Classification of the length of study based on the student characteristics and academic performance in FMIPA Unsyiah

Fitriana A R*, J Aida, N Salwa and A Rusyana
Department of Statistics, Syiah Kuala University, Banda Aceh, Indonesia

*E-mail: fithriana.ar@fmipa.unsyiah.ac.id

Abstract. The ideal length of study for undergraduate students is about four to five years, but the fact is there are still many students of FMIPA Unsyiah takes longer time to graduate. This research aims to classify the factors that affect that length of study based on the provided data in FMIPA Unsyiah, which are some data of characteristics and educational background of the students. The statistical methods used are CHAID (Chi-Squared Automatic Interaction Detection). CHAID method will divide the data into groups of smaller based linkages between response variable with the explanatory variables. This research used the data from undergraduate students who have graduated from the period of January 2007 to August 2014. It is found that the significant factors influencing the students’ length of study are the GPA, the education level of supervisor, the final exam score, the gender, and the program of study. It is also found that the students whose length of study in time and average are students with excellent GPA score, received grade A in their final exam, and are female. While students within late length of study category are students in satisfactory GPA with the level education of their final writing supervisors are doctorate. In terms of classification accuracy, the percentage of accuracy of the model is 61.4%.

1. Introduction
The formal education process required the period of study in accordance with each level of the education. The faculty of Mathematics and Sciences (FMIPA) of Syiah Kuala University (Unsyiah) has developed S1 undergraduate program of study such as Mathematics, Physics, Chemistry and Biology. To graduate, the students need to pass at least 144 credits and maximum 160 credits. Based on the data in FMIPA Unsyiah, there are still many students who do not finish the study on time (4-5 years). From the data of undergraduated (S1) who graduated from January 2007 until August 2014, there are only 8% of students graduated in 8 semesters, 38% of students graduated in 9-10 semesters, and students who graduated after 10 semesters is 54%. From this information it would be interesting to classify the factors that influence the length of study based on the characteristics and educational background of the S1 students of FMIPA Unsyiah.

One suitable statistical method for classifying is CHAID (Chi-Squared Automatic Interaction Detection) method. CHAID method can show graphical tree structure in the form of diagram explaining the relationship between categorical response variable with categorical explanatory variable so easy to interpret. Some of the factors from provided data are used to see the effect on the length of study of S1 FMIPA Unsyiah students in the form of characteristics and educational background. The characteristics and educational background are gender, high school location area, program of study (major), educational level of supervisor, Grade Point Average (GPA), and the final exam score. These factors are used as explanatory variables in this study. While the period of student study (length of study) is categorized into three categories: fast (in time), average and late. The data of the length of
study is used as a response variable to determine the relationship between the response variables and the explanatory variables.

The CHAID method is commonly known as the Classification Tree Method. This method divides the data into smaller groups based on the relationship between response variables and explanatory variables. The CHAID method generally works on the relationship between response variables and some explanatory variables and then classifies the samples accordingly. The purpose of this study is to determine the factors and characteristics that are suspected to affect the length of study of S1 students of FMIPA Unsyiah.

2. Literature reviews

2.1. Length of study
Syiah Kuala University assigns 36 credits minimum for the first year students, and it should be completed in packages. For the following years, the credits taken depend on the Grade Point Average (GPA). The minimum credits for the undergraduate program is 144 (one hundred and forty-four), which are scheduled in 8 (eight) semesters with a maximum completion period of 14 (fourteen) semesters. If there are other issues emerged, the decisions about this length of study matters, then there is policy which be adjusted accordingly [1].

2.2. CHAID method
CHAID is an iterative technique that examines each of the explanatory variables used in the classification and arranges them based on the statistical significance level of Chi-square on the response variable [2]. CHAID is used to form segmentation that divides a sample into two or more different groups based on a certain criterion. This is then passed on by dividing the groups into smaller groups based on other explanatory variables.

