CHAPTER I

INTRODUCTION

1.1 Background

Statistical learning refers to a set of tools used to model understand data sets [5]. Along with the development of technologies and growing volume of data, the importance of extracting knowledge from these complex data sets are getting more apparent. In recent years there has been an increased interest in using data mining in various fields, one of them being educational field. One educational problems that can be solved with data mining is the prediction of students' academic performances, whose goal is to predict an unknown variable (outcome, grades or scores). One of them are Osmanbegovic and Suljic, who used Naive Bayes, Multilayer Perceptron, and J48 Decision Tree to predict final students' grades (fail and pass) in course ”Business Informatics” [10]. Using the regression methods, Kotsiantis and Pintelas predicted a student's marks (pass and fail classes) [7].

Generally, data mining can be defined as mechanisms and techniques utilized to extract information from huge amount of data. The origin of data mining can be traced back to late 1980s when the term began to be used within the research community [1]. By the early 1990s data mining was commonly referred as a sub-process within a larger process called Knowledge Discovery in Databases (KDD) [1]. The most commonly used definition of KDD is that attributed to Fayyad et al. : ”The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” [14]. Other sub-processes that form part of the KDD process are data preparation and the analysis or visualization of results [14].

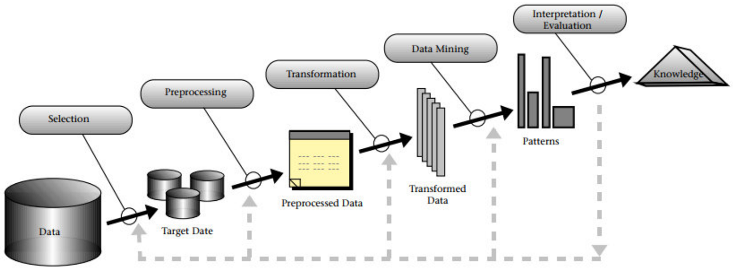


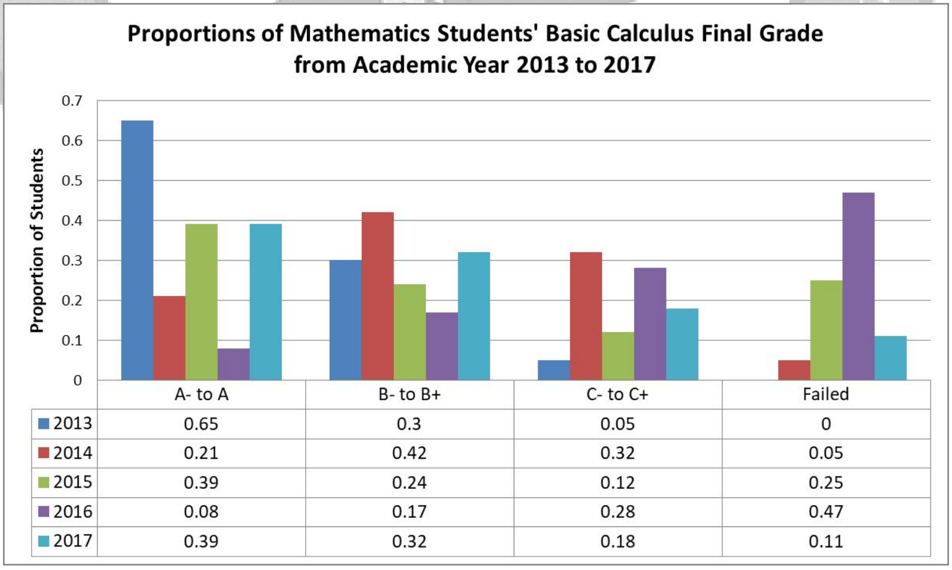
Figure 1.1. Knowledge Discovery in Databases

Basically, there are four different learning styles in data mining applications;

classification, association, prediction, and clustering [1]. The main concern of classification is the construction of ”classifiers” that can be applied to ”unseen” data to categorize that data into groups or classes. From this definition alone, classification has parallels with clustering, in which clustering’s main concern is to categorize data into groups or classes. However classification requires pre-labeled training data from which the classifiers can be built [1]. Therefore classification is sometimes referred to as supervised learning while clustering is referred to as unsupervised learning. The classifiers can take many forms: decision trees, Support Vector Machines (SVM) as first proposed by Vapnik [2]. The most influential decision tree generation algorithm with respect to data mining is Quinlan’s C4.5 algorithm [11]. Other notable classification techniques include Naïve Bayes by Hand and Yu [15].

