Homework 6

First analysis of your data:

Q1 What is the number of good students who have already graduated? Good students are those, which graduated with GPA greater or equal to 3.0.

```
student_profile[(student_profile.STATUS=='G') & (student_profile.GPA>=3.00)].shape
```

```
(1439, 12)
```

Number of good students who have already graduated is 1,439.

Q2 What is the number of bad students who have already graduated? Bad students are those, which graduated with GPA less than 3.0.

```
student_profile[(student_profile.STATUS=='G') & (student_profile.GPA<3.00)].shape
```

```
(846, 12)
```

Number of bad students who have already graduated is 846.

Q3 For each department, give the number of students who have already graduated.

```
student_profile_G['DEPARTMENT'].value_counts()
```

```
Electrical Engineering       799
Civil Engineering            400
Environmental Engineering    375
Chemical Engineering         349
Machanic Engineering         200
Computer Engineering         162
Name: DEPARTMENT, dtype: int64
```

```
student_profile_G['DEPARTMENT'].value_counts(normalize=True) * 100
```

```
Electrical Engineering       34.967177
Civil Engineering            17.505470
Environmental Engineering    16.411379
Chemical Engineering         15.273523
Machanic Engineering          8.752735
Computer Engineering          7.089716
Name: DEPARTMENT, dtype: float64
```

Department Name	Number of occurences	Percentage
Computer Engineering	162	7.09
Civil Engineering	400	17.51
Electrical Engineering	799	34.97
Chemical Engineering	349	15.27
Machanic Engineering	200	8.75
Environmental Engineering	375	16.41

Q4 What is the total number of first year students?

```
student_profile[student_profile.STATUS=='N'].shape
```

```
(260, 12)
```

Total number of first year students is 260.

Data Pre-processing:

Q5 What are all necessary tables to be used in this data-mining project?

All necessary tables to be used in this data-mining project is STUDENT_PROFILE and STUDENT_GRADE

Training Phase:

Q6 Using Decision tree

Q6.1 What is the class label attribute? How many classes to be predicted?

Class label attribute is Department, 6 classes to be predicted.

Department Name
Computer Engineering
Civil Engineering
Electrical Engineering
Chemical Engineering
Machanic Engineering
Environmental Engineering

Q6.2 Split data into training set (70%, stratified sampling) and test set (30%, stratified sampling). What is the number of training records?

1006	6812	Chemical En...	m	
1007	6814	Electrical Eng...	m	

ExampleSet (1,007 examples, 2 special attributes, 9 regular attributes)

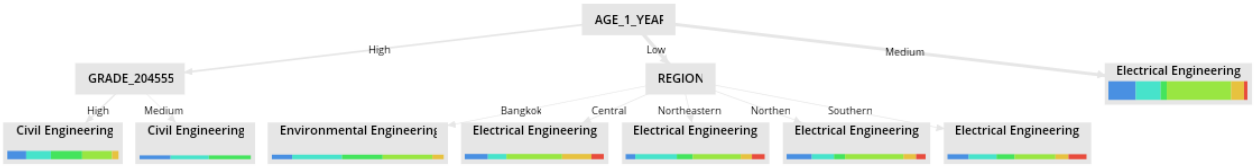
Number of training records is 1007.

Q6.4 What is the accuracy of the model with the hold-out method? Using the following parameters: Criterion: gini_index, minimal size for split: 55, minimum leaf size: 50

accuracy: 38.66%

Accuracy of the model is 38.66%.

Q6.5 How many levels are in your decision tree?



This decision tree has 2 levels.

Q6.6 Explain your tree in terms of predictors to predict Department

if AGE_1_YEAR = HIGH and GRADE_204555 = HIGH then Civil Engineering

if AGE_1_YEAR = HIGH and GRADE_204555 = MEDIUM then Civil Engineering

if AGE_1_YEAR = LOW and REGION = Bangkok then Environmental Engineering

if AGE_1_YEAR = LOW and REGION = Central then Electrical Engineering

if AGE_1_YEAR = LOW and REGION = Northeastern then Electrical Engineering

if AGE_1_YEAR = LOW and REGION = Northern then Electrical Engineering

if AGE_1_YEAR = LOW and REGION = Southern then Electrical Engineering

if AGE_1_YEAR = MEDIUM then Electrical Engineering

Q6.7 Which attribute(s) is the most related to Department?

AGE_1_YEAR is the most related to department.

Q6.8 Give precision and recall for each class?

accuracy: 38.66%

	true Chemical Engineer...	true Environmental Eng...	true Civil Engineering	true Electrical Engineeri...	true Machanic Engineer...	true Computer Enginee...	class precision
pred. Chemical Engine...	0	0	0	0	0	0	0.00%
pred. Environmental En...	3	12	4	12	6	0	32.43%
pred. Civil Engineering	12	24	32	19	8	0	33.68%
pred. Electrical Enginee...	60	55	19	123	25	18	41.00%
pred. Machanic Engine...	0	0	0	0	0	0	0.00%
pred. Computer Engine...	0	0	0	0	0	0	0.00%
class recall	0.00%	13.19%	58.18%	79.87%	0.00%	0.00%	

Department Name	Precision	Recall
Computer Engineering	0.00%	0.00%
Civil Engineering	33.68%	58.18%
Electrical Engineering	41.00%	79.87%
Chemical Engineering	0.00%	0.00%
Machanic Engineering	0.00%	0.00%
Environmental Engineering	32.43%	13.19%

Q6.9 Which major(s) is the most accurate? Why?

Electrical Engineering is the most accurate because at the first level of decision tree predict electrical engineering in 1 of 3 labels and at the second level of decision tree predict electrical engineering in 4 of 7 labels

That because 34.97% of all data is Electrical Engineering.