CHAID is a statistical technique designed to divide samples into groups and subgroups that are separated from each other based on categorical results (ordinal) and some categorical explanatory variables (ordinal) [3]. The results of the analysis are carried out in stages where the most significant explanatory variables are used to divide the entire sample into two or more subgroups that are separated from each other. Basically, from several CHAID definitions above it can be concluded that CHAID is a method to classify categorical data where the purpose of the procedure is to divide the data set into subgroups based on its exploratory variables [4].

3. Research methodology

3.1. Data sources and determination of number of samples
The population in this study is all graduates of S1 FMIPA Unsyiah who graduated in January 2007 until August 2014. The data collected has not exposed gender and length of study (in months). The number of students who graduated from January 2007 to August 2014 is 1034 students. The determination of the number of samples in this study using the method of Non-Probability Sample (Selected Samples). The sample selection in this way ignores the principles of probability. The method used for this sampling is based on a purposive sampling sample that is taken only by consideration of the study itself, which considers the desired elements already present in the sample members taken.

3.2. The variables
Response variable used is a categorical variable. It is length of students’ study within 3 categories; fast, average, and late. Individual and environment are a hierarchical structure consists of units grouped at different levels [5]. Environment has strong influence in finishing of study. While the explanatory variables consist of seven variables which are individual dan environment variables, see Table 1.
Table 1. Research variables

| No | Variable                      | Scale and Category                      |
|----|-------------------------------|----------------------------------------|
| 1  | Length of study (Y)          | Ordinal                                |
|    |                               | 1 = Fast (≤ 8 semester)                 |
|    |                               | 2 = Average (8 < semester ≤ 10)        |
|    |                               | 3 = Late (> 10 semester)               |
| 2  | Gender (X1)                  | Nominal                                |
|    |                               | 1 = Female                             |
|    |                               | 2 = Male                               |
| 3  | Location of Senior High School (X2) | Nominal                           |
|    |                               | 1 = Non Banda Aceh                     |
|    |                               | 2 = Banda Aceh                         |
| 4  | Program of Study (X3)        | Nominal                                |
|    |                               | 1 = Mathematics                        |
|    |                               | 2 = Physics                            |
|    |                               | 3 = Chemistry                          |
|    |                               | 4 = Biology                            |
| 5  | Education Degree of Supervisor (X4) | Ordinal                           |
|    |                               | 1 = Non Doktorate                      |
|    |                               | 2 = Doktorate                          |
| 6  | GPA (X5)                     | Ordinal                                |
|    |                               | 1 = Satisfactory (≤ 2.75)              |
|    |                               | 2 = Good (2.76-3.50)                   |
|    |                               | 3 = Excellent (≥ 3.51)                 |
| 7  | Final Exam Grade (X6)        | Ordinal                                |
|    |                               | 1 = A                                  |
|    |                               | 2 = B+ & B                             |
|    |                               | 3 = C+ & C                             |

3.3. Research procedure

The steps of data analysis conducted in this research are as follows:
1. Analyzing the data descriptively.
2. Analyzing data with CHAID method to get the factors that influence the length of study of undergraduate students in FMIPA Unsyiah, using the following steps:
   a. Determining the category variables for each variable. The variable can be seen in Table 1.
   b. Combining the categories when there are more than two categories of the explanatory variable. Chi-square test is used to combine these categories, and the hypothesis is as follows:
      H0: there is no association between the two variables (mutual independence)
      H1: there is an association between the two variables
      If the test is not significant, then the category is merged and this step is repeated until there is no more insignificant category value. If the test is significant, it does not need to be combined and the smallest p-value adjusted will be calculated with the various categories for each explanatory variable.
   c. Selecting the insulator of the explanatory variable that has the smallest adjusted p-value. If the smallest adjusted p-value is greater than some significance levels of the insulator, then the insulation of the explanatory variable is stopped, and the node becomes the last node. Thus, a tree diagram can be generated.
3. Interpreting the data; concluding by interpreting the CHAID tree diagrams.
4. Results and discussion

