UNIVERSITI PUTRA MALAYSIA

MODELING STUDENTS' BACKGROUND AND ACADEMIC PERFORMANCE WITH MISSING VALUES USING CLASSIFICATION TREE

By

NORSIDA BINTI HASAN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

December 2014

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

MODELING STUDENTS' BACKGROUND AND ACADEMIC PERFORMANCE WITH MISSING VALUES USING CLASSIFICATION TREE

By

NORSIDA BINTI HASAN

December 2014

Chair: Mohd Bakri Adam, Ph.D.

Faculty: Institute for Mathematical Research

Student's academic performance is a prime concern to high level educational institution since it will reflect the performance of the institution. The differences in academic performance among students are topics that has drawn interest of many academic researchers and our society. One of the biggest challenges in universities decision making and planning today is to predict the performance of their students at the early stage prior to their admission. We address the application of inferring the degree classification of students using their background data in the dataset obtained from one of the high level educational institutions in Malaysia. We present the results of a detailed statistical analysis relating to the final degree classification obtained at the end of their studies and their backgrounds. Classification tree model produce the highest accuracy in predicting student's degree classification using their background data as compared to Bayesion network and naive Bayes. The significance of the prediction depends closely on the quality of the database and on the chosen sample dataset to be used for model training and testing. Missing values either in predictor or in response variables are a very common problem in statistics and data mining. Cases with missing values are often ignored which results in loss of information and possible bias. Surrogate split in standard classification tree is a possible choice in handling missing values for large dataset contains at most ten percent missing values. However, for dataset contains more than 10 percent missing values, there is an adverse impact on the structure of classification tree and also the accuracy. In this thesis, we propose classification tree with imputation model to handle missing values in dataset. We investigate the application of classification tree, Bayesian network and naive Bayes as the imputation techniques to handle missing values in classification tree model. The

investigation includes all three types of missing values machanism; missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Imputation using classification tree outperform the imputatation using Bayesian network and naive Bayes for all MCAR, MAR and MNAR. We also compare the performance of classification tree with imputation with surrogate splits in classification and regression tree (CART). Fifteen percent of student's background data are eliminated and classification tree with imputation is used to predict student's degree classification. Classification tree with imputation model produces more accurate model as compared to surrogate splits.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

PERMODELAN LATARBELAKANG DAN PENCAPAIAN AKADEMIK PELAJAR DENGAN NILAI HILANG MENGGUNAKAN POKOK KLASIFIKASI

Oleh

NORSIDA BINTI HASAN

December 2014

Pengerusi: Mohd Bakri Adam, Ph.D.

Fakulti: Institut Penyelidikan Matematik

Pencapaian akademik pelajar menjadi keutamaan di institusi pengajian tinggi kerana ia mencerminkan prestasi institusi tersebut. Perbezaan pencapaian aka-demik di kalangan pelajar sentiasa menjadi topik perbincangan yang menarik minat ramai penyelidik dan masyarakat umum. Di dalam kajian ini, analisis statistik memperlihatkan perkaitan di antara pencapaian akademik pelajar semasa bergraduat dan latarbelakang mereka. Salah satu daripada cabaran besar yang dihadapi oleh pembuat dasar serta perancangan universiti hari ini adalah untuk meramal pencapaian pelajar semasa awal kemasukan mereka ke universiti. Kami menangani aplikasi penafsiran klasifikasi ijazah pelajar menggunakan data latarbelakang dalam set data yang diperolehi daripada salah satu Institusi Pengajian Tinggi Awam (IPTA) di Malaysia. Kami paparkan hasil analisis statistik yang terperinci berkaitan dengan klasifikasi ijazah yang diperolehi semasa tamat pengajian berdasarkan latarbelakang mereka. Model pokok klasifikasi menghasilkan kejituan tertinggi berbanding dengan rangkaian Bayesian dan Bayes naif. Signifikasi ramalan sangat bergantung kepada kualiti pangkalan data serta bergantung juga kepada sampel yang akan digunakan untuk model latihan dan model pengujian. Nilai hilang samada dalam pembolehubah peramal atau pembolehubah tindakbalas merupakan masalah yang biasa dalam bidang statistik dan perlombongan data. Kes-kes nilai hilang yang selalunya diabaikan menyebabkan kehilangan maklumat dan boleh meghasilkan keputusan yang berpihak. Pemisah gantian (*surrogate split*) dalam pokok klasifikasi piawai boleh menjadi pilihan semasa mengendalikan nilai-nilai yang hilang bagi set data besar yang mengandungi paling banyak 10 peratus nilai hilang. Walau bagaimanapun bagi set data yang mengandungi lebih daripada 10 pratus nilai hilang, terdapat impak yang buruk ke atas struktur pokok klasifikasi dan kejituan klasifikasi. Di dalam tesis ini, kami mencadangkan

