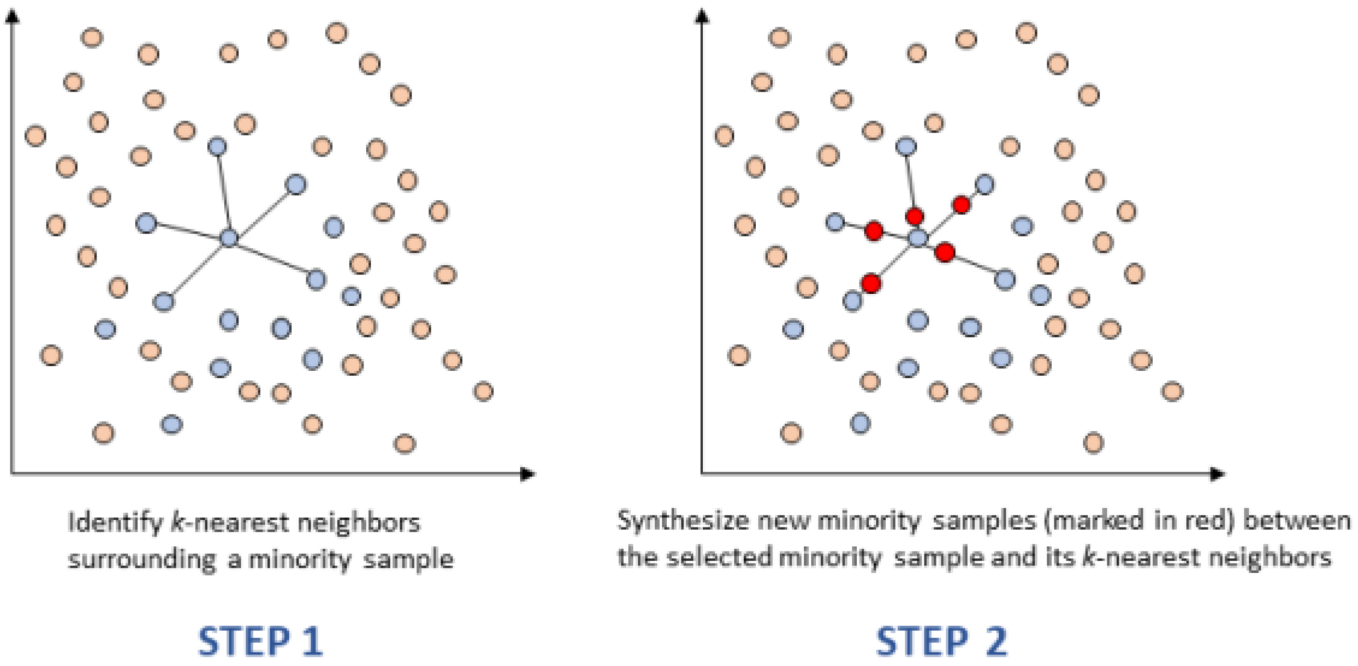
3. Methodology

3.3 Data preparation

Data Sampling

SMOTE

Repeat again from step 1 with other minority samples until reach designed sample size.