4.1. Descriptive analysis

The total explanatory variables data collected in this research is 1034 undergraduate students of FMIPA Unsyiah who graduated from January 2007 to August 2014. The explanatory variables are GPA, education level of the supervisor, final exam grade, gender, program of study, and high school location. The descriptive of the explanatory variables can be seen in Table 2. Based on this table, it can be seen that the percentage of each category for all explanatory variables for the late category of length of study relatively higher compared to the percentage of students in the fast and average category of students’ length of study.

|                      | Length of Study (%) | Total  |
|----------------------|---------------------|--------|
|                      | Fast    | Average | Late   |        |
| GPA                  |         |         |        |        |
| Satisfactory*        | 0.00    | 2.03    | 22.82  | 24.85  |
| Good                 | 6.67    | 33.27   | 30.75  | 70.70  |
| Excellent            | 1.64    | 2.42    | 0.39   | 4.45   |
| Education Degree of  |         |         |        |        |
| Supervisor           |         |         |        |        |
| Non doctorate*       | 3.68    | 22.34   | 31.24  | 57.25  |
| Doctorate*           | 4.64    | 15.38   | 22.73  | 42.75  |
| Final Exam Score     |         |         |        |        |
| A                    | 5.61    | 18.67   | 11.99  | 36.27  |
| B+&B*                | 2.71    | 19.05   | 37.81  | 59.57  |
| C+&C*                | 0.00    | 0.00    | 4.16   | 4.16   |
| Gender               |         |         |        |        |
| Female*              | 6.48    | 32.21   | 35.78  | 74.47  |
| Male*                | 1.84    | 5.51    | 18.18  | 25.53  |
| Program of Study     |         |         |        |        |
| Mathematics*         | 0.58    | 10.15   | 16.83  | 27.56  |
| Physics*             | 2.51    | 7.64    | 13.73  | 23.89  |
| Chemistry*           | 2.80    | 9.38    | 10.83  | 23.02  |
| Biology*             | 2.42    | 10.54   | 12.57  | 25.53  |
| Location of Senior   |         |         |        |        |
| High School          |         |         |        |        |
| Non BandaAceh*       | 5.80    | 26.79   | 35.49  | 68.09  |
| BandaAceh*           | 2.51    | 10.93   | 18.47  | 31.91  |
| Respondents          | 8       | 38      | 54     | 100    |

* The percentage of the late length of study is greater than the fast and average length of study.

4.2. Analysis of CHAID

Respondents collected in this study are 1034 undergraduate students from FMIPA Unsyiah who graduated from January 2007 until August 2014. There are 58 graduates whose length of study more than the maximum semester allowed in undergraduate program (more than 14 semesters). They are not involved in this research. So the amount of data used in this study is 976 students from 4 undergraduate courses that graduated in January 2007 until August 2014.

The results of data analysis with CHAID method; the model summary section shows how the arrangement used to construct the model of the classification tree, also including the variables used in the analysis. In the explanatory variables, the procedure automatically excludes explanatory variables that significantly affect the final model. Therefore, it’s seen in the table of the specification section there are six explanatory variables, whereas, in the result, there are only five explanatory variables that significantly give effect to the final model. The explanatory variables are GPA, educational level of supervisor, final exam score, type of gender, and program of study. The variable of high school location is not included because of no significant effect on the final model.
Maximum tree depth in the specification section is to limit the growth of classification tree. In the table, the value of maximum tree depth is 3 is the value of the limit that has been set automatically. While the minimum case in branch node (minimum number of cases on the branch node) is 100 and the minimum case in leaf node is 50.

The result of the formation of the model is a classification tree with 11 nodes, six last nodes (terminal node) with the depth level of 3. The GPA is the best explanatory variable used to divide and explain the variable of length of study as the response variable. The tree diagram produced from the CHAID analysis can be seen in Figure 1.