model pokok klasifikasi dengan imputasi untuk menangani nilai hilang dalam set data. Kami mengkaji penggunaan pokok klasifikasi, rangkaian Bayesian dan Bayes naif sebagai teknik imputasi untuk menangani nilai hilang dalam model pokok klasifikasi. Kajian ini meliputi kesemua tiga jenis mekanisma nilai hilang: hilang sepenuhnya secara rawak (MCAR), hilang secara rawak (MAR) dan hilang bukan secara rawak (MNAR). Imputasi menggunakan pokok klasifikasi mempunyai kejituan mengatasi imputasi menggunakan rangkaian Bayesian dan Bayes naif bagi kesemua mekanisma iaitu MCAR, MAR dan MNAR. kami juga membandingkan pencapaian model pokok klasifikasi dengan imputasi dengan kaedah pemisah gantian dalam pokok klasifikasi dan regresi piawai (CART). Lima belas peratus daripada data latarbelakang pelajar dihapuskan dan model pokok klasifikasi dengan imputasi digunakan untuk meramalkan kelas ijazah pelajar. Model pokok klasifikasi dengan imputasi menghasilkan model yang lebih jitu berbanding dengan pemisah gantian.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CART	Classification and Regression Tree
STPM	Malaysian Higher School Certificate
PKPG	In-service Teacher Education Programme
KDPK	In-service Teachers with Diploma in Special Education Programme
MCAR	Missing Completely At Random
MAR	Missing At Random
MNAR	Missing Not At Random
RRP	Random Recursive Partitioning
ITree	Imputation Tree
UPSI	Universiti Pendidikan Sultan Idris
FB	Faculty of Languages
FPE	Faculty of Business and Economics
FSKPM	Faculty of Cognitive Science and Human Development
FSM	Faculty of Music
FSS	Faculty of Sports Science
FSSK	Faculty of Human Sciences
FST	Faculty of Science and Technology
FTMK	Faculty of Information Technology and Communication
CGPA	Cumulative Gred Point Average
FP	False Positive
FN	False Negative
TP	True Positive
TN	True Negative

CHAPTER 1

INTRODUCTION

1.1 Student's Academic Performance

Student performance is a prime concern to high level educational institution since it will reflect the performance of the institution. Researchers and educators conducted many studies and experiments to determine the factors that affect student's performance. Socio-demographic characteristics such as age, gender, marital status, family status, ethnicity and previous achievement are shown to affect their undergraduate academic performance (Brown and Burkhardt, 1999; Clayton and Cate, 2004; Stevens et al., 2004; Ding et al., 2006; Ismail and Othman, 2006; Lietz, 2006; Gibb et al., 2008).

One of the biggest challenges in university decision making and planning today is to predict the performance of their students at the early stage prior to their admission. This is not an easy task but the findings is important to assist the university in determining future policy on student admissions and to provide the necessary plans to improve student performance. One of the significant facts in universities is the explosive growth of students' information in databases system. As the amount of these data increasing rapidly, the interest has grown in tapping these data to extract the hidden information that is valuable to the management. The discipline concern with this task is known as data mining. Data mining techniques can be used to extract meaningfull information and to develop significant relationships among variables stored in the students' background data.