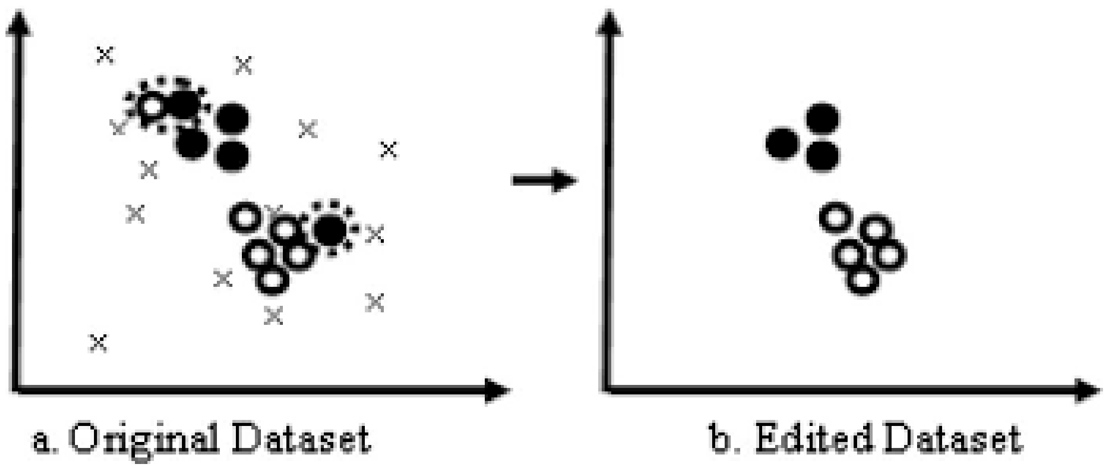
3. Methodology

3.3 Data preparation

Data Sampling

SMOTEENN = SMOTE + ENN

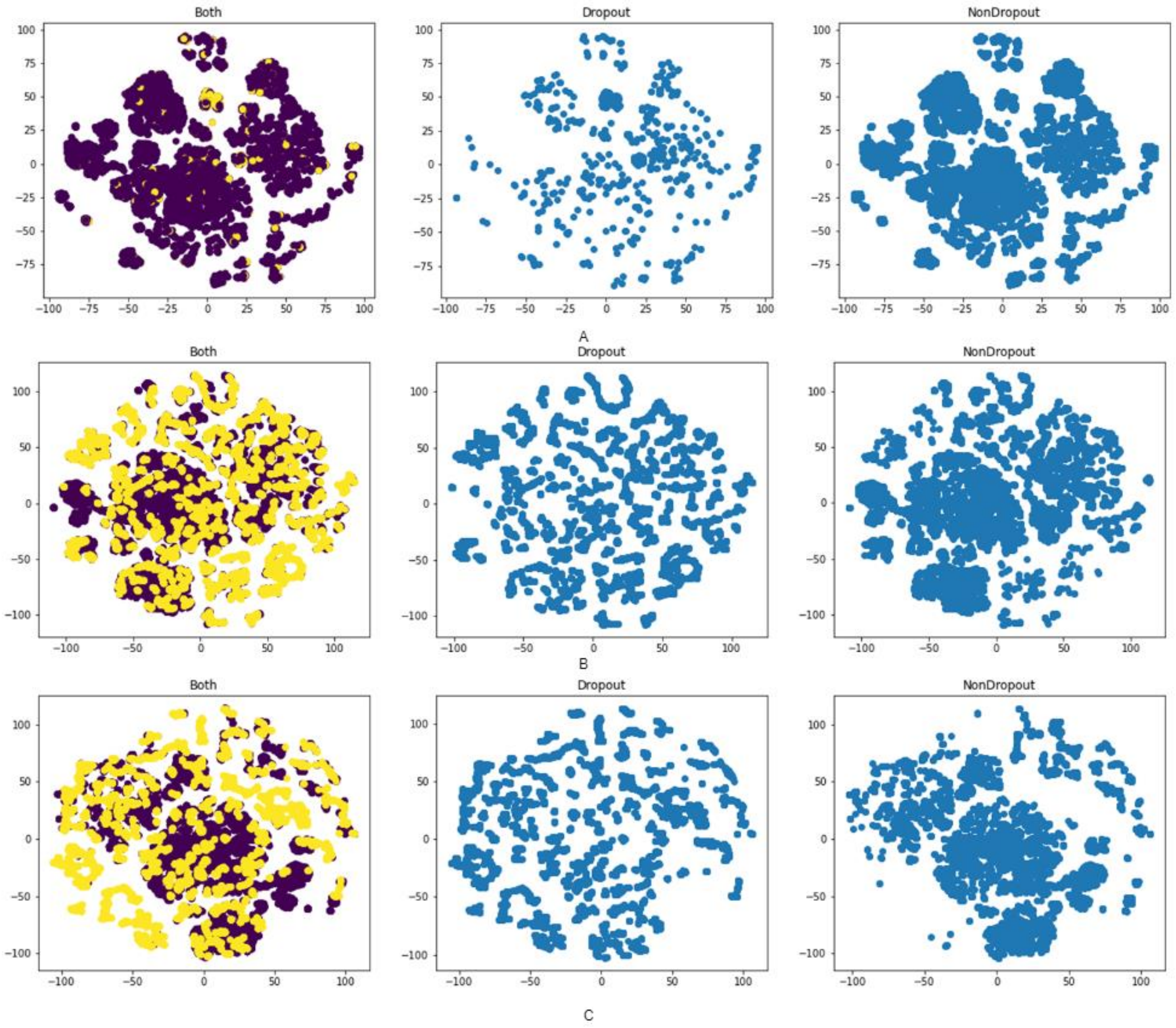
In Each iteration of SMOTE, Edited Nearest Neighbor will start after the minority data was synthetized in order to remove the new minority which was much differ from samples