From the above classification tree diagrams can be interpreted as follows:

a. The tree diagram has 11 nodes of node 0 (length of study), node 1 (satisfactory GPA), node 2 (GPA is good or excellent), node 3 (level of non-doctoral education), node 4 (doctoral education level) node 5 (final exam A), node 6 (final exam (B + & B) or (C + & C)), node 7 (female), node 8 (male), node 9 (mathematics majors) , and node 10 (majoring in physics, chemistry, and biology),
with the number of node terminals of 6, nodes 3, 4, 7, 8, 9, and 10 which means there are 6 segments of the length of study of S1 FMIPA Unsyiah students.

b. The significant variables affecting the length of study are GPA, education level of supervisor, final exam grade, gender, and program of study.

The result of the CHAID analysis tree diagram in Figure 2 shows that p-values and Chi-square test statistics of each explanatory variable considered to be related to the response variable can be summarized in Table 3.

Table 3. The p-values and chi-square test are explanatory variables of the tree diagram

| Variable                     | p-value | Chi-square test |
|------------------------------|---------|-----------------|
| GPA                          | 0.000   | 201.886         |
| Education Level of Supervisor| 0.002   | 9.394           |
| Final Exam Score             | 0.000   | 46.372          |
| Gender                       | 0.003   | 9.131           |
| Program of Study             | 0.001   | 15.150          |

The hypotheses in this test are as follows:
H0: There is no relationship between length of study and GPA, education level of supervisor, final exam grade, gender, and program of study.
H1: There is a relationship between length of study and GPA, education level of supervisor, final exam grade, gender, and program of study.

From Table 3, it can be seen that if the decision is made based on the p-value, where all these values are less than $\alpha = 0.05$; i.e. 0.000; 0.002; 0.000; 0.003; and 0.001; then reject H0. It can be concluded that there is a relationship between the length of study with GPA, the education level of the supervisor, final exam grade, gender, and the program of study.

c. Each terminal node can be interpreted as one segment of the length of study of FMIPA S1 students of Unsyiah.

CHAID analysis tree diagram in Figure 1 above explained that the length of study of FMIPA Unsyiah S1 graduated in January 2007 through August 2014 can be divided into several segments. While to know the most influential variables on each category of length of study can be seen from the largest percentage of terminal nodes. In the data there are six segments are as follows:

Table 4. Segmentation and percentage of length of study of S1 FMIPA Unsyiah students

| Terminal node | Segment | Characteristic                                                                 | Percentage of Student |
|---------------|---------|--------------------------------------------------------------------------------|-----------------------|
| 3             | 1st (node 0,1,3) | Students whose length of study are fast with satisfactory GPA and the level of education of the supervisor is not a doctorate. | 0.0%                  |
|               |         | Students whose length of study are average with satisfactory GPA and the education level of their supervisor is not a doctorate. | 14.8%                 |
|               |         | Students with late of length of study, with GPA satisfactory GPA, and the education level of their supervisor is not a doctorate. | 85.2%                 |
| 4             | 2nd (node 0,1,4) | Students whose length of study are fast with GPA are satisfactory and the education level of their supervisor is doctorate. | 0.0%                  |
In Table 4, it is predicted that the highest number of students whose length of study in fast category is the 3rd segment with 17.3%, that is, students who have fast length of study with good GPA or excellent get final exam A, and female. Segment whose length of study has percentage of 60.8% is 3rd segment also, that is student of average length of study with GPA is good or excellent get the final exam A, and female. While the segment of the late length of study is the second segment, the students with late length of study with a satisfactory GPA and educational level of their thesis supervisor is doctoral has a percentage of 96.9%.