Classification is the most familiar and most effective data mining technique used to classify and predict values, Educational Data Mining (EDM) included. Therefore, it was used in this research paper to analyze collected students’ information through a survey, and provide classifications based on the collected data to predict and classify a student’s performance in Statistics and Calculus courses at the end of the semester.

In Universitas Pelita Harapan, a data of Mathematics students’ final grades in Basic Calculus from 2013 to 2017 are recorded, and the proportions of students attaining a certain grade are shown in Figure 1.2.



	A- to A	B- to B+	C- to C+	Failed
🟦 2013	0.65	0.3	0.05	0
🟥 2014	0.21	0.42	0.32	0.05
🟩 2015	0.39	0.24	0.12	0.25
🟪 2016	0.08	0.17	0.28	0.47
🟦 2017	0.39	0.32	0.18	0.11

Figure 1.2. Proportions of Mathematics Students’ Basic Calculus Final Grade from Academic Year 2013 to 2017

From Figure 1.2, it can be seen that batch 2013 has the highest proportion of students attaining final grades between A- and A at 0.65 with no students failing the course. Batch 2016 has the highest proportion of students failing the course at 0.47 and the lowest proportion of students attaining final grades between A- and A at 0.08. the proportion of students failing the course is seen to be rising from year 2013 peaking at 2016 and fell at the year 2017. Aside from that, no clear trend can be seen from the past 4 years.

In order to get more information on what affects a student's performance in these important, basic courses, classification models to predict a student's final grade based on selected attributes using data mining techniques will be constructed.

1.2 Problem Statement

The main problem that will be discussed on this thesis is whether the models developed using data mining techniques can be used to predict a student's final grade accurately and tell us the relationships between attributes involved and final grade. Specific research questions include as followings.

1. What are the attributes that have significant effect on a student's performance in Statistics and Calculus courses?
2. How do these attributes affect the final grade?
3. Are there any relationship between these attributes?
4. How do the models built using data mining techniques compare to each other?
5. Which model can give the best performance?

1.3 Objectives

The main objective of this thesis is to develop models that can accurately predict a student's final grade in Statistics and Calculus courses, and determine the relationships of the attributes involved. Listed below are the outline of our main objective in this thesis.

1. Identify and analyze academic and non-academic attributes that affect a student's performance in Statistics and Calculus courses.
2. Develop models to predict final grade based on these attributes.
3. Implement models to a test data.

4. Evaluate the performances of developed models.
5. Describe and Interpret the results.

To achieve the objectives described above, This thesis proposed several methods, which are listed below.

1. Review related research and literature.
2. Collect data by survey and divide them into training data and test data.
3. Use training data to develop classification models using data mining techniques.
   (a) Naïve Bayes,
   (b) Decision Tree,
   (c) Support Vector Machine,
   (d) deriving the loss function and gradient approximate equation.
4. Verify the performance of developed models by using test data, describe and interpret the result.

1.4 Restrictions and Assumptions

1. The sample data is taken from students of batch 2017 and 2018 who are enrolled in a department under Faculty of Science and Technology.
2. The sample taken in our training data are independent and identically distributed.
3. The input variables/attributes are independent to each other as a condition for Naive Bayes algorithm.
4. The students are assumed to have the same basic knowledge of mathematics when they were enrolled in Statistics and Calculus courses.
5. The learning environment is assumed to be the same for each classes of Statistics and Calculus courses, and therefore doesn't have an effect on the performance of the students.

6. It is assumed that the amount of time spent on doing academic related activities outside of lectures will have a positive relationship with the amount of time spent doing individual course related studies.
7. It is assumed that the students filled the questionnaire truthfully.

1.5 Benefits

There are several benefits of this research, both theoretical and practical, including the followings.

1.5.1 Theoretical Benefits

As a real-life application of data mining techniques that can be evaluated and analyzed for further improvements.

1.5.2 Practical Benefits

Both lecturers and students can have an insight of significant and not significant variables, academic or not, that influence the students' performance in related courses, and can use that information for their own benefits.

1.6 Thesis Structure

The writing structure of this thesis is as detailed below.

1. Chapter I describes the background, problem statement, objectives, restrictions, and methodology.
2. Chapter II describes the basic theories used in this research. It consists of the definition of statistical learning and the theories behind the data mining techniques (Naïve Bayes, Decision Tree, and Support Vector Machine) that will be used in this research.
3. Chapter III describes the steps taken to reach the objectives stated in Chapter I. It consists of steps and algorithms for data collecting, data processing, model building, and model evaluation.
4. Chapter IV contains the results obtained from applying the steps and algorithms explained in Chapter III and analysis of the results..

5. Chapter V summarizes achieved results, draw conclusions and proposes directions for further research.