3. Methodology

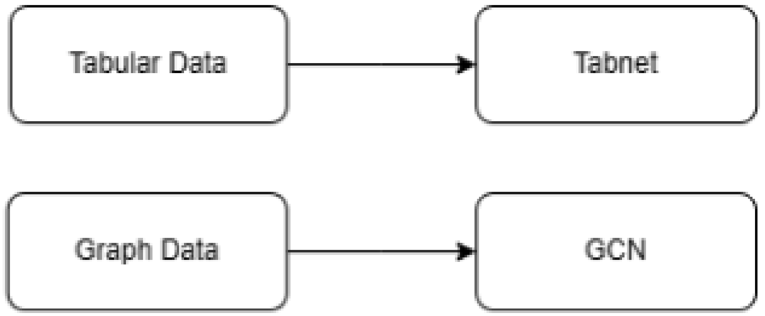
3.3 Data preparation

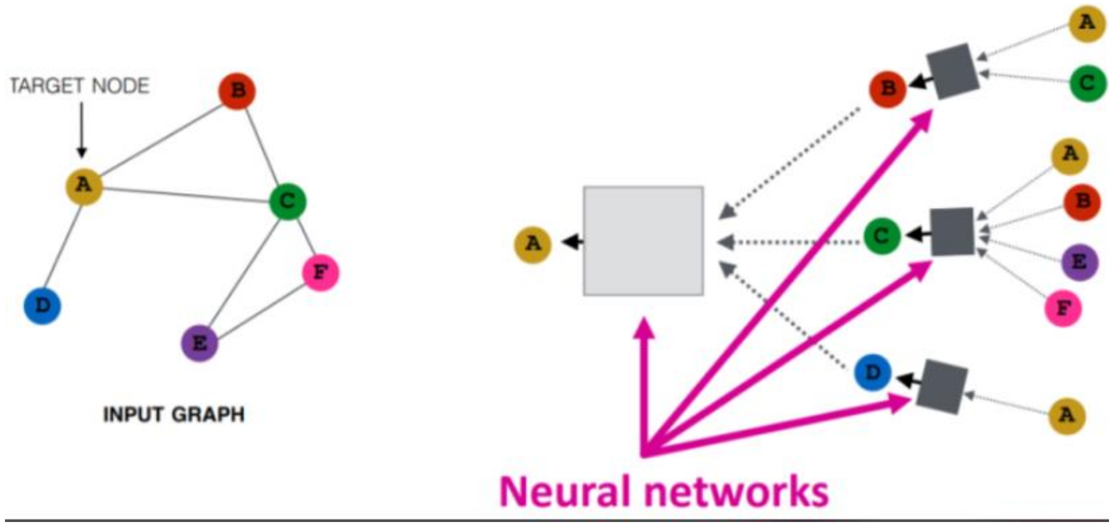
Data Sampling

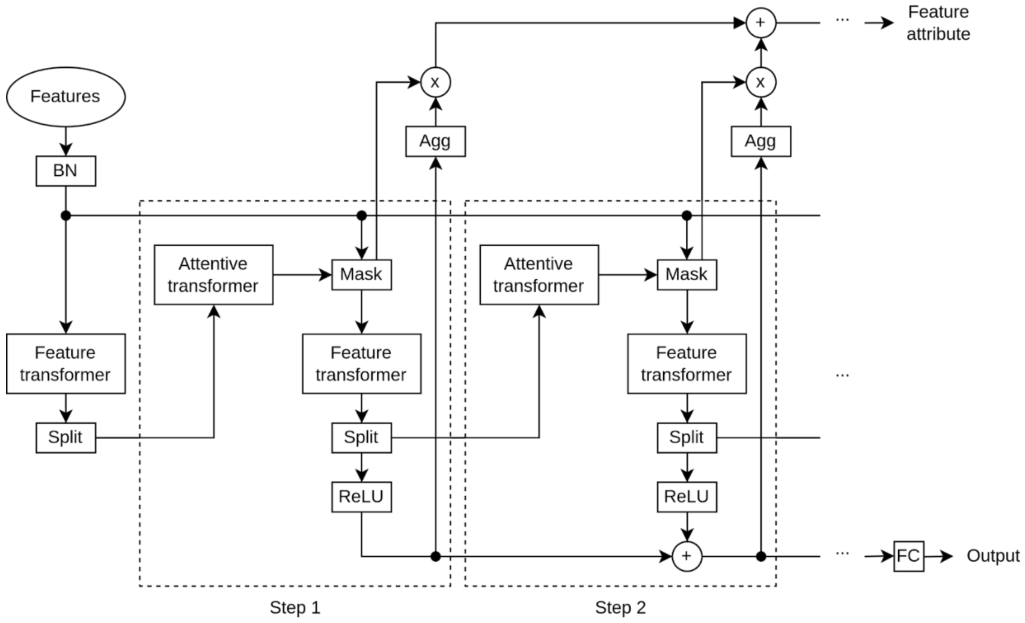


3. Methodology

3.4 Predictive model







3. Methodology

3.4 Predictive model

Cross Entropy loss

$$\boxed{\boldsymbol{CE(p_t) = -\log(p_t)}}$$

, where $p_t = \begin{cases} p & if\ y = 1 \\ 1 - p & otherwise \end{cases}$

$y = \{\pm 1\}$ specifies the ground-truth class

$p \in [0,1]$: model's estimated probability for class y=1

Focal loss

$$\boxed{\boldsymbol{FL(p_t) = -\alpha_t(1 - p_t)^{\gamma}\log(p_t)}}$$

, where

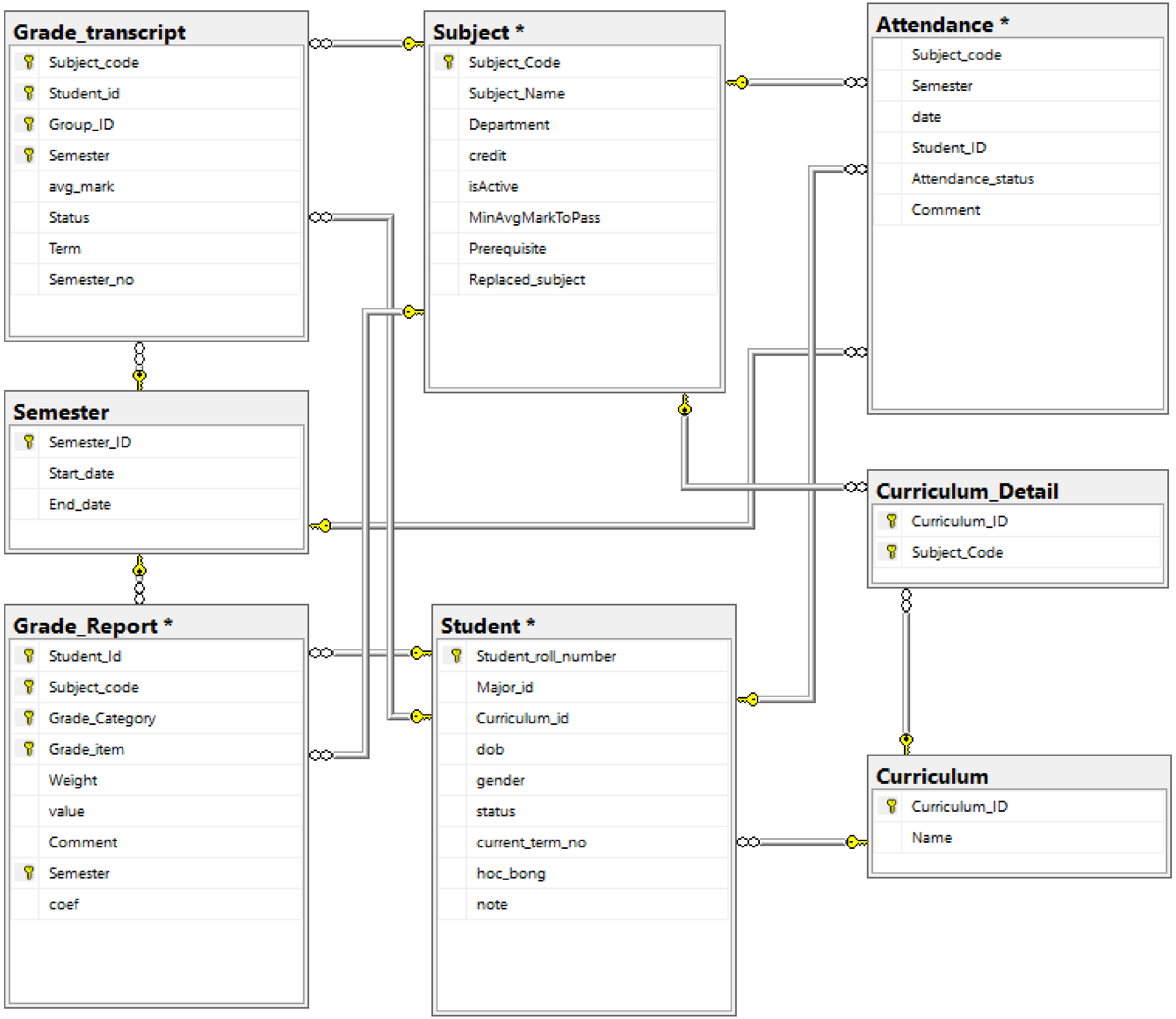
γ help focus on hard sample

α_t help to balance the loss according to the number of samples