| Terminal node | Segment          | Characteristic                                                                 | Percentage of Student |
|---------------|------------------|-------------------------------------------------------------------------------|-----------------------|
|               |                  | Students whose length of study are average with a satisfactory GPA and the education education of thesis supervisor is doctorate. | 3.1%                  |
|               |                  | Students with late length of study, with GPA are satisfactory and the levels of education of their supervisor are doctorate. | 96.9%*                |
| 7             | 3rd (node 0,2,5,7) | Students whose length of study are fast with GPA are good or excellent, final exam A, and female. | 17.3%*                |
|               |                  | Students whose length of study is average, with GPA good or excellent, final exam A, and female. | 60.8%*                |
|               |                  | Students with late length of study with GPA are good and excellent, final exam A, and female. | 22.0%                 |
| 8             | 4th (node 0,2,5,8) | Students whose length of study is fast with GPA is good or excellent, get the final exam A, and male. | 15.6%                 |
|               |                  | Students whose length of study is average, with GPA is good or excellent, final exam A, and male. | 40.0%                 |
|               |                  | Students with late length of study, with GPA are good or excellent, get final exam A, and male. | 44.4%                 |
| 9             | 5th (node 0,2,6,9) | Students whose length of study is fast with GPA is good or excellent, get final exam (B+ & B) or (C+ & C) from mathematics major. | 0.0%                  |
|               |                  | Students whose length of study is average, with GPA is good or excellent, final exam grade (B+ & B) or (C+ & C) from mathematics major. | 41.4%                 |
|               |                  | Students with late length of study with GPA are good or excellent, get final exam (B+ & B) or (C+ & C) from mathematics major. | 58.6%                 |
| 10            | 6th (node 0,2,6,10) | Students whose length of study is fast with GPA is good or excellent, get final exam (B+ & B) or (C+& C), majoring in physics, chemistry, and biology. | 10.7%                 |
|               |                  | Students whose length of study are late with GPA are good or excellent, receiving final exam grade (B + or B) or (C+ & C) from a physics, chemistry, and biology program of study. | 44.1%                 |
|               |                  | Students with late length of study with GPA are good or excellent receive final exam grade (B+ & B) or (C+& C) from a major in physics, chemistry, and biology. | 45.2%                 |

* The highest percentage of the length of study in each category.
From the results of the CHAID method has been known segments that are formed from the data alumni of FMIPA Unsyiah. The results of this classification can also be used to identify segments of students who are still active. It is known from the alumni data that the 2nd segment has a late length of study, so in active students who have characteristics such as those segments need to be aware if they have a tendency to be late in their studies or have a late length of study.

The results obtained from the risk and classification tables mention the evaluation of how well the working model can be. The predicted result gives a high error; the result will not be good. The risk estimate of 0.386 indicates that category predictions by the model (fast, average or late length of study) are not satisfying for the research. So the risk of misclassification for the length of study is 38.6%. The classification table also shows the appropriate value, the percentage of accuracy of the model classifies the length of study appropriately by 61.4%, see Table 5.

Table 5. Misclassification observed and predicted study length

| Observed | Predicted | PercentCorrect |
|----------|-----------|----------------|
|          | Fast      | Average  | Late |                |
| Fast     | 0         | 44       | 42   | 0.0%           |
| AverageL| 0         | 155      | 235  | 39.7%          |
| late     | 0         | 56       | 444  | 88.8%          |
| OverallPercentage | 0% | 26.1% | 73.9% | 61.4% |

5. Conclusions
The results of CHAID analysis on undergraduate students of FMIPA Unsyiah who graduated in January 2007 until August 2014 are:
1. The predicting factors that significantly affect the length of study are GPA, education level of the final writing supervisors, the final exam scores, the gender, and the program of study.
2. Characteristics of categories of students whose length of study in time (fast) is similar to those with have average one, which are; students with excellent GPA score, received grade A in their final exam, and are female. While the characteristics of students within late length of study category are students in satisfactory GPA with the level education of their final writing supervisors are doctorate.

References
[1] Tim Penyusun 2010 Panduan Administrasi Akademik Program Sarjana dan Diploma (Aceh: Universitas Syiah Kuala)
[2] Gallagher C A, Monroe H M and Fish J L 2000 An Iterative Approach To Classification Analysis. https://www.casact.org/pubs/dpp/dpp90/90dpp237.pdf
[3] Dudley W N and Dilorio C S J 1999 Nursing Research 48 53
[4] Eherler D and Lehmann T 2001 Responder Profiling with CHAID and Responcy Analysis. http://www.citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.8533&rep=rep1&type=pdf.
[5] Zulvia P, Kurnia A and Soleh A M 2017 AIP Conference Proceedings 1827 020004