1.2 Classification Tree

In this thesis, we applied classification tree because it produced the best accuracy as compared to naive Bayes and bayesian network. Classification and Regression tree (CART) is a supervised learning method that constructs a flow-chart-like tree as the classification model from the data and uses the tree model to classify the future data. Classification tree is a flow-chart-like tree structure consists of one root, branches, nodes and leaves. Classification tree analysis is a form of binary recursive partitioning where a node (parent node) in a decision tree, can only be split into two child nodes. The term "recursive" refers to the fact that the binary partitioning process can be applied over and over again (Breiman et al., 1984).

Classification tree is usually obtained in two steps. Initially a large tree is grown using a greedy algorithm, and then this tree is pruned by deleting bottom nodes through a process of statistical estimation. The greedy algorithm typically grows a tree by sequentially choosing splitting rules for nodes on the basis of maximizing some fitting criterion. All possible splits consist of possible splits of each predictor variable. This step generates a sequence of trees, each of which is an extension of previous trees. A single tree is then selected by pruning the largest tree according

to a model selection criterion such as cost-complexity pruning, cross-validation, or even multiple tests of whether two adjoining nodes should be collapsed into a single node (Breiman et al., 1984). This pruning process ensures the tree which fits the information in the learning dataset, but does not overfit the information.

The CART begins with the entire sample of student's data. This entire sample is heterogeneous, consisting of all students. It then divides up the sample according to a splitting rule and a goodness of split criterion. Each internal node has an associated splitting rule which uses a predictor variable to assign observations to either its left child node or right child node. The splitting rules for our sample are question of the form, "Is the FACULTY F2, F3 or F6?" or put more generally, is $X \in d$, where X are some variables and d is some elements within that variable. If the criterion is satisfied, we follow the division to the left and if the said criterion is not satisfied, we follow the division to the right. Such questions are used to divide or split the sample. The CART algorithm considers all possible variables and all possible values in order to find the best split. The best split refers to the question that splits the data into two parts with maximum homogeneity (Breiman et al., 1984). Maximum homogeneity of child nodes is defined by impurity function $i(t)$ which is equivalent to the maximization of change of impurity function Δi_t as shown by

$$\Delta i_t = i(t_p) - P_l i(t_l) - P_r i(t_r),$$

where

- t_p is a parent node,
- $i(t_p)$ is the impurity measure for the parent node,
- P_l is the proportion of the samples in node t that go to the left node t_l,
- P_r is the proportion of the samples in node t that go to the right node t_r,
- $i(t_l)$ is the impurity measure for left child node,
- $i(t_r)$ is the impurity measure for right child node.

Since the parent node is constant for any split, then, the maximization problem is equivalent to minimizing the following expression

$$P_l i(t_l) + P_r i(t_r). \tag{1.1}$$

Equation (1.1) implies that CART will compare different splits and determines which of these will produce the most homogeneous subsamples. Common measures are:

1.3 Problem Statements

Student's performance is a prime concern to high level educational institution because it will reflect the performance of the institution. The differences in academic performance among students are a topic that has drawn interest of many academic

researchers and our society. However, the student's performance is not encouraging since less than 4 percent of student in public university in Malaysia obtained first class degree classification upon graduation (Graduate Tracer Study Report 2009, Retrieved 14/11/2012).

Even though there is a weak relationship between employees performance with CGPA as reported by Hashim (2012), employers usually use the students academic performance as the selection criteria to shortlist the candidates for the interview. Hashim (2012) also stated that several well-established companies in Malaysia limit their recruitment only to those students who achieve 3.00 CGPA and above. Therefore, the biggest challenges in university decision making and planning today is to understand the student's performance pattern and then to predict the performance of the students at the early stage prior to their admission. To our knowledge, there is no study has yet been made to model student's background data from all faculties in a university to classify and predict the final degree classification. The findings can assist the university in determining future policy on student admissions and to provide the necessary plans to improve student performance.