3. Methodology

3.5 Data description

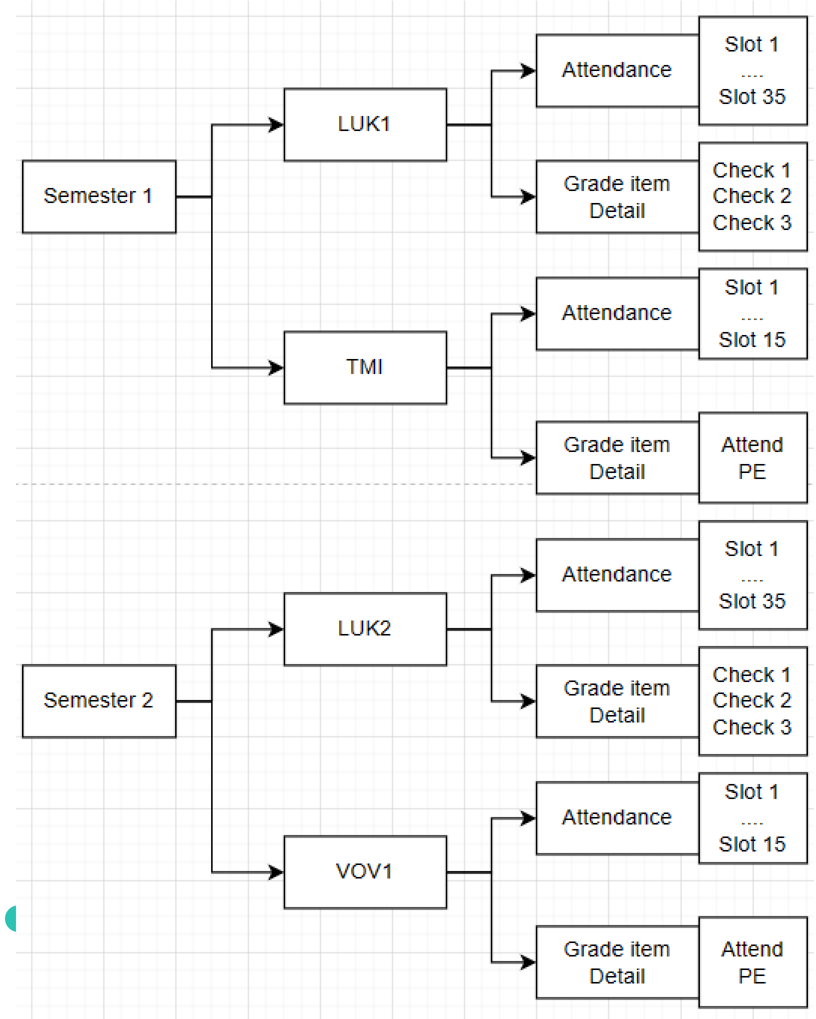
Dataset schema



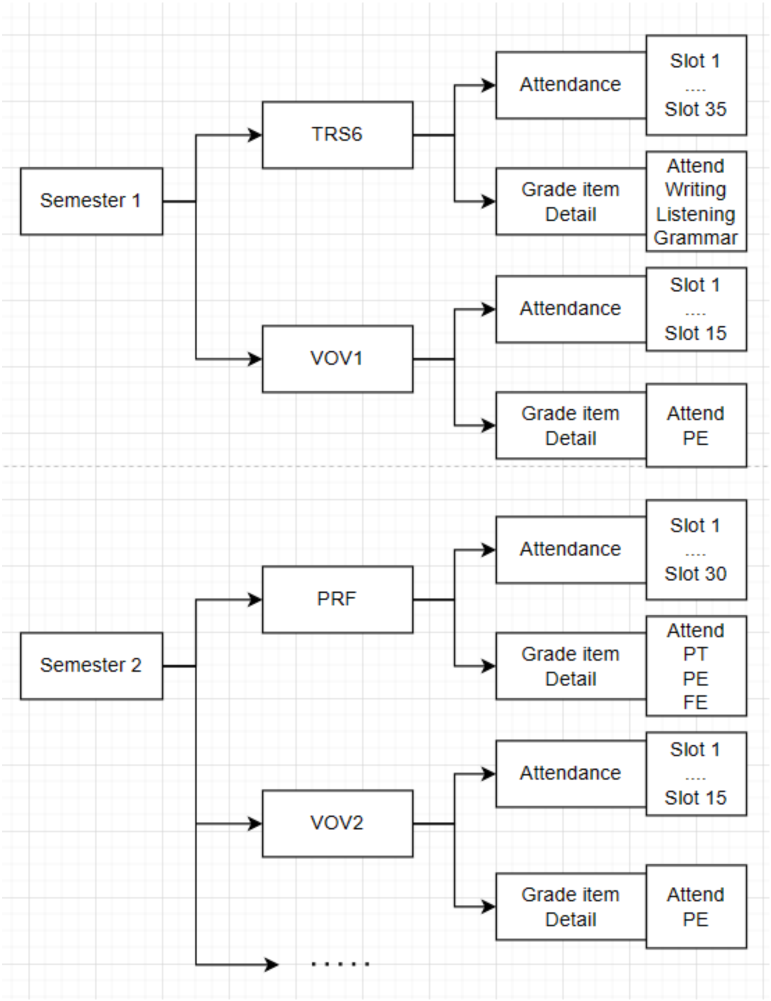
3. Methodology

3.5 Data description

Student 1

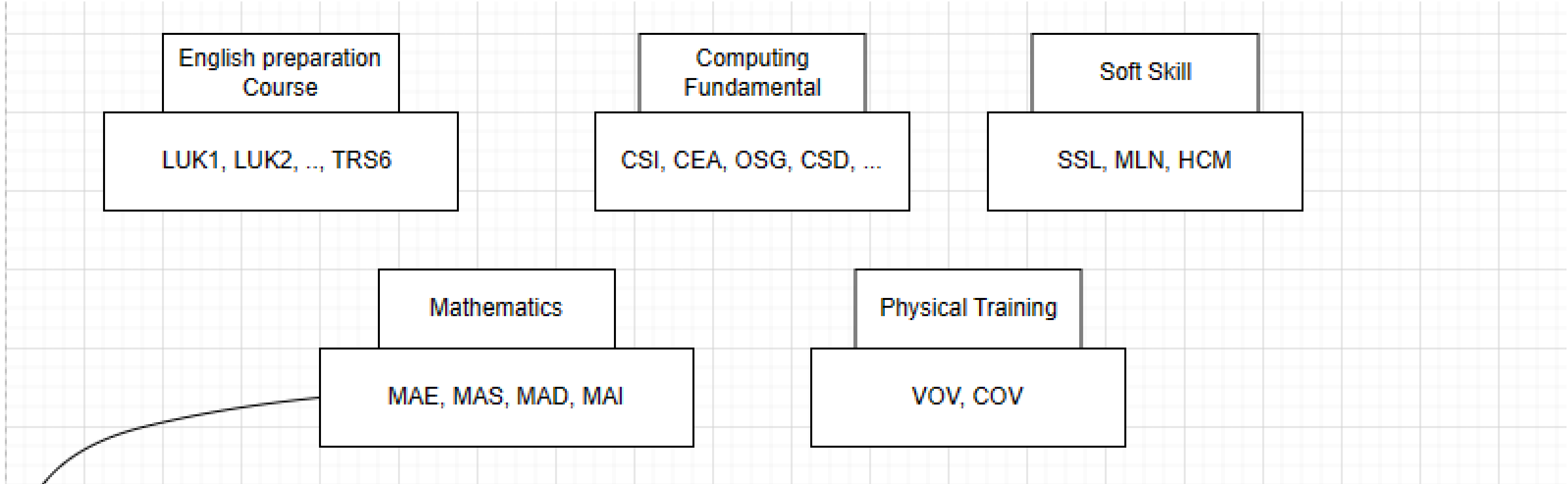


Student 2



3. Methodology

3.5 Data description

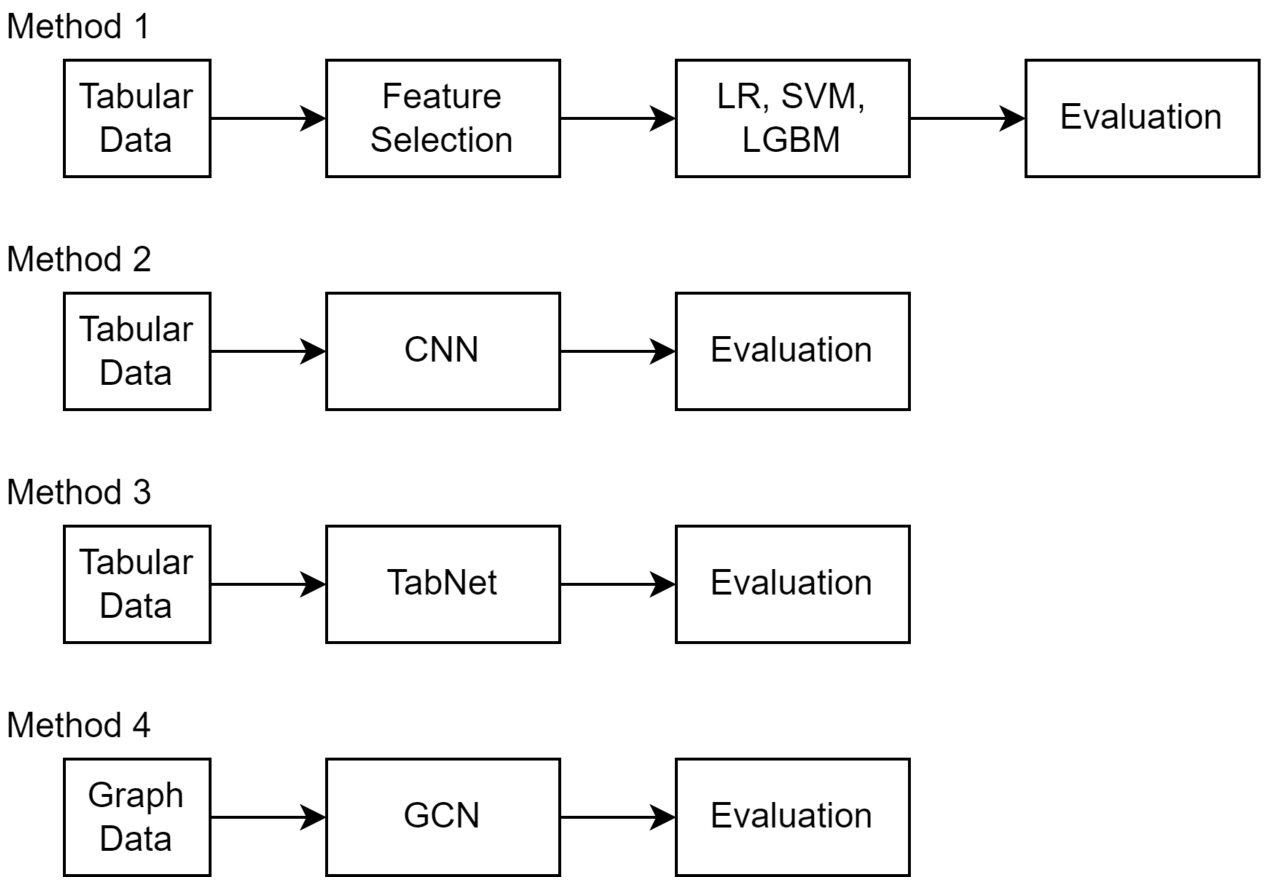


Subject_Code	Department	credit	isActive	MinAvgMarkToPass	Prequisite	Replaced_subject
MAS291	Mathematics	3	True	5	MAE101, MAC101	MAS202

4. Experimental and result

4. Experimental and result

4.1 Experimental Design



4. Experimental and result

4.1 Experimental Design

Config model

Hyperparameter	Values
# layers Conv	2
Size of FC	32
Size of Conv	128
Activation function	ReLU, Sigmoid
