Name ID

Lab-practice on Imbalanced Data Classification

Submit to: 1) Student ID and lab hand

DAKDL University Case Study

DAKDL managers noticed that student major selection is very important factor of his/her success. So, they decided to apply analytics techniques in order to support students in selecting the most appropriate major.

They managers wish to determine the *characteristics* of good students for each major. At DAKDL, there are 6 majors of engineering, which are computer, electrical, mechanical, chemical, environmental and civil. Good students (GPA >= 3.0) are those that obtain good *grade point average* (GPA) for their graduation for a specific major.

More precisely, DAKDL managers want to determine which majors to recommend based on profiles of students. *Profiles* of students are combinations of attributes values such as:

- first year age,
- gender,
- region,
- first year *performance*: grade obtained in each course in the first year,
- status [G = Graduated, N = New (First-Year Student), R = Re-grade]

Several methods are possible to reach the objective. A simple method is to construct a classification model to predict the most appropriate major for a first-year student (Status = G).

Your task is to do necessary steps for pre-processing data and to discover patterns for each major. Discovered patterns can help students to select the appropriate major (among 6 possible majors) according to their characteristics when they enter the second year.

Name ID

Data exploration:

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

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

Q3 For each department, show number of good students who have already graduated?

Department Name	Number of occurrences
Computer_Engineering	
Civil_Engineering	
Electrical_Engineering	
Chemical_Engineering	
Mechanical_Engineering	
Environ_Engineering	

Q4 Is the data Balanced? What is the Imbalance-ratio (Number of Majority/ Number of Minority) between each pair of majority/minority class?

Majority-Class	Minority-Class	Imbalance-ratio

Q5 What is the total number of first year students to be predicted ?

Name ID

Data Pre-processing:

Q6 Show your pre-processing steps?

Training Phase:

Build a classification model in order to predict which majors should be appropriate to which students. Training data consist of good students already graduated.

Notice that good students are those, which graduated with GPA greater or equal to 3.0. Following is the list of attributes necessarily for construction data classification model:

1. Gender
2. Age-at-first-year
3. Region
4. Department
5. Grade obtained for each course during the first year study (204111,204222,204333,204444,204555,204666)

ID (**Key**)	Gender	Age_1_year	Region	Dept	204111	204222	204333	204444	204555	204666
37058063	male	17	Central	Civil-Eng	2	2.5	3	2	3	3.5
37058167	male	18	South	Electrical-Eng	2	2.5	3	2	3	3.5
…………	….				…..	…..				

Q4 Model construction. Use 10-fold cross-validation for evaluating your model

Q4.1 What is the class-label attribute? How many classes to be predicted?

Q4.2 Give three best attributes the predict major?

Q4.3 What is the accuracy of your model?

Q4.4 Give precision of the most accurate major? Give precision of the least accurate major? Which major is the most accurate? Why?

Q5 Accuracy improvement => Your score will be based-on the best accuracy Obtained

Q5.1 Explain in detail your steps of improving accuracy.

Q5.2 What is your final accuracy? Give precision of the most accurate major? Give precision of the least accurate major?

Q6 Use of Model to Predict Unseen

Q6.1 Determine appropriate majors for the following students?

Student-ID	Major
5342	
5364	
5381	
5881	

Submit => 1) Student_ID_answer.pdf (containing your answers)
2) Student_ID_model_1.xml (or .ipynb)
3) Student_ID_model_2.xml (or .ipynb)