The significance of the prediction depends closely on the quality of the database and on the chosen sample dataset to be used for model training and testing. Unfortunately, missing values either in predictor or in response variables are a very common problem in statistics and data mining. Cases with missing values are often ignored and standard methods for complete data are run on the remaining data cases. If the rate of missing values is less than 1 percent, missing values are considered trivial, 1 percent to 5 percent missing values are considered manageable, 5 percent to 15 percent missing values require sophisticated methods to handle and more than 15 percent may severely impact any kind of interpretation (Acuna and Rodriguez, 2004; Peng et al., 2005). To our knowledge, there is no study has yet been made of sensitivity of missing data in the classification tree structure and classification accuracy with big sample size.

Case deletion method discards valuable information about features that are observed which results in loss of information and possible bias (Shafer, 2002; Little and Rubin, 2002). One effective way of dealing with missing values is to impute them with some reasonable value before proceed with inference. The key to imputation techniques is to substitute with the most probable values and meanwhile preserve the joint relationships between variables. Imputation by a constant using mean or mode values will ignore the between-attribute relationships and assumes that all missing values represent the same value, probably leading to considerable distortions. Surrogate split in standard classification tree is a possible choice for large dataset contains at most ten percent missing values. However, for dataset contains more than 20 percent missing values, there is an adverse impact on the accuracy of the classification tree (Peng et al., 2005). Peng et al. (2005); Saar-Tsechansky and Provost (2007) showed that imputation methods are able to increase the accuracy in classifcation model. However, these research are limited to missing completely at random (MCAR). Tree-based approach for missing values

imputation was proposed by Vateekul and Sarinnapakorn (2009). However, this method is applicable for quantitative data.

In this thesis, we propose classification tree model with imputation to handle missing values in dataset. We investigate the application of classification tree, Bayesian network and naive Bayes as the imputation techniques to handle missing values in classification tree model. The investigation includes all three types of missing values machanism; missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

1.4 Research Objectives

The main objective of this research is to develop an accurate model to predict student's academic performance using their background data with the present of missing values. To achieve the objective, the following sub-objectives are adopted:

1. To propose classification tree model with imputation to handle dataset with missing data.
2. To propose an imputation method for three types of missing data mechanism: MCAR, MAR and MNAR.
3. To propose the predictor variable for student's academic performance.

1.5 Research Contributions

There are three main contibution of this research:

1. Classification tree model with missing data imputation for predicting the student's academic performance based on their background data.
2. Imputation method for three types of missing data mechanism: MCAR, MAR and MNAR.
3. Predictor variables for student's academic performance.

1.6 Organization of Thesis

This thesis contains seven chapters; Introduction, Literature Review, Research Methodology, Data Pre-processing and Missing Data Injection, Model Development, Experimental Results and Conclusion and Future Work. The details of the chapter are as follow:

Chapter 1 provides an overview of the thesis, such as background studies, problem statement, objectives and reseach contribution.

Chapter 2 presens the literature reviews on the existing work to determine the factors that affect student's performance. This description is particularly focused on socio-demographic characteristics such as age, gender, marital status, family

status and ethnicity. We present an overview of the major data mining techniques used in predicting student's academic performance. Classification tree is the common method for mining student's data. However it is sensitive to the presence of missing values. The reviews on sensitivity of missing values and how to handle missing values in data mining are also presented.

Chapter 3 provides the methodology applied in this study. Research framework including data, data pre-processing and missing data injection, model design, model development and model implementation are briefly explained in this chapter.