Optimizer	Adam
Batch	32
Dropout	0.8
Learning rate	1e-2

CNN

Hyperparameter	Values
n_steps	3
n_a, n_d	8
n_independent	2
n_shared	2
Optimizer	Adam
Batch	256
VBS	128
Learning rate	1e-2

TabNet

Hyperparameter	Values
# layers GCN	2
Size of FC	16
Activation function	Sigmoid
Optimizer	Adam
Batch	32
Dropout	0.7
Learning rate	1e-2

GCN

4. Experimental and result

4.2 Experimental Data

<table>
<tr><th rowspan="2">Tabular format</th><th colspan="3"># Students</th><th rowspan="2"># Features</th></tr>
<tr><th>Total</th><th>Non-dropout</th><th>Dropout</th></tr>
<tr><td>IT dataset</td><td>7836</td><td>7458</td><td>378</td><td>266 features: 5 demographic features, 29 performance features x 9 Departments</td></tr>
<tr><td>English preparation dataset</td><td>21429</td><td>20443</td><td>986</td><td>34 features: 5 demographic features and 29 features of English Preparation Subjects.</td></tr>
</table>

4. Experimental and result

4.2 Experimental Data

Graph format	# Students			# features
	Total	Non-dropout	Dropout	
IT dataset	**7836**	**7458**	**378**	• 29 features each node • # nodes depend on the subject student have learnt • Nodes that share the same department have adjacent edges

4. Experimental and result

4.3 Evaluation metric

Because of the dataset imbalance makes those metrics biased toward non-dropout class results. Therefore, we use the macro average for evaluation since the metric is the means of each class evaluation individually:

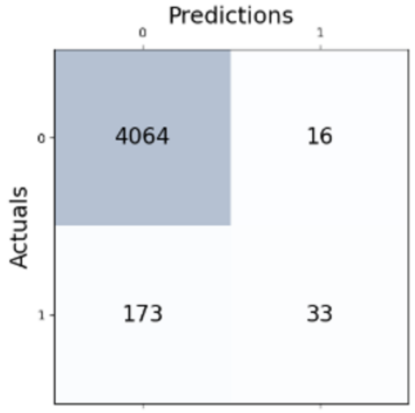
$$PrecisionMacroAvg = \frac{(Prec_1 + Prec_2 + \cdots + Prec_n)}{n}$$

$$RecallMacroAvg = \frac{(Recall_1 + Recall_2 + \cdots + Recall_n)}{n}$$

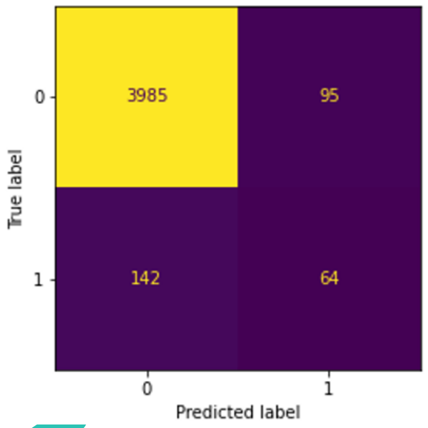
4. Experimental and result

4.4 Result

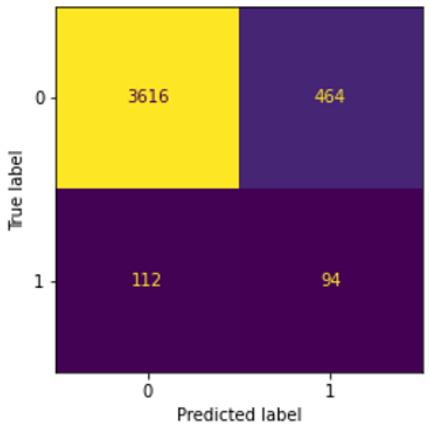
4.4.1 English preparation experience



CNN + Focal loss



LGBM



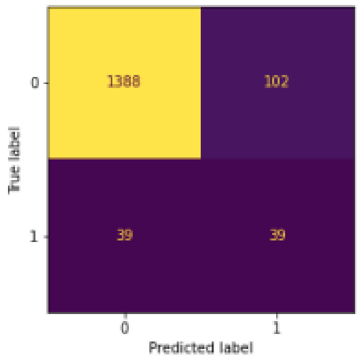
TabNet

	Accuracy	Precision-macro	Recall-macro	F1-macro
LR	0.74	0.54	0.68	0.52
SVC	0.76	0.54	0.67	0.52
LGBM	0.94	0.68	0.64	**0.66**
LGBM + Pearson	0.95	0.74	0.60	0.64
LGBM + Chi2	0.90	0.59	0.65	0.61
CNN	0.82	0.56	0.70	0.56
TabNet	0.73	0.56	**0.72**	0.52
CNN + Focal	**0.956**	**0.81**	0.57	0.61
Tabnet + Focal	0.953	0.75	0.60	0.64

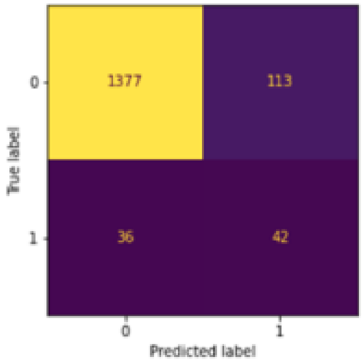
4. Experimental and result

4.4 Result

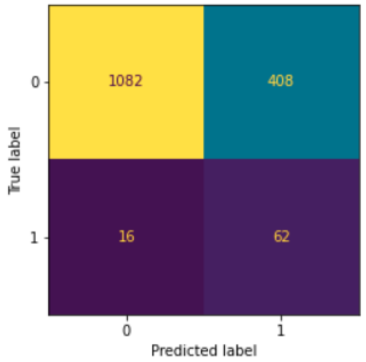
4.4.2 Information technology experience



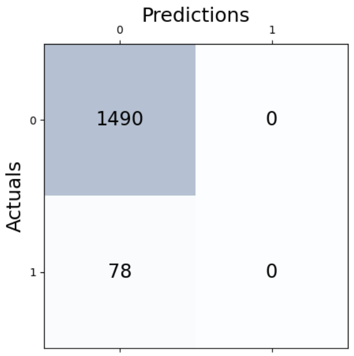
LGBM



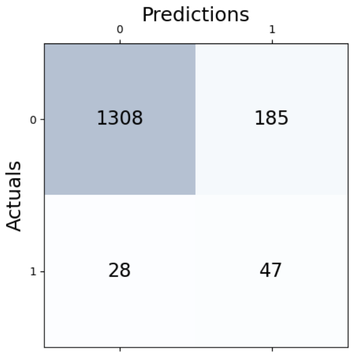
LGBM+Pearson



TabNet



CNN + Focal



GCN

	Accuracy	Precision-macro	Recall-macro	F1-macro
LR	0.78	0.55	0.69	0.54
SVC	0.81	0.55	0.69	0.56
LGBM	0.91	0.62	0.71	**0.65**
LGBM+Pearson	0.90	0.62	0.73	**0.65**
LR + Pearson	0.70	0.52	0.74	0.56
LGBM + Chi2	0.90	0.61	0.70	0.63
CNN	0.858	0.58	0.71	0.60
TabNet	0.73	0.55	**0.75**	0.53
GCN	0.864	0.60	**0.75**	0.62
CNN + Focal	**0.95**	0.47	0.5	0.48
TabNet + Focal	0.94	**0.72**	0.58	0.61
GCN + Focal	**0.95**	0.47	0.5	0.48