Chapter 4 presents the data pre-processing and missing data injection. The descriptive data analysis is carried out to investigate the relationship between categorical variables of student's academic performance according to their gender, university academic intake category, age and race. Data mining techniques namely classification tree, Bayesian network and naive Bayes are applied to student's background data to predict student's degree classification. We also simulate the student's background data using the correlation of the actual data, then, we simulate the three types of missing data mechanism (MCAR, MAR and MNAR). The influence of missing values in classification tree, Bayesian network and naive Bayes are then investigated by removing levels of student's background data.

Chapter 5 provides a detailed explaination on the development of classification tree with imputation model. The imputation of missing values using three imputation techniques; classification tree, Bayesian Network and naive Bayes are explained. All three imputation techniques are implemented on datasets having three types of missing values mechanism; MCAR, MAR and MNAR.

Chapter 6 presents the results of experiments applied to real student's background and academic performance dataset to evaluate the performance of proposed algorithms.

Chapter 7 gives concluding remarks and directions of future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Factors Affecting Academic Performance

The improvement of students' performance is one of the priority goals in education. In the past few decades, researchers and educators conducted many studies and experiments to determine the factors that affect students' performance. Socio-demographic characteristics such as age, gender, marital status, family status, ethnicity, previous achievement are shown to affect their undergraduate academic performance (Brown and Burkhardt, 1999; Clayton and Cate, 2004; Stevens et al., 2004; Ding et al., 2006; Ismail and Othman, 2006).

Upon completing upper secondary education, there are several routes to higher education in Malaysia. The first route is a pre-university or matriculation programme and the examination result will be used for entrance to public universities. The second route is a two-year sixth form of secondary education programme, where students take a centralized examination known as the Malaysian Higher School Certificate (Sijil Tinggi Pelajaran Malaysia) or STPM. The third route is for in-service teacher to obtain the bachelor degree. In-service Teacher Education Programme (Program Khas Pensiswazahan Guru) or PKPG is joint programme with the Ministry of Education which offering special degree programmes to non-graduate teachers. The fourth route is a special degree programme known as In-service Teachers with Diploma in Special Education Programme (Program Pensiswazahan Guru-guru lepasan Kursus Diploma Perguruan Khas) or KDPK. The fifth routes to this higher educational institution are second channel, special case or diploma holder. Each route has its own education system, varying in standard of grading.

The first socio-demographic variable considered in this study is gender. According to Ismail and Awang (2008), the effects of gender differences were significant to distinguish between the higher and lower achievers or medium and lower achievers with girls were more likely as compared to boys to be in the higher group of achievers but not between the high and medium achievers. Ding et al. (2006) used a longitudinal multilevel modeling to examine overall performance of two school districts in two different states in the United States. The results suggested that both male and female students demonstrated the same growth trend in mathematics performance over time, but females' mathematics grade-point average was significantly higher than males.

Some researchers investigated the relationship between race and academic achievement but they showed conflicting results. For example, Clayton and Cate (2004) discovered that students from different ethnic backgrounds differed in the levels of academic achievement. They found that whites and hispanics students were outperform asian americans students in the MBA program at Northern Kentucky University. Study on hispanic and caucasian students done by Stevens et al. (2004)

reported a statistically different motivation and performance across ethnicity. On the other hand, Ismail and Othman (2006) found that the role of ethnicity in explaining the differences in academic performance in three faculties at University of Malaya, Malaysia was not significant.

Findings from previous study on the role of age in influencing the academic achievement were similarly contested. Although Graham (1991) found that the age is insignificantly correlated with academic performance, Peiperl and Trevelyan (1997); Ali et al. (2013) discovered a negative correlation between age and performance, with younger students performed better than older ones. Archer et al. (1999) and Hayes et al. (1997) agree with Graham (1991) i.e. the older students perform at par with younger students.

2.2 Meta Analysis of Students' Performance Between Gender

During the last two decades, the issue of gender differences in academic performance has been addressed by carrying out meta-analyses of studies on academic performance such as in Lindberg et al. (2010), Gibb et al. (2008) and Lietz (2006) but they showed conflicting results.