5. Conclusion

5. Conclusion

- We have **transformed** the raw database into features for the ML/DL
- **Partitioned** the dropout problem into 2 phases: English preparation, Main course phase
- Proposed methods achieves 72% and 75% recall macro in the English preparation and IT first-semester datasets (better than other methods in our study)

⇒ The proposed method has created hidden features to help separate the characteristics of dropout students (RQ2)

Limitation

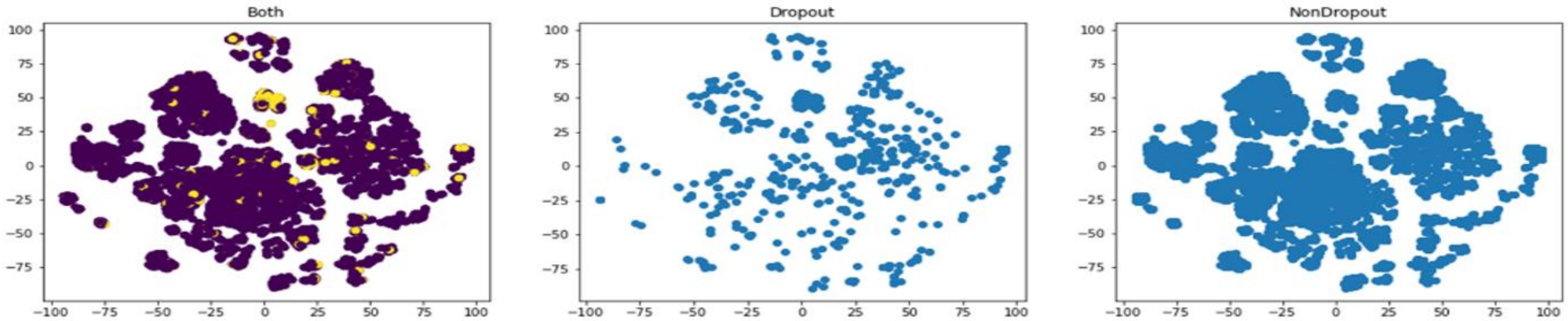
- Dataset imbalance (dropout students only take up 5% of the total dataset)
- Missing data (because of the change in curriculum over semesters)

⇒ Generate a dataset for a few missing features

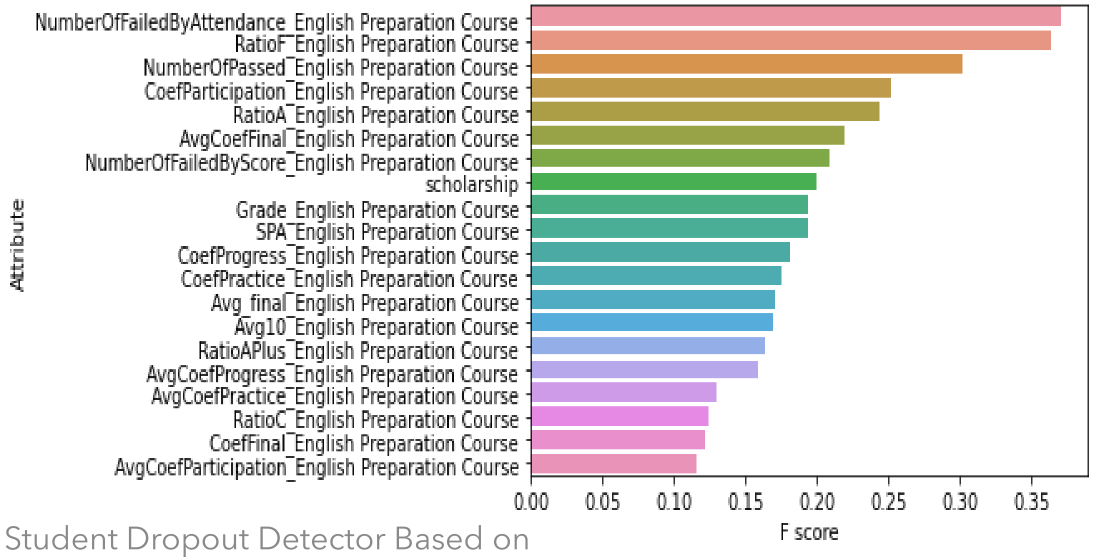
- Choosing when to suggest the prediction is also a challenging problem for us to solve

5. Conclusion

In addition, from our result, students' performance does not significantly influence dropout status by visualizing the dataset in t-SNE space, features of students' performance in each class are mixed. (RQ1)



With feature ranking measurement, attendance failed status has a **significant relationship** with dropout status. But we cannot use attendance-related features due to the lack of a dataset in attendance features.



5. Conclusion

The thesis contribution is:

- Analysis of students' performance and determine factors influencing Dropout.
- Investigate the influence of grade categorical and subject department in dropout prediction problems.
- Construct students' performance dataset based on subject department and grade detail.
- Analyze the efficiency of machine learning algorithms, convolution, graph neural networks, and tabular learning in academic dropout prediction problems.

Future work:

The binary state of dropout status also obstructs our studies because there are too few dropouts and the probability of those remaining student's dropouts in the following semesters. A solution for this problem is constructing a dropout rate representing the student’s dropout probability.

Thank you