For example, Gibb et al. (2008) reported a tendency for females to score better than males on standardised tests and to achieve more school and post-school qualifications. Lietz (2006) discovered that females outperformed males in a meta-analysis of large-scale studies between 1970 and 2002 in the area of reading achievement at the secondary school level.

Studies by Naderi et al. (2009) discovered that males and females perform similarly in mathematics. Lindberg et al. (2010) perform meta-analysis using 242 studies published between 1990 and 2007 and also reported that males and females perform similarly in mathematics. Naderi et al. (2009) revealed that there is no significant difference between CGPA and gender among undergraduate students.

2.3 Predicting Academic Performance Using Classification and Regression Tree

Classification or regression tree is the most common method used to model and predict the students' performance (Kumar and Vijayalakshmi, 2011; Vialardi et al., 2009; Vandamme et al., 2007; Ibrahim and Rusli, 2007; Nghe et al., 2007; Adeyemo and Kuye, 2006; Al-Radaideh et al., 2006). The capabilities of data mining in the context of higher educational system were presented by Delavari et al. (2008) who proposed an analytical guideline to enhance the current decision making processes in higher education institutions and then applying data mining techniques to discover new explicit knowledge which could be useful for the decision making

processes.

A comparative study among data mining tools for predicting the academic performance showed that they are able to produce similar levels of accuracy. However, findings from the study on the best data mining tool in predicting the academic performance are similarly contested. Nghe et al. (2007) compared the accuracy of decision tree and Bayesian network algorithm for predicting the academic performance of undergraduate and postgraduate students at Can Tho University (CTU), Viet Nam and Asian Institute of Technology (AIT), Thailand. Despite of the diversity of these two student populations was very different, both decision tree and Bayesian network were able to achieve similar levels of accuracy for predicting student performance. Their findings showed that the decision tree consistently outperformed the Bayesian Network algorithm with 3 to 12 percent more accuracy.

The prediction of secondary student grades using previous school grades, demographic, social and other school related data was carried out by Cortez and Silva (2008). Four data mining methods namely decision trees, random forests, neural networks and support vector machines were tested on 1044 students' data. The study revealed that student achievement was highly affected by previous performance and the tree based algorithms outperform other methods. The current research on application of decision tree in education is done by Kumar and Vijayalakshmi (2011) to predict students performance in the final exam using their internal assessment data.

The application of decision tree, Artificial Neural Network and linear discriminant analysis to predict the low-risk, medium-risk and high-risk groups of students was done by Vandamme et al. (2007). The accuracy of prediction obtained was not particularly good due to the difficulty to classify students into groups before the first university examinations. Overall total correct classification rate for linear discriminant analysis was 57.35 percent, followed by neural network approach with 51.88 percent and the least accurate was decision tree with 48.65 percent. Ibrahim and Rusli (2007) constructs three prediction models: Artificial Neural Network, decision tree and linear regression to predict the final CGPA of the students. Their result showed that all three models produced more than 80 percent of accuracy with artificial neural network outperforms the other models.

A recommendation system based on classification tree to assist students to make decision on their academic itineraries was proposed by Vialardi et al. (2009). It provides support for the students at Universidad de Lima in course enrolment based on the pattern of previous students with similar achievement.

In real world applications of large amount of data might have a skewed distribution or imbalanced class categories. Dataset is imbalance if one of the classes having more sample than other classes. Cassification of data becomes difficult as most of data mining algorithm are more focusing on classification of major sample while ignoring or misclassifying. Sampling is a method to solve the imbalanced dataset at the data-preprocessing stage. There are two sampling technique that are widely

used namely under-samplig and over-sampling. The process of removing a sample is known as under-sampling while the process of adding a sample is known as oversampling (Longadge et al., 2013).

The random undersampling method is less favorable as it can potentially remove certain important examples (Jo and Japkowicz, 2004; Batista et al., 2004; Chawla et al., 2002). One of the widely used over-sampling method is proposed by Chawla et al. (2002) who used KNN to generate synthetic data for minority class.

2.4 Missing Data and Imputation using Classification Tree

Data mining process deals significantly with prediction, classification, pattern recognition, and association. The significance of the analysis depends closely on the quality of the database and on the chosen sample dataset to be used for model training and testing. However, in either data mining or statistics methods, base-formed models are designed assuming that data are usually complete. That is, all of the values in the data matrix, with the rows being the observations (instances) and the columns being the variables (attributes), are observed.

However, missing data either in predictor or response variables are a very common problem. Cases contain missing responses, or with missing predictor variables are often ignored, and standard methods for complete data are run on the remaining data cases. Case deletion discards valuable information about features that are observed which results in loss of information and possible bias (Shafer, 2002; Little and Rubin, 2002). According to Allison (2002), case deletion must only be performed if analysis of the remaining complete sample will not lead to biased estimates. However, it is rather wasteful since it usually decreases the information content of the sample. For all these reasons, its use is generally not recommended.

The sensitivity of missing data depends on the pattern of the missing data and the mechanisms that led to missing data. According to Diggle and Kenward (1994) and Little and Rubin (2002), there are three types of missing data mechanism. The first type defined as missing completely at random (MCAR) arises when the probability of missingness of an observation is unrelated to any variables under study. The second type is missing at random (MAR), occurs when response probability does not depend on the missing variables. The third type is non-ignorable missingness or known as missing not at random (MNAR), if the response probabilities depend on missing variables. In this research, we study the influence of MCAR, MAR and MNAR to the classification tree.

There are many approaches to handle problem of missing values as describe in Han and Kamber (2001). The first approach is to ignore the cases containing missing values (also known as listwise deletion) which may cause loss of information. The second approach is to fill in the missing values manually, which is infeasible tasks due to the size of the datasets. The third approach is to impute the missing values by a constant and the fourth approach is to impute the most probable value to fill

the missing ones. Imputation is a strategy for handling missing data that is widely used in the statistical inference. The key to imputation techniques is to substitute some reasonable value for each missing data before proceed with inference. Imputation by a constant will ignore the between-attribute relationships and assumes that all missing values represent the same value, probably leading to considerable distortions.

The classification tree can also be used to impute missing values by using terminal node values as imputed values for missing data. The use of classification tree for multivariate imputation is a relatively new approach. One advantage of classification tree for imputation is the nonparametric nature of this approach. Moreover, compared to parametric technique, nonparametric technique, is relatively less sensitive to outliers. Comparison of existing missing data method in various classification tree algorithm was done by Ding and Simonoff (2010). They study the effectiveness of six popular missing data methods: probabilistic split, complete case method, grand mode/mean imputation, separate class, surrogate split and complete variable method for binary response data. They showed that the relationship between the missingness and the dependent variable as well as the exixtence of missing values in the testing dataset are the main criteria to distinguish the different missing data methods.

Imputation class created by the CART with weighted sequential hot deck imputation methodology within the imputation classes is the best method discussed in Creel and Krotki (2006). Ssali and Marwala (2008) also used decision tree combined with neural network to impute missing data. Iacus and Porrob (2007) introduced the applications of the random recursive partitioning (RRP) that generates a proximity matrix which can be used in non-parametric matching problems such as hot-deck missing data imputation and average treatment effect estimation. The most recent work on imputation using decision tree was done by Vateekul and Sarinnapakorn (2009) who developed an imputation algorithm known as Imputation Tree (ITree) to handle missing in quantitative dataset. The review on various types of missing data methods and software in the context of a motivating example from a large health services research dataset can be found in Hortan and Kleinman (2007).

2.5 Conclusion

The literature reviews shows that socio-demographic characteristics such as age, gender, marital status, family status and ethnicity are some of the factors that affect student's performance. Classification tree model is the common method for mining student's data to predict their academic performance. The study on dealing with missing values is very important because the tree model is very sensitive to